Form Approved Through 11/30/2010

PF 12152040 Application

Do not exceed character length restrictions indicated. MAR 2 8 20(

PI: KOSOROK, MICHAEL R
1 P01 CA142538-01
Dual:
IRG: ZCA1 SRC(99)

Council: 10/2009
Received: 01/28/2009

1. TITLE OF PROJECT *(Do not exceed 81 characters, including spaces and punctuation.)*
Statistical Methods for Cancer Clinical Trials

2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION ☐ NO ☒ YES *(If "Yes," state number and title)*
Number: PAR-09-025 Title: National Cancer Institute Program Project (P01) Applications

3. PROGRAM DIRECTOR/PRINCIPAL INVESTIGATOR		New Investigator ☒ No ☐ Yes
3a. NAME (Last, first, middle) Kosorok, Michael R.	3b. DEGREE(S) PhD MS MM	3h. eRA Commons User Name MKOSOROK
3c. POSITION TITLE Professor and Chair	3d. MAILING ADDRESS *(Street, city, state, zip code)* Department of Biostatistics CB #7420, McGavran-Greenberg Hall University of North Carolina at Chapel HillChapel Hill, NC 27599-7420	
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Biostatistics		
3f. MAJOR SUBDIVISION School of Public Health		
3g. TELEPHONE AND FAX *(Area code, number and extension)* TEL: (919) 966-8107 FAX: (919) 966-3804	E-MAIL ADDRESS: Kosorok@bios.unc.edu	

4. HUMAN SUBJECTS RESEARCH	4a. Research Exempt	If "Yes," Exemption No.
☐ No ☒ Yes	☒ No ☐ Yes	

4b. Federal-Wide Assurance No.	4c. Clinical Trial	4d. NIH-defined Phase III Clinical Trial
FWA-4801	☒ No ☐ Yes	☒ No ☐ Yes

5. VERTEBRATE ANIMALS ☒ No ☐ Yes	5a. Animal Welfare Assurance No.

6. DATES OF PROPOSED PERIOD OF SUPPORT *(month, day, year—MM/DD/YY)*		7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT	
From	Through	7a. Direct Costs ($)	7b. Total Costs ($)	8a. Direct Costs ($)	8b. Total Costs ($)
12/01/09	11/30/14	$2,444,504	$2,865,864	$12,350,257	$14,408,618

9. APPLICANT ORGANIZATION	10. TYPE OF ORGANIZATION
Name: University of North Carolina at Chapel Hill Address: Office of Sponsored Research Administrative Office Building Suite 2200 104 Airport Drive, CB #1350 Chapel Hill, NC 27599-1350	Public: → ☐ Federal ☒ State ☐ Local Private: → ☐ Private Nonprofit For-profit: → ☐ General ☐ Small Business ☐ Woman-owned ☐ Socially and Economically Disadvantaged
	11. ENTITY IDENTIFICATION NUMBER 156-6001393A1 DUNS NO. 608195277 Cong. District 4

12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE	13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION
Name: John Gallagher Title: Interim Director Address: Office of Sponsored Research Administrative Office Bldg., Suite 2200, CB #1350 104 Airport Drive, Chapel Hill, NC 27599-1350 Tel: (919) 966-3411 FAX: (919) 962-3352 E-Mail: resadminosr@unc.edu	Name: Tony G. Waldrop Title: Vice Chancellor for Research & Econ. Dev. Address: Office of Sponsored Research, UNC-CH AOB, Suite 2200, CB #1350 104 Airport Dr., Chapel Hill, NC 27599-1350 Tel: (919) 966-3411 FAX: (919) 962-3352 E-Mail: resadminosr@unc.edu

14. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.

SIGNATURE OF OFFICIAL NAMED IN 13. *(In ink. "Per" signature not acceptable.)* Acting For Tony G. Waldrop

DATE 3/28/09

Use only if preparing an application with Multiple PDs/PIs. See http://grants.nih.gov/grants/multi_pi/index.htm for details.

Contact Program Director/Principal Investigator (Last, First, Middle): Kosorok, Michael R., et al.

3. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR

3a. NAME (Last, first, middle)	3b. DEGREE(S)	3h. NIH Commons User Name
Davidian, Marie	PhD MS	davidian

3c. POSITION TITLE
Professor

3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT
Statistics

3f. MAJOR SUBDIVISION
College of Physical and Mathematical Sciences

3d. MAILING ADDRESS *(Street, city, state, zip code)*
Department of Statistics
Campus Box 8203
2501 Founders Drive
Raleigh, NC 27695-8203

3g. TELEPHONE AND FAX *(Area code, number and extension)*
TEL: (919) 515-1940 FAX: 919) 515-7591

E-MAIL ADDRESS:
davidian@stat.ncsu.edu

3. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR

3a. NAME (Last, first, middle)	3b. DEGREE(S)	3h. NIH Commons User Name
George, Stephen L.	MD MES	georg001

3c. POSITION TITLE
Professor

3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT
Biostatistics and Bioinformatics

3f. MAJOR SUBDIVISION
School of Medicine

3d. MAILING ADDRESS *(Street, city, state, zip code)*
2424 Erwin Road, Suite 802, Room 8037
Duke University Medical Center
Durham, NC 27705

3g. TELEPHONE AND FAX *(Area code, number and extension)*
TEL: 919) 681-2224 FAX: (919) 668-9335

E-MAIL ADDRESS:
stephen.george@duke.edu

3. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR

3a. NAME (Last, first, middle)	3b. DEGREE(S)	3h. NIH Commons User Name

3c. POSITION TITLE

3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT

3f. MAJOR SUBDIVISION

3d. MAILING ADDRESS *(Street, city, state, zip code)*

3g. TELEPHONE AND FAX *(Area code, number and extension)*
TEL: FAX:

E-MAIL ADDRESS:

3. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR

3a. NAME (Last, first, middle)	3b. DEGREE(S)	3h. NIH Commons User Name

3c. POSITION TITLE

3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT

3f. MAJOR SUBDIVISION

3d. MAILING ADDRESS *(Street, city, state, zip code)*

3g. TELEPHONE AND FAX *(Area code, number and extension)*
TEL: FAX:

E-MAIL ADDRESS:

PROJECT SUMMARY (See instructions):

The overall scientific goal of this ambitious program project is to develop highly innovative methods for cancer clinical trials that can hasten successful introduction of effective new therapies into practice. The method of approach is to leverage recent advances in statistical and computational science to create new clinical trial designs and data analysis tools that resolve many of the key scientific limitations of current clinical trial methodology. The projects focus on practical design and analysis problems in Phase II and Phase III clinical trials, the problem of missing data and efficient use of prognostic information, post-marketing surveillance and comparative effectiveness research using clinical trial data, pharmacogenetics and individualized therapies, and the potential of dynamic treatment regimens to improve cancer treatment. The proposed clinical trial design and analysis innovations have the potential to change the prevailing clinical trial paradigm and greatly increase the rate of discovery and translation of new treatments into clinical practice. Our multi-institutional approach includes an effective and energetic process for intense, coordinated implementation, communication and dissemination of results, including developing new software for practical implementation of the newly developed methods. Our comprehensive and novel approach will lead to significant improvements in cancer clinical trial practice that will result in improved health of cancer patients.

RELEVANCE (See instructions):

The proposed program project aims to dramatically improve the efficiency of the cancer clinical trial process in order to improve the health and longevity of cancer patients. This is extremely important to public health since almost all biomedical advances in cancer treatment must pass through the clinical trial process before becoming accepted clinical practice.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location			
Organizational Name: University of North Carolina at Chapel Hill			
DUNS: 608195277			
Street 1: Office of Sponsored Research, CB #1350	Street 2: 104 Airport Dr., Suite 2200		
City: Chapel Hill	County: Orange	State: NC	
Province:	Country: USA	Zip/Postal Code: 27599-1350	
Project/Performance Site Congressional Districts: NC04			

Additional Project/Performance Site Location			
Organizational Name: North Carolina State University			
DUNS: 04-209-2122			
Street 1: Research Admin/ SPARCS	Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514		
City: Raleigh	County: Wake	State: NC	
Province:	Country: USA	Zip/Postal Code: 27695-7514	
Project/Performance Site Congressional Districts: NC02			

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location

Organizational Name: Duke University

DUNS: 044387793

Street 1: Hock Plaza | Street 2: Box 2716 Med Ct.

City: Durham | County: Durham | State: NC

Province: | Country: USA | Zip/Postal Code: 27705

Project/Performance Site Congressional Districts: NC04

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Kosorok, Michael R.	Michael_Kosorok	UNC-CH	Lead PD/PI
Davidian, Marie	Davidian2	NC State University	PD/PI
George, Stephen L.	Georg001	Duke University	PD/PI
Auman, J. Todd	NA	UNC-CH	Co-Investigator
Blackwell, Kimberly L.	NA	Duke University	Co-Investigator
Bondell, Howard D.	NA	NC State University	Co-Investigator
Boos, Dennis D.	dennis_boos	NC State University	Co-Investigator
Cai, Jianwen	Jianwen_Cai	UNC-CH	Project 1 Leader
Carpenter, William R.	Wcarpenter	UNC-CH	Co-Investigator
Chu, Haitao	Hchu11	UNC-CH	Co-Investigator
Crawford, Jeffrey	Crawf006	Duke University	Co-Investigator

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project
Begg, Colin	Cornell University	Member, External Adv. Comm.
Murphy, Susan	University of Michigan	Chair, External Adv. Comm.
Parmigiani, Giovanni	Johns Hopkins University	Member, External Adv. Comm.

Human Embryonic Stem Cells [X] No [ ] Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

KEY PERSONNEL (CONTINUED)

Name	eRA Commons User Name	Organization	Role on Project
Febbo, Philip G.	FEBBO001	Duke University	Co-Investigator
Fine, Jason P.	Jasonp3p	UNC-CH	Co-Investigator
Goldberg, Richard M.	RICHARD_GOLDBERG	UNC-CH	Co-Investigator
Harpole, David H.	HARPO002	Duke University	Co-Investigator
Ibrahim, Joseph G.	JOE_IBRAHIM	UNC-CH	Co-PD/PI
Jung, Sin-Ho	Jung0005	Duke University	Co-PD/PI
Lin, Danyu	DANYU_LIN	UNC-CH	Project 4 Leader
Liu, Yufeng	NA	UNC-CH	Co-Investigator
McLeod, Howard L.	Hmcleod	UNC-CH	Co-Investigator
Owzar, Kouros	KOWZAR	Duke University	Co-Investigator
Pang, Herbert	Oxbert	Duke University	Co-Investigator
Sandler, Robert S.	ROBERT_SANDLER	UNC-CH	Co-Investigator
Socinski, Mark A.	NA	UNC-CH	Co-Investigator
Spector, Neil L.	NA	Duke University	Co-Investigator
Stefanski, Leonard A.	NA	NC State University	Co-Investigator
Tsiatis, Anastasios	butch_tsiatis	NC State University	Co-PD/PI
Tzeng, Jung-Ying	jytzeng	NC State University	Co-Investigator
Wang, Wei	wei_wang	UNC-CH	Co-Investigator
Wang, Xiofei	XIAOFEI.WANG	Duke University	Co-Investigator
Wright, Fred A.	Fred_Wright	UNC-CH	Co-Investigator
Zeng, Donglin	Donglin_Zeng	UNC-CH	Co-Investigator
Zhang, H. Helen	NA	NC State University	Co-Investigator
Zhou, Haibo	Haibo_Zhou	UNC-CH	Co-Investigator

DISTRIBUTION OF PROFESSIONAL EFFORT (%) ON THIS APPLICATION

PARTICIPATING INVESTIGATOR	PROJECT 1	PROJECT 2	PROJECT 3	PROJECT 4	PROJECT 5	CORE A	CORE B	CORE C	APPLICATION TOTAL
UNC									
Dr. Michael R. Kosorok (Principal Investigator)	5		10	10	10	15*			50
Dr. J. Todd Auman				5					5
Dr. Jianwen Cai	15*	10				5			30
Dr. William R. Carpenter			5						5
Dr. Haitao Chu			10				5		15
Dr. Jason P. Fine		10	5						15
Dr. Richard M. Goldberg	5								5
Dr. Joseph G. Ibrahim	10	5	15*			10	5		45
Dr. Danyu Lin				15*		5		10	30
Dr. Yufeng Liu				5					5
Dr. Howard L. McLeod				5					5
Dr. Robert S. Sandler			5						5
Dr. Mark A. Socinski					5				5
Dr. Wei Wang				5					5
Dr. Fred A. Wright				10					10
Dr. Donglin Zeng	5			5	5				15
Dr. Haibo Zhou	5								5
NCSU									
Dr. Marie Davidian		15*			5	10	5	15*	50

PARTICIPATING INVESTIGATOR	PROJECT 1	PROJECT 2	PROJECT 3	PROJECT 4	PROJECT 5	CORE A	CORE B	CORE C	APPLICATION TOTAL
Dr. Howard D. Bondell		10	5	5	5				25
Dr. Dennis D. Boos		5			5				10
Dr. Leonard A. Stefanski		5			5				10
Dr. Anastasios Tsiatis	5	10			15*	10	5		45
Dr. Jung-Jing Tzeng				10					10
Dr. H. Helen Zhang	5	5	5	5	5				25
DUKE									
Dr. Stephen L. George	10					10	15*		35
Dr. Kimberly L. Blackwell	5								5
Dr. Jeffrey Crawford	5								5
Dr. Philip G. Febbo				5					5
Dr. David H. Harpole				5					5
Dr. Sin-Ho Jung	10	5		10		10			35
Dr. Kouros Owzar				5			10	10	25
Dr. Herbert Pang	5			5					10
Dr. Neil L. Spector		5							5
Dr. Xiofei Wang	10								10

*Project Leader/Core Director

RESEARCH GRANT
TABLE OF CONTENTS

Section I *Page Numbers*

Section II

Individual Research Project 1

PROGRAM OVERVIEW

STATISTICAL METHODS FOR CANCER CLINICAL TRIALS

Lead PD/PI: Michael R. Kosorok, PhD
PD/PI: Marie Davidian, PhD
PD/PI: Stephen L. George, PhD

PROGRAM OVERVIEW

1 INTRODUCTION TO RESUBMISSION/REVISION APPLICATION - N/A

2 SPECIFIC AIMS

The overall scientific goal of this ambitious program project is to develop highly innovative methods for cancer clinical trials that can hasten successful introduction of effective new therapies into practice by exploiting recent advances in statistical and computational science. The key questions to be addressed are:

1. How can we leverage new statistical and computational techniques to yield major improvements in the performance of Phase II and Phase III cancer clinical trials?
2. How can we use new statistical and computational methods to solve missing data problems in clinical trials and to maximize effective use of prognostic information to improve efficiency?
3. How can we fully utilize existing clinical trial data to effectively monitor for rare adverse events and to find candidate treatment rules that can lead to improved care of cancer patients?
4. How can we use new statistical and computational techniques to incorporate and leverage genetic and genomic information in both pre-clinical studies and clinical trials?
5. How can we design new clinical trials that both discover and evaluate individualized treatment strategies that factor in past treatment experience and prognostic information to yield optimal patient outcomes?

Each of these five questions forms the basis for one of the five proposed individual, interrelated research projects. While these questions are not specific clinical hypotheses, they are questions about how to test hypotheses in the clinical trial framework. The proposed program project seeks to develop statistical and computation techniques for efficiently and validly testing many kinds of hypotheses that can arise in cancer clinical science. This includes discovering both new possibilities and clearer limitations on the kinds of hypotheses that can be evaluated using clinical data. For each question, we will use cutting edge statistical and computational techniques to create new design and analysis tools that are theoretically valid for a broad and practical range of clinical trial hypotheses. These new tools will be evaluated for practical utility in both comprehensive simulation studies and analyses of existing clinical trial data. Theoretical validity is important because this provides assurance of both the internal and external validity (reproducibility) of the new tools in a general context to all potential future studies that may or may not be comparable to the simulated or existing data used for evaluation. The tools will be modified and refined until they pass all of these tests and will then be implemented in a user-friendly manner on several widely available software platforms and disseminated to the public with accompanying training documentation. Two cores will provide services to facilitate the simulation studies, analyses, and dissemination of software and training materials. A third core will provide overall leadership and administration to coordinate and optimize the many interactions between projects and cores. Although the program project cannot address all aspects of cancer clinical trials, the proposed research is comprehensive and integrated and maximizes the combined biostatistical methods skills represented in this trans-institutional collaboration. The proposed research will advance cancer clinical trial methodology in a fundamental and paradigm-changing manner that has the potential to improve significantly the health and well-being of cancer patients.

3 BACKGROUND AND SIGNIFICANCE

3.1 Global Significance of Program

One of the major challenges of clinical research in cancer is the bottleneck between laboratory research and clinical practice. Only a very few candidate treatments make it to human clinical trials and only 10% of treatments making it to trials demonstrate enough efficacy to be approved for marketing (see Food and Drug Administration, 2004; Högberg, 2005). Given the vast resources required to conduct clinical trials, this disconnect between laboratory and clinic is a serious economic as well as scientific roadblock. This issue does not appear to be the result of limitations in basic science or technology for drug discovery. Astonishing progress in basic science in

the past few decades and new technologies such as high throughput screening have resulted in the production of $> 20,000$ new compounds annually (Högberg, 2005); however, this has not yielded a proportional growth in new drugs approved for humans. In the opinion of the Food and Drug Administration (FDA), "the applied sciences needed for medical product development have not kept pace with the tremendous advances in the basic sciences... In many cases, developers have no choice but to use the tools and concepts of the last century to assess this century's candidates" (Food and Drug Administration, 2004). This statement highlights the urgent need for innovations in methods for the design and analysis of clinical trials that address the opportunities presented by the rapidly evolving scientific landscape.

Through their significant involvement in cancer research (see Section **5.2**), much of which is with the Cancer and Leukemia Group B (CALGB), investigators on this project are well-acquainted with and inspired by both specific studies and general settings where new design and analysis approaches would have a profound impact. For example, as pharmcogenomics is increasingly incorporated into CALGB trials such as protocols 80303 (pancreatic cancer), 40101 (breast cancer), and 90401 (prostate cancer), the resulting insights on the genetic basis for drug disposition and response carry enormous potential for individualizing therapy. However, the volume, complexity, and types of data arising from emerging technologies in these studies present statistical and computational challenges that cannot be addressed with current methods. New techniques for design and analysis are critical if the promise of pharmacogenomics is to be realized. Similarly, collection of both time to death or relapse and longitudinal biomarkers such as prostate specific antigen (PSA) and measures such as quality of life (QOL) has become commonplace in CALGB trials such as protocol 90401 and 49907 (breast cancer), and characterizing the relationship between the event time and the longitudinal progression is an important objective. Statistical techniques are required that specifically address this goal in the cancer trial context. As a final example, there is growing recognition that treatment of cancer is an ongoing process involving a sequence of treatment decisions; e.g., Grossi et al. (2008) refer to the successive choices of first-, second-, and third-line therapy based on evolving patient characteristics in non-small cell lung cancer as a "treatment algorithm" and call for research to identify the best such algorithm yielding the greatest overall benefit. Study of the entire sequence of treatment decisions requires a fundamentally new approach to design and analysis; the design of CALGB studies such as protocol 19808 in acute myelogenous leukemia, which involves a series of randomizations at treatment decision points, offers a promising starting point.

The research projects that comprise this proposed program respond directly to these and other key challenges posed by this new era of cancer research. The overarching theme of the program is to effect a paradigm shift in cancer clinical trial methodology by exploiting recent advances in the statistical and computational basic sciences to yield a new generation of clinical trial tools for discovering and evaluating promising new cancer treatments. Key such advances include the genomics revolution, high throughput screening tools such as microarrays, individualized therapy and personalized medicine, dynamic treatment regimes, statistical learning methodologies such as machine and reinforcement learning, Bayesian statistical methods, methods for time-to-event data, and statistical advances in high dimensional data analysis and modeling using semiparametric methods and empirical process theory. The assembled collaborators in this program, spanning three institutions, have collectively contributed to most, if not all, of these recent advances. The importance and need for the proposed new statistical methods for cancer clinical trials in public health is tremendous and will improve the lives of patients with many kinds of cancer and will be applicable to many other diseases, including cardiovascular disease and HIV infection.

3.2 Prior Collaborations

The project investigators have a number of existing collaborations on topics related to the proposed research that have resulted in either publications or funded grants. Examples of collaborations among University of North Carolina at Chapel Hill (UNC-CH) personnel in biostatistical methodology areas related to the project include Drs. Cai and Kosorok on semiparametric methods and time-to-event data (Song et al., 2008a); Drs. Cai and Zeng on joint modeling of longitudinal and time-to-event data (Zeng and Cai, 2005); Drs. Ibrahim and Zeng on missing data (Chen et al., 2007); Drs. Chu, Ibrahim, and Sandler on statistical methodology in cancer (Qu et al., 2008); Drs. Fine and Kosorok on semiparametric methods (Lee et al., 2005) and on microarray methods (Ma et al., 2006); and Drs. Lin and Zeng on statistical genetics (Lin and Zeng, 2006). At North Carolina State University (NCSU), such collaborations include Drs. Davidian and Stefanski on longitudinal data (Huang et al., 2009), Drs. Davidian and Tsiatis on joint modeling of longitudinal and time-to-event data (Tsiatis and Davidian,

2004) and clinical trials (Tsiatis et al., 2007; Zhang et al., 2008), and numerous other joint publications. Duke University collaborations include Drs. George and Jung on clinical trial methods (Jung and George, 2009); Drs. Wang and Pang on cancer biomarkers (Wang et al., 2009); and many others.

There are also significant collaborations related to the proposed program among personnel across two or more of the three institutions, including Drs. Fine (UNC-CH) and Tsiatis (NCSU) in the areas of semiparametric methods and time-to-event data (Fine and Tsiatis, 2000), Drs. Liu (UNC-CH) and Zhang (NCSU) on statistical model selection (Zhang et al., 2008), and Drs. Zhou (UNC-CH) and Wang (Duke) in the area of semiparametric methods (Wang and Zhou, 2006). Drs. Davidian and Tsiatis (NCSU) are adjunct faculty in the Duke Department of Biostatistics and Bioinformatics and collaborate regularly with Duke faculty on ongoing research projects. Moreover, Drs. Kosorok, Cai, Fine, Ibrahim, Wright, and other investigators at UNC-CH teamed up with Drs. Davidian and Tsiatis at NCSU to formulate the Biostatistics Core of the recently funded TraCS Institute at UNC-CH, established by the NIH Clinical and Translational Science Award (CTSA, 1 UL1 RR025747-01) to UNC-CH and partners (including NCSU). Overall, there exist many strong research collaborations among the program investigators across the three institutions.

UNC-CH, NCSU, and Duke are all within a 25-mile radius and together form the Research Triangle of North Carolina. Research Triangle Park (RTP), centrally located to all three universities, is home to numerous research organizations, businesses, and institutes, several of which are collaborative ventures involving two or more of the universities, including the National Institute of Statistical Sciences (NISS). NISS shares its RTP building with the Statistical and Applied Mathematics Sciences Institute (SAMSI), a partnership between the three universities and NISS funded by the National Science Foundation that sponsors numerous research programs attracting top international researchers. SAMSI programs are a unique resource that project personnel have and will continue to exploit. For example, Drs. Davidian and Tsiatis (NCSU) and Drs. Kosorok, Lin, and Zeng (UNC-CH) participated in a 2007 SAMSI summer program on "Dynamic Treatment Regimes and Multi-stage Decision-Making," where they initiated collaborations on dynamic treatment regimes and reinforcement learning in clinical trials that form the basis for Project 5 of the proposed program project. With the National Institute of Environmental Health Sciences in RTP and statistical software company SAS Institute in nearby Cary, the region has many important resources and opportunities that facilitate statistical research collaboration.

3.3 Sequence of Events Leading to Program Application

Some of the main events, in approximately chronological order, leading up to the proposed program project are listed below. This is not an exhaustive list of all of the scientific precursors to the application, as there are far too many of these to enumerate explicitly, but rather is an overview of the key material developments:

June 2007: Drs. Davidian and Tsiatis (NCSU) and Drs. Kosorok, Lin, and Zeng (UNC-CH) participate in the above SAMSI program and initiate collaborations that formed the basis for Project 5.

July 2007: The UNC Center for Innovative Clinical Trials (CICT) in the Gillings School of Global Public Health is formed, with Dr. Ibrahim as Director and initial funding and support provided by the School as part of the Gillings Innovation Laboratories initiative. Dr. Ibrahim initiates meetings with FDA officials and others to identify current key problems in clinical trial methodology. Drs. Kosorok, Cai, Lin and others at UNC-CH, Drs. Davidian and Tsiatis (NCSU), and Drs. George and Jung (Duke) are members of the CICT. Resulting collaborations form the basis for aims in Projects 1, 2 and 3.

October 2007: Inspired by the SAMSI program, Drs. Kosorok and Zeng (UNC-CH) form the "Reinforcement Learning Group," which meets weekly to work on reinforcement learning, dynamic treatment regimes, statistical learning, and related high dimensional problem in biostatistics. Participants include Drs. Kosorok, Fine, Liu, Wang and Zeng. Research from this collaboration has yielded key components of Project 3, 4, and 5.

March 2008: Dr. Ram Tawari from the National Cancer Institute (NCI) approaches Drs. Davidian and Tsiatis (NCSU) to encourage the Research Triangle universities to collaborate on a statistical methods program project because of the strong research and NCI funding records of investigators at UNC-CH, NCSU, and Duke. Dr. Davidian discusses this with Dr. Kosorok (UNC-CH). Drs. Ibrahim and Kosorok from UNC-CH meet with Dr. George and others at Duke to discuss collaborations in clinical trials methodology.

April 2008: Drs. Davidian and Tsiatis (NCSU) and Drs. Cai, Ibrahim, Kosorok, and Lin (UNC-CH) meet at NISS to begin work on a program project in cancer clinical trials. Drs. Kosorok and Zeng (UNC-CH) attend with

Drs. Davidian and Tsiatis (NCSU) the "Atlantic Coast Symposium on the Mathematical Sciences in Biology and Biomedicine" organized by Dr. Davidian. The resulting collaboration contributed to the aims of Project 5.

May 2008: The proposed five individual research projects are identified and outlined. Dr. Eric Feuer, Chief of the NCI Statistical Research and Applications Branch, is contacted and begins advising the group. After additional discussions with NCI, the program project concept is assigned to the Cancer Therapy and Evaluation Program under the direction of Dr. Heng Xie, who becomes our primary NCI contact and adviser.

June 2008: Drs. George and Jung (Duke) are invited to join the group because of their interest and expertise in clinical trial methods as identified during the March 2008 meeting and through previously established collaborations between NCSU and Duke. The Steering Committee as currently proposed is formed. A program project wiki is established at NCSU and is used for organizing research activities. This wiki is the prototype for the proposed program wiki to be housed at UNC-CH.

July–October 2008: Extensive research and planning meetings of the Steering Committee occur to formulate a more complete program project proposal with its aims fleshed out as well as the three proposed cores. Writing is initiated, and additional investigators are identified. Arrangements are made to present the program project plans to NCI for feedback.

November 2008: The program project concept is presented to NCI on November 10, 2008. Constructive feedback and positive encouragement is provided by Dr. Xie and his associates at NCI. Writing continues.

December 2008: The letter-of-intent is submitted and approved at NCI. Writing continues. Drs. Kosorok, Davidian, and George obtain significant institutional commitments of support for the project.

January 2008: The grant proposal is completed and submitted to NIH for review.

3.4 Advantages of a Group Effort

While most of the investigators have common interests in clinical trial methodology, there is tremendous diversity in research areas, modes of problem solving, and resources among this group of scientists at UNC-CH, NCSU, and Duke. This simultaneous unity of purpose and diversity of skills, along with our geographic proximity, make us uniquely qualified to solve the extremely challenging technical and applied problems identified in the proposed research and thereby to make a significant contribution to clinical trial methodology. The extensive practical scientific expertise at Duke through experience with the CALGB brings valuable focus and wisdom to the program. The extensive experience in model selection, dynamic treatment regimes, and related areas of applied statistics and biostatistics at NCSU is crucial to the success of the program. The extensive expertise at UNC-CH in clinical trial methodology, time-to-event analysis, empirical processes, and genomics also plays a critical role. These topics represent only a small subset of the many areas of biostatistical expertise represented on this project. Each individual and each institution bring something to this program that is important to the program's success. The combined strength of this group effort is essential to achieve high-impact, paradigm-changing advances—rather than only incremental improvements—in statistical methods for cancer clinical trials.

4 LIST OF PROJECT AND CORE COMPONENTS

The following are the five individual research project titles and leadership proposed in this program project:

Project 1: Innovative Clinical Trial Design and Analysis. Project Leader: Dr. Cai (UNC-CH). Project co-Leaders: Drs. George (Duke) and Ibrahim (UNC-CH). Project co-Investigators: Drs. Blackwell (Duke), Crawford (Duke), Goldberg (UNC-CH), Jung (Duke), Kosorok (UNC-CH), Pang (Duke), Tsiatis (NCSU), Wang (Duke), Zeng (UNC-CH), Zhang (NCSU), and Zhou (UNC-CH).

Project 2: Methods for Missing and Auxiliary Data in Clinical Trials. Project Leader: Dr. Davidian (NCSU). Project co-Leaders: Drs. Ibrahim (UNC-CH) and Tsiatis (NCSU). Project co-Investigators: Drs. Bondell (NCSU), Boos (NCSU), Cai (UNC-CH), Fine (UNC-CH), Jung (Duke), Spector (Duke), Stefanski (NCSU), and Zhang (NCSU).

Project 3: Methods for Post Marketing Surveillance and Comparative Effectiveness Research. Project Leader: Dr. Ibrahim (UNC-CH). Project co-Leaders: Drs. Chu (UNC-CH), Kosorok (UNC-CH) and Zhang (NCSU). Project co-Investigators: Drs. Bondell (NCSU), Carpenter (UNC-CH), Fine (UNC-CH), and Sandler (UNC-CH).

Project 4: Methods for Pharmacogenomics and Individualized Therapy Trials. Project Leader: Dr. Lin (UNC-CH). Project co-Leaders: Drs. Jung (Duke), Kosorok (UNC-CH) and Owzar (Duke). Project co-Investigators: Drs. Auman (UNC-CH), Bondell (NCSU), Febbo (Duke), Harpole (Duke), Liu (UNC-CH), McLeod (UNC-CH), Pang (Duke), Tzeng (NCSU), Wang (UNC-CH), Wright (UNC-CH), Zeng (UNC-CH), and Zhang (NCSU).

Project 5: Methods for Discovery and Analysis of Dynamic Treatment Regimes. Project Leader: Dr. Tsiatis (NCSU). Project co-Leaders: Drs. Davidian (NCSU) and Kosorok (UNC-CH). Project co-Investigators: Drs. Bondell (NCSU), Boos (NCSU), Socinski (UNC-CH), Stefanski (NCSU), Zeng (UNC-CH), and Zhang (NCSU).

These projects are supported by three cores:

Core A: Administrative Core. Core Director: Dr. Kosorok (UNC-CH). Core co-Directors: Drs. Davidian (NCSU), George (Duke), Ibrahim (UNC-CH), Jung (Duke) and Tsiatis (NCSU). Core Contributors: Drs. Cai (UNC-CH), Lin (UNC-CH), and Owzar (Duke).

Core B: Data Compilation Core. Core Director: Dr. George (Duke). Core co-Directors: Drs. Cai (UNC-CH), Chu (UNC-CH), Davidian (NCSU), Owzar (Duke), and Tsiatis (NCSU).

Core C: Computational Resource and Dissemination Core. Core Director: Dr. Davidian (NCSU). Core co-Directors: Drs. Lin (UNC-CH) and Owzar (Duke).

5 PRELIMINARY STUDIES

5.1 Qualification of Investigators

Michael R. Kosorok, PhD, lead PD/PI, Core A Director, co-Leader for Projects 3, 4 and 5, and co-Investigator for Project 1. Dr. Kosorok is Professor and Chair of Biostatistics and Professor of Statistics and Operations Research at UNC-CH. He is also Director of the Biostatistics Core of the UNC-CH TraCS Institute (CTSA). His expertise is in clinical trials, survival analysis, microarrays, statistical learning, empirical processes, and semiparametric inference, and he has written a text on the last two topics (Kosorok, 2008). He is an elected Fellow of the American Statistical Association (ASA) and the Institute of Mathematical Statistics (IMS).

Marie Davidian, PhD, PD/PI, Project 2 Leader, Core C Director, Project 5 co-Leader, and co-Director for Cores A and B. Dr. Davidian is William Neal Reynolds Professor in the Department of Statistics and Director of the Center for Quantitative Sciences in Biomedicine at NCSU and Adjunct Professor of Biostatistics and Bioinformatics at Duke. She also serves as Executive Editor of *Biometrics*, regarded by many to be the top journal in the field of biostatistics. Her expertise is in longitudinal data, missing data, biomedical modeling, clinical trials, and semiparametric methods. She is an elected Fellow of the ASA and IMS.

Stephen L. George, PhD, PD/PI, Core B Director, Project 1 co-Leader, and Core A co-Director. Dr. George is Professor of Biostatistics and Bioinformatics at Duke and Director of Biostatistics for the Duke Comprehensive Cancer Center (DCCC) and for the CALGB. His expertise is in clinical trials, translational science, and prognostic and predictive models. He is an elected Fellow of the ASA.

Joseph G. Ibrahim, PhD, co-PD/PI, Project 3 Leader, co-Leader for Projects 1 and 2, and co-Director for Cores A and B. Dr. Ibrahim is Alumni Distinguished Professor of Biostatistics at UNC-CH, the Lineberger Comprehensive Cancer Center (LCCC) Director of Biostatistics, and the Director of the UNC CICT. His expertise is in Bayesian methods, missing data, clinical trials, and cancer genomics. He has published two texts on Bayesian methods (Ibrahim et al., 2001; Chen et al., 2008). He is an elected Fellow of the ASA and IMS.

Sing-Ho Jung, PhD, co-PD/PI, co-leader for Projects 1, 2, and 4, and co-Director for Core A. Dr. Jung is Professor of Biostatistics and Bioinformatics at Duke. Dr. Jung serves as the Director of the CALGB Biostatistics unit. His expertise is in survival analysis, various types of clustered and longitudinal data analysis, design and analysis methods for Phase II cancer clinical trials, microarrays, and proteomics.

Anastasios A. Tsiatis, PhD, co-PD/PI, Project 5 Leader, Project 2 co-Leader, co-Director for Cores A and B, and co-Investigator for Project 1. Dr. Tsiatis is Drexel Professor of Statistics at NCSU and Adjunct Professor of Biostatistics and Bioinformatics at Duke. His expertise is in survival analysis, causal inference, clinical trials, and semiparametric methods, and is author of a text on semiparametric methods (Tsiatis, 2006). He is

an elected Fellow of the ASA and IMS.

Jianwen Cai, PhD, Project 1 Leader, Project 2 co-Investigator, and Core A Contributor. Dr. Cai is Professor and Associate Chair of Biostatistics at UNC-CH. Her expertise is in clinical trials, survival analysis, and semiparametric methods. She is an elected Fellow of the ASA.

Danyu Lin, PhD, Project 4 Leader, Core C co-Director, and Core A Contributor. Dr. Lin is Dennis Gillings Distinguished Professor of Biostatistics at UNC-CH. His expertise is in clinical trials, survival analysis, genomics, and semiparametric methods. He is an elected Fellow of the ASA and IMS.

The above investigators constitute the Steering Committee, which provides overall scientific leadership. Table 1 gives the complete list of program project investigators with percent effort by institution, project and core. The statistical methodology experts in addition to the Steering Committee members are Drs. Chu, Fine, Liu, Wright, Zeng, and Zhou (UNC-CH); Drs. Bondell, Boos, Stefanski, Tzeng, and Zhang (NCSU); and Drs. Owzar, Pang, and X. Wang (Duke). Dr. W. Wang (UNC-CH) is a computer scientist, Dr. Carpenter (UNC-CH) is a health policy and management expert, and Drs. McLeod and Auman (UNC-CH) are pharmacologists. Investigators with cancer clinical and translational expertise include Drs. Goldberg, Sandler, and Socinski (UNC-CH) and Drs. Blackwell, Crawford, Febbo, Harpole and Spector (Duke). Details on the qualifications of these investigators are presented elsewhere in the project and core narratives.

5.2 Cancer and Clinical Trial Experience and Collaborations

The investigators at UNC-CH have considerable experience in clinical trials, both in cancer and in other diseases. With over 15 years of experience in cancer clinical trials, Dr. Ibrahim is currently Biostatistics Core Director at the LCCC at UNC-CH, Biostatistical Core Leader of the GI SPORE grant and the melanoma program project at the LCCC, and co-director of the Biostatistics Core of the UNC-CH Breast SPORE grant. Dr. Ibrahim has been heavily involved in the design and analysis of Phase I and Phase II clinical trials at LCCC. Previously, He worked for 8 years at the Dana-Farber Cancer Institute (DFCI) and with the Eastern Cooperative Oncology Group (ECOG) and was the senior and lead statistician on the Gastrointestinal Committee and the Melanoma Committee in ECOG. Dr. Ibrahim was instrumentally involved in the design and analysis of the pivotal melanoma Phase III clinical trials E1684, E1690, E1697, and E1694, which ultimately led to FDA approval for the use of high-dose interferon as the standard treatment in high-risk melanoma. Drs. Kosorok, Cai, Chu, and Fine also have extensive clinical trials methods experience in cancer and in other diseases. Dr. Kosorok was chair of the Data Safety Monitoring Committee for the intramural program of the National Institute of Child Health and Human Development from 2001-2006.

Dr. George (Duke) has nearly 40 years experience in cancer clinical trials, including 7 years in the Department of Bioinformatics at the MD Anderson Cancer Institute; 1 year at the European Organization for Research on the Treatment of Cancer (EORTC) in Brussels, where he established and was first Director of the EORTC Data Center; 12 years as Director of Biostatistics at St. Jude Children's Research Hospital; and 20 years as Director of Biostatistics at the DCCC, the last 18 years also as Group Statistician and Director of the Statistical Center for the CALGB. He has served on several advisory committees and study sections for the NCI, on numerous Data and Safety Monitoring Boards (DSMBs) for government and industry sponsored cancer clinical trials, and for 4 years on the FDA Oncology Drug Advisory Committee (ODAC). Dr. George has also been active in the Society for Clinical Trials, including serving a term as President. Dr Jung (Duke) has worked at NCI-designated cancer centers at Mayo Clinic, Duke, and Indiana University, serving as Director of the Biostatistics Core at the latter. He has also worked in the North Central Cancer Treatment Group; the American College of Surgeons Oncology Group, for which he was acting group statistician; and CALGB, for which he is currently Director of the Biostatistics Unit as well as faculty statistician for the Lymphoma and Imaging Committees and the Cancer Prevention Subcommittee. Dr. Owzar is Director of the CALGB Bioinformatics Unit, faculty statistician for Pharmacology and Experimental Therapeutics and Transplant committees and represents the Statistical Center on the CALGB Correlative Science Advisory Committee. He is also involved in design and analysis of pharmacogenomics studies in cancer and is a member of the Pharmacogenetics of Anticancer Agents Research Group of the Pharmacogenetics Research Network. Dr. Wang is faculty statistician for the Respiratory Committee and the Cancer Control and Health Outcome Committee in CALGB, and Dr. Pang is faculty statistician on the Respiratory and Oncology Nursing committees in CALGB and currently serves as a statistical reviewer

Table 1: Percent effort of all investigators by institution, project and core: * denotes Steering Committee members and project and core leadership, † denotes PD/PIs and co-PD/PIs, and ‡ denotes Executive Committee members.

Institution	Name	Project					Core			Total
		1	2	3	4	5	A	B	C	
UNC-CH	Dr. Michael R. Kosorok*†‡	5		10	10	10	15*			50
	Dr. J. Todd Auman				5					5
	Dr. Jianwen Cai*	15*	10				5			30
	Dr. William R. Carpenter			5						5
	Dr. Haitao Chu			10				5		15
	Dr. Jason P. Fine		10	5						15
	Dr. Richard M. Goldberg	5								5
	Dr. Joseph G. Ibrahim*†	10	5	15*			10	5		45
	Dr. Danyu Lin*				15*		5		10	30
	Dr. Yufeng Liu				5					5
	Dr. Howard L. McLeod				5					5
	Dr. Robert S. Sandler			5						5
	Dr. Mark A. Socinski					5				5
	Dr. Wei Wang				5					5
	Dr. Fred A. Wright				10					10
	Dr. Donglin Zeng	5			5	5				15
	Dr. Haibo Zhou	5								5
	UNC-CH Total	45	25	50	60	20	35	10	10	255
NCSU	Dr. Marie Davidian*†‡		15*			5	10	5	15*	50
	Dr. Howard D. Bondell		10	5	5	5				25
	Dr. Dennis D. Boos		5			5				10
	Dr. Leonard A. Stefanski		5			5				10
	Dr. Anastasios A. Tsiatis*†	5	10			15*	10	5		45
	Dr. Jung-Ying Tzeng				10					10
	Dr. H. Helen Zhang	5	5	5	5	5				25
	NCSU Total	10	50	10	20	40	20	10	15	175
Duke	Dr. Stephen L. George*†‡	10					10	15*		35
	Dr. Kimberly L. Blackwell	5								5
	Dr. Jeffrey Crawford	5								5
	Dr. Philip G. Febbo				5					5
	Dr. David H. Harpole				5					5
	Dr. Sin-Ho Jung*†	10	5		10		10			35
	Dr. Kouros Owzar				5			10	10	25
	Dr. Herbert Pang	5			5					10
	Dr. Neil L. Spector		5							5
	Dr. Xiaofei Wang	10								10
	Duke Total	45	10		30		20	25	10	140
	Grand Total	100	85	60	110	60	75	45	35	570

for the Cancer Protocol Committee at Duke.

Dr. Tsiatis (NCSU) has extensive experience in cancer clinical trials and cancer research. From 1979-1981, he worked with Dr. George at St. Jude Children's Research Hospital on the design and analysis of childhood leukemia trials. Dr. Tsiatis was on the faculty at DFCI from 1981-1990, where he collaborated on design and analysis of trials in lymphoma and multiple myeloma; in addition, he was affiliated with ECOG, for which he served as Coordinating Statistician from 1982-1984. Dr. Tsiatis was the primary statistical consultant for the World Health Organization Breast Self Examination Study, a group randomized trial conducted in the former Soviet Union. From 1999-2004, he was a member of the DSMB for the CALGB. Dr. Tsiatis also has considerable experience in clinical trials in HIV infection, cardiovascular disease, and diabetes, and has served on numerous DSMBs. Dr. Davidian served as Senior Statistician on numerous HIV clinical trials through her affiliation with the AIDS Clinical Trials Group in 1994-1996, she currently collaborates regularly on design and analysis of cardiovascular disease trials at Duke Clinical Research Institute (DCRI).

5.3 Preliminary Results

Project 1: Innovative Clinical Trial Design and Analysis. The team of investigators has made extensive progress in developing statistical methods for novel clinical trial and analysis issues that provides an ideal starting point for developing the proposed new methodology. This includes development of joint models of longitudinal and time-to-event data (Brown and Ibrahim, 2003a, 2003b, 2005; Ibrahim et al., 2004; Tsiatis and Davidian, 2004; Zeng and Cai, 2005; Chi and Ibrahim, 2006, 2007) and development of methods for study designs with case-cohort sampling or correlated outcomes for time-to-event data. See, for example, Cai and Zeng (2004, 2007), who developed sample size and power methodology for a case-cohort design; Wang and Zhou (2006), who developed methods of using prognostic data to improve designs of clinical studies; and Jung and Jeong (2003), who studied the weighted rank test under cluster randomization and Jeong and Jung (2006) under subunit randomization. Gangnon and Kosorok (2004) proposed closed form sample size formulæ for general clustered data. See also Song et al. (2008a). Other related preliminary work by our team includes methods for comparing multiple arms obtained from multistage randomized Phase II trials (Jung and George, 2008) and a new semiparametric method for covariate adjustment that separates modeling of covariate relationships from estimation of the treatment effect (Tsiatis et al., 2008).

Project 2: Methods for Missing and Auxiliary Data in Clinical Trials. In Zhang et al. (2008), Drs. Davidian and Tsiatis outline a broad framework for covariate adjustment in clinical trials for general types of outcomes and two or more treatments based on taking a semiparametric perspective that places minimal restrictions on the nature of the data. A general strategy for deriving estimators for parameters of interest when data are missing due to censoring using the theory of semiparametrics has been outlined in Sections 9.3 and 10.4 of Tsiatis (2006). Dr. Ibrahim and co-workers have developed case-deletion measures for assessing the influence of several observations for a variety of statistical models for missing data (Cho et al., 2009). Dr. Fine has proposed sensitivity analysis methods in the special case of longitudinal analysis of a binary outcome, as in E1684, where there is potentially informative drop-out (Todem and Fine, 2008). These innovative results, along with the other preliminary results described in more detail in Project 2, are collectively a very advantageous starting point for the proposed research of Project 2.

Project 3: Methods for Post Marketing Surveillance and Comparative Effectiveness Research. Dr. Ibrahim and colleagues have developed powerful Markov chain Monte Carlo approaches for models with latent effects such as those arising in meta-analyses of clinical trials (Chen et al., 2006). Investigators have also made preliminary progress on developing statistical methods for diagnostic accuracy studies in meta-analysis using random effects models in the presence of a gold standard (Chu and Cole 2006; Chu and Guo 2009). Dr. Kosorok developed two new technical tools in empirical processes that pave the way for developing methods of inference for semiparametric techniques for rare time-to-event data: a more flexible central limit theorem (see Theorems 11.16 and 11.18 in Kosorok, 2008) and a novel compound Poisson process methodology (Kosorok and Song, 2007; Song et al., 2008b). These preliminary results, and others described in more detail in Project 3, will greatly facilitate progress on the proposed research.

Project 4: Methods for Pharmacogenomics and Individualized Therapy Trials. The investigators have been at the forefront of developing statistical methods to detect haplotype-disease associations in cross-sectional, case-control, and cohort studies (e.g., Lin et al., 2005; Lin and Zeng, 2006; Zeng et al., 2006) and to analyze

untyped SNPs in case-control studies (Lin et al., 2008). Dr. Lin's software interface HAPSTAT and SNPMStat have been downloaded and utilized by many researchers. The team also has considerable experience in variable selection and model building in various contexts. Lu and Zhang (2007), Zhang and Lu (2007), and Johnson et al. (2008) studied moderate-scale variable selection under semiparametric regression models for censored event times data, while Zhang et al. (2008) proposed a new shrinkage method based on the supnorm penalty for variable selection in multiclass support vector machines for high dimensional data. Dr. Wright has developed a novel approach to handling significance in high dimensional data, recently described in the context of genetic association testing (Ghosh et al., 2008), that is generally applicable. Although the approach shares some similarities with an approach simultaneously published by Zhong and Prentice (2008), simulations showed that a standard confidence interval procedure of Zhong and Prentice has incorrect confidence coverage in some settings, while the proposed approach appears to be uniformly correct.

Project 5: Methods for Discovery and Analysis of Dynamic Treatment Regimes.

Drs. Davidian and Tsiatis have developed methods for estimation of mean outcomes for dynamic treatment regimes from sequentially randomized studies (Lunceford, Davidian, and, and Tsiatis, 2002; Wahed and Tsiatis, 2004, 2006), and have developed other methods for inference on dynamic treatment regimes (Johnson and Tsiatis, 2004, 2005; Zhang et al., 2009). Drs. Kosorok and Zeng have undertaken a preliminary assessment of the proposed reinforcement learning methods for discovering optimal dynamic treatment regimes using a simple computer model for a generic cancer (Zhao et al., 2008) based on a simple difference equation that balances a chemotherapeutic agent's efficacy and toxicity. Based on data from 1000 patients simulated from the model in a clinical trial of the type proposed in Project 5, they used the methods to determine the optimal timing and dosing of the agent over a 6 month period. Figure 1 presents the disease severity of 200 new simulated patients receiving the optimal treatment rule (dark solid line) compared to them receiving 10 different fixed dose regimens (dotted lines), showing how the optimal rule determined by the proposed methods is clearly superior after 6 months of treatment. These results demonstrate the considerable promise of the methods proposed in Project 5.

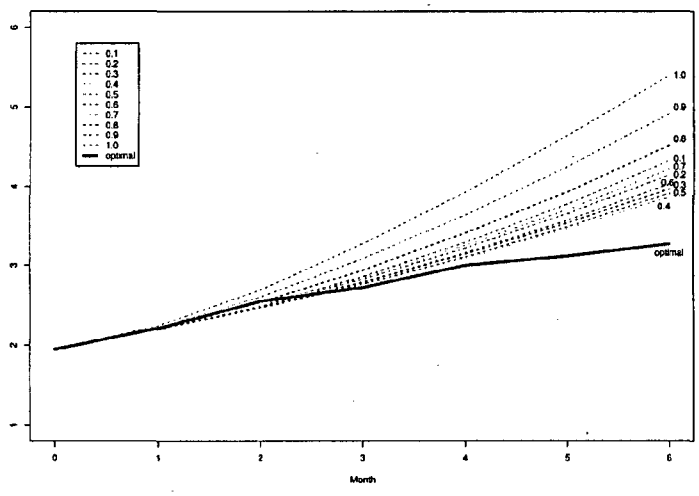


Figure 1: Disease severity (lower is better) as a function of time (in months) for the optimal treatment based on reinforcement learning (dark solid line) versus fixed dose options (dotted lines).

6 OVERALL RESEARCH DESIGN AND METHODS

The overall research strategy is to discover, create, evaluate, validate, and disseminate new statistical methodology to solve important design and analysis problems related to cancer clinical trials in order to significantly advance cancer public health. Our group of scientists from UNC-CH, NCSU, and Duke is uniquely experienced and prepared to take on this challenge. The discovery and creation stages of the proposed research will be accomplished through individual and collaborative problem solving and sharing of ideas. This aspect is a natural extension of our strong track record in developing novel statistical methods to solve important health science problems, combined with our close attention to cancer clinical trial practice and experience. The discovery and creation stages will both feed into and benefit from the evaluation and validation stages in order to produce new statistical design and analysis tools that work in practice. The evaluation and validation stages of our research will involve theoretical assessment, simulation studies, and analyses of existing data. These three phases are of necessity inter-dependent, because discoveries made in one phase will influence and illuminate lines of attack for another phase. The theoretical assessment phase is extremely important, as this provides

insight into internal and external validity of the proposed methodology that transcends any conclusions obtainable from simulation studies or data analyses. The reason for this is that theoretical results can provide insight into how a methodology will perform in broad generality for future clinical data sets, trials, and analyses, not just for observed data. In contrast, simulation and data based evaluations can only provide insight into performance for an observed data set or an observed collection of data sets. Nevertheless, simulation studies and analyses of existing data are also crucially important, because they both calibrate the theoretical results and provide a critical reality check. They also provide motivation and direction in methodological development, as mentioned above, as well as practical insights that can prevent spending too much time on false leads. We wish to emphasize, however, that the collective experience and ability of our research team to provide theoretical rigor in combination with simulation and data evaluations is uniquely strong and is a key feature of our proposal. The fact that we have researchers spanning theoretical, computational, and clinical perspectives enables the strongest possible research thrust and greatest likelihood of success in achieving the proposed research goals.

The dissemination stage consists of developing, refining, and testing software implementing the new design and analysis tools and creating tutorials and methods of dissemination using our program project web site. We are committed to putting our new technology into the hands of practitioners to improve public health.

We next present a description of several cross-cutting biostatistical research themes that run through at least two or more of the proposed projects. Part of our overall research strategy is to collaborate across projects on these themes for increased efficiency; our critical mass of expertise across the three combined institutions on these themes is a powerful and unique strength of the project. We will then present brief summaries of the methods for the individual research projects and cores, including descriptions of opportunities for intra-program collaboration and relevant cross-cutting research themes, as well as core support applicable to that project or core.

6.1 Cross-Cutting Statistical Research Themes

Statistical Learning. Statistical learning methods, including machine learning and reinforcement learning, are powerful data mining methods for classification and regression that can be used to develop high dimensional rules for making decisions about, for example, what treatment to give what patient. Statistical learning methods play an important role in Aim 3.ii of Project 1, Aim 5 of Project 3, Aims 2 and 4 of Project 4, and all Aims of Project 5. Investigators with expertise in statistical learning include Drs. Kosorok, Lin, Wright, Liu, and W. Wang (see the book by Wang and Yang, 2005) of UNC-CH; Drs. Bondell, Davidian, Tsiatis, Tzeng, and Zhang at NCSU; and Drs. Jung, Owzar, Pang, and X. Wang at Duke.

Model and Variable Selection. Model and variable selection is a challenging but ubiquitous concern in many areas of biostatistics. The consequences of ignoring the model or variable selection process in statistical inference can lead to serious biases and misleading conclusions. This general area plays an important role in at least two aims or sub-aims in all of the research projects. Investigators with expertise in model and variable selection include Drs. Cai, Ibrahim, and Fine at UNC-CH; Drs. Bondell, Boos (see the book by Boos and Stefanski, 2006), Stefanski, and Zhang at NCSU; and Drs. Jung and X. Wang at Duke.

Statistical Genetics and Genomics. This is an extremely exciting and dynamic area and is especially challenging because of the very high dimensional data involved. The general area plays an important role in Aim 1 of Project 2 and in two or more aims in Projects 3, 4 and 5. Investigators with expertise in Statistical Genetics and Genomics include Drs. Kosorok, Lin, Wright, and Zeng at UNC-CH; Dr. Tzeng at NCSU; and Drs. Owzar and Pang at Duke.

Semiparametric Inference and Empirical Processes. Semiparametric models are statistical models with both a parametric component for scientific interpretability and a nonparametric component for robustness to biological complexities such as can arise in medical data. Empirical processes are high dimensional statistical summaries of data associated with many important and complex statistical settings, including inference for semiparametric models. This general research area plays fundamental roles in almost all of the aims in the program project and is arguably the most ubiquitous cross-cutting theme. Investigators with expertise in semiparametric inference and empirical processes include Drs. Cai, Fine, Ibrahim, Kosorok (see the book by Kosorok, 2008), Lin, Liu, Zeng, and Zhou at UNC-CH; Drs. Davidian, Tsiatis (see the book by Tsiatis, 2006), and Zhang at UNC-CH; and Drs. Jung and X. Wang at Duke.

Bayesian Methods. Bayesian methodology is an important and fundamental branch of statistics that contributes broadly to many medical research areas and is beginning to have an impact in cancer clinical trials because of its ability to handle complicated data models. This general research area plays fundamental roles in Aim 4 of Project 2, Aims 1–3 of both Projects 3 and 4, and Aim 1 of Project 5. Investigators with expertise in Bayesian Methods include Drs. Chu, Ibrahim (see the books by Ibrahim et al., 2001; Chen et al., 2008), and Lin at UNC-CH; Dr. Tzeng at NCSU; and Drs. Owzar and Pang at Duke.

6.2 Individual Research Project and Core Methods

Project 1: Innovative Clinical Trial Design and Analysis. In this project, we will develop new statistical methodology to address issues in the design and analysis of clinical trials and investigate their analytical and empirical behavior. Related software will be developed. We have three specific aims:

1. Develop methods for design and sample size calculation for longitudinal and joint models for longitudinal and survival data. We will develop methods for design issues, such as sample size and power considerations, for investigating treatment effect on both time-to-event and longitudinal processes and the effect of a longitudinal process on time-to-event. For example, in many cancer clinical trials, both time to death (or relapse) and longitudinal QOL measures are collected. If treatment has an effect on the longitudinal process, and the longitudinal process has an effect on the survival time, then the longitudinal biomarker is in the casual pathway and can potentially be used as a surrogate endpoint for the death time. We will consider settings ranging from a single univariate longitudinal process and a univariate time-to-event process to the complex case of multivariate longitudinal and time-to-event processes.

2. Develop statistical methodology for the design and analysis of group randomized cancer prevention trials with survival and recurrent event outcomes. We will develop sample size and power calculations for treatment effect on time-to-event and recurrent event outcomes in a group randomized trial setting.

3. Develop statistical methodology for cancer drug development. We will address important statistical issues in the oncology drug development pathway, including three specific sub-aims: (i) Develop methods for the design and analysis of clinical trials of targeted therapy. We will develop new targeted therapy clinical trial designs and analysis methods, including "enrichment" designs in which some, but not all, of the patients without the target are randomized, and compare the operating characteristics and costs of these designs to fully targeted designs. (ii) Develop designs for Phase II trials that are predictive of Phase III trial success. We will develop new methods for Phase II trials, particularly randomized Phase II trials, and Phase II/III clinical trials, and assess their operating characteristics, costs, and predictive ability for subsequent phase III trials. We will also gather information on both combination and non-combination therapies in Phase II studies and subsequent Phase III studies to build prediction models using machine learning and other nonparametric classification methods. (iii) Develop methods for the design and analysis of partially randomized clinical trials. We will develop new semiparametric empirical likelihood methods for the analysis of such trials to adjust for selection bias and to improve efficiency. We will also work closely with collaborators on Aim 1 of Project 2 on an alternative approach.

Intra-Program Collaboration. Because missing data are routine in clinical trials, the proposed research for Project 1 will benefit from the missing data methodology developed in Project 2. The design results of Project 1 will be useful for developing trials for evaluating candidate treatments obtained from the methodology developed in Projects 4 and 5. Project 1 will need Core C to assist with development of efficient code for simulation studies. Core B will be needed to provide the existing data sets for evaluation. Core C will prepare and disseminate the software that implements the new methods. Cross-cutting research themes include Statistical Learning for Aim 3.ii; Model and Variable Selection for both Aims 1 and 3.ii; and Semiparametric Inference and Empirical Processes for Aims 2, 3.ii and 3.iii. Core A will facilitate the various modes of interaction needed between Project 1 and the other projects and cores.

Project 2: Methods for Missing and Auxiliary Data in Clinical Trials. We will develop statistical methodologies to exploit prognostic auxiliary information and to provide frameworks for analyses in the presence of missing data that will affect notably the strength and impact of inferences possible from current cancer clinical trials. Our approach both acknowledges and takes advantage of the recent advent of novel biomarkers and emerging genomic technologies that may yield important new baseline predictors of primary clinical outcomes, the increasing emphasis on analyses of longitudinal progression of markers such as measures of QOL, recent

advances in semiparametric methods, and the routine complications of missing information and subject drop-out. We have the following four aims:

1. Develop methods to improve efficiency of inferences in randomized cancer clinical trials using auxiliary covariates. Although auxiliary baseline information is routinely collected on trial participants in addition to clinical endpoints, "adjusting" for auxiliary covariates has engendered considerable controversy because of the temptation under the usual regression approach to inspect different model fits and choose that leading to the most dramatic estimated treatment effect, resulting in potentially misleading conclusions. We will address this difficulty by developing new approaches based on state-of-the art semiparametric methods that properly incorporate model uncertainty in this setting to obtain correct inferences, including for the case where key auxiliary information is missing for some subjects. We will also extend the methods to analysis of the more complex partially randomized trial design studied in Aim 3 of Project 1.

2. Develop methods for primary and longitudinal analyses in the presence of drop-out. A routine feature of cancer trials is drop-out, where subjects are lost prior to the end of follow-up, so that data intended to be collected are missing subsequent to the time of drop-out. We will extend the methods in Aim 1 to this setting. Many cancer trials also involve analyses of longitudinal measures such as QOL and PSA, which are complicated by drop-out. A promising, recent development in semiparametric methods is the concept of "doubly robust" methods, which uses models for both the longitudinal data and the drop-out mechanism but requires only one of the two models to be correct. We will utilize this new technology to develop both efficient and robust methods for longitudinal analysis under these conditions.

3. Develop diagnostic measures for joint models for longitudinal and survival data in the presence of nonignorably missing data. Cancer trials may involve studies of the association between longitudinal markers and clinical outcomes such as relapse-free survival or death, and a popular framework for analysis is that of "joint models" for the longitudinal data and time-to-event outcome, also studied in Aim 1 of Project 1. Because of their complexity, however, these frameworks rely heavily on the correctness of models for the full data. In order to help data analysts facing this issue, we will develop new, previously unavailable diagnostic techniques for these models when there may be nonignorable missing outcome and/or covariate data.

4. Develop inference methods for sensitivity analyses of missing data. A major challenge when intended data are missing is that it is impossible to evaluate whether or not the missing at random assumption is justified based on the observed data, and models for nonignorable missingness mechanisms cannot be entirely verified based on observed data. Because analyses may be predicated on such models, misleading inferences may result. We will develop rigorous inferential methods that formally acknowledge this non-identifiability of the missingness model as well as the need to explore a range of plausible models to gauge sensitivity of inferences.

Intra-Program Collaboration. Because missing and auxiliary data arise in many clinical trial contexts, the results of this project, especially the general foundational approach developed in Aim 4, will impact and be of benefit to all of the other projects. Aims 2 and 3 of Project 2 will directly benefit Aim 1 of Project 1. As mentioned above, Aim 1 of Project 2 will be directly useful in the trial methodology studied in Aim 3 of Project 1. Project 2 will need Core C to assist with efficient code for the extensive simulation studies. Core B will provide existing data sets for calibration and evaluation of the new methods. Core C will also be needed to prepare and disseminate the software implementing the new methods. Cross-cutting research themes are Model and Variable Selection for all of the aims; Statistical Genetics and Genomics for Aim 1; Semiparametric Inference and Empirical Processes for Aims 1, 2 and 3; and Bayesian methods for Aim 3. Core A will facilitate the various modes of interaction needed between Project 2 and the other projects and cores.

Project 3: Methods for Post Marketing Surveillance and Comparative Effectiveness Research. We will develop, test, and evaluate new statistical methodology for Bayesian meta-analysis of cancer clinical trials; design, sample size, and power approaches for future studies using meta-analytic models; meta-analysis of diagnostic tests; regression analysis of rare adverse events; and identifying optimal individualized therapies. We will pursue the following five specific aims:

1. Develop methodology for Bayesian meta-analysis of cancer clinical trials. We will develop Bayesian parametric and semiparametric models for meta-analysis for aggregated data, time-to-event data, discrete data, and longitudinal data. To this end, we will consider: 1) Normal random effects models and a novel Bayesian derivation of the Q function for assessing heterogeneity across different studies for aggregated data; 2) Ran-

dom effects generalized linear models for continuous or discrete data; 3) Mixed effects models for longitudinal data; and 4) Random effects Cox models with gamma process priors for time-to-event data. We will incorporate missing covariates and/or responses in all these models for various data types.

2. Develop methodology for Bayesian trial design using meta-analytic models. We will develop a new Bayesian approach of sample size determination (SSD) for design of non-inferiority clinical trials using the models developed in Aim 1. First, we will extend the fitting and sampling priors of Wang and Gelfand (2002) to Bayesian SSD using meta-analytic models with a focus on controlling type I error, type II error, and power. Secondly, we will develop simulation-based Bayesian SSD using meta-analytic random effects generalized linear and linear mixed models, and random effects Cox models with gamma process priors.

3. Develop meta-analytic methodology for diagnostic tests without a gold standard. We will first develop statistical methods for estimating accuracies of two and multiple (i.e., $\geq$3) diagnostic tests in a meta-analysis in the absence of a gold standard using maximum likelihood and full Bayesian methods. We will then reanalyze the meta-analysis data of 17 studies to evaluate the accuracy of microsatellite instability testing (MSI) and mutation analysis (Chen et al., 2005), and a multi-center data set from NCI Colorectal Cancer Family Registry Study to evaluate the accuracy of 10 biomarkers in predicting Lynch syndrome as well as other data sets.

4. Develop methodology for regression analysis of rare adverse events for post-marketing safety evaluation. First, we will develop semiparametric methods of inference for evaluating drug and risk factor effects for rare time-to-event outcomes in cancer clinical trials and cancer epidemiological studies. Secondly, we will develop semiparametric methods of inference for extremely rare time-to-event outcomes. Thirdly, we will extend both of the results to the adjudicated endpoint setting. Fourthly, we will extend these results to the meta-analytic setting involving collections of clinical studies, registry data and health insurance claims data.

5. Develop methodology for identifying optimal individualized therapies from existing clinical trial data using meta-analysis, utility functions, classification and regression. We will develop a general inferential tool for determining optimal individualized therapies for cancer based on meta-analysis of cancer clinical trials. The approach involves a novel multi-attribute utility function for accommodating complex time-to-event information, as well as cost and quality of life considerations. We will develop rigorous inferential procedures for finding optimal treatments as a function of genomic as well as other prognostic factors. To achieve this, we will use both traditional modeling and high dimensional statistical learning and regression techniques.

Intra-Program Collaboration. Aim 2 of this project can provide insights into the clinical trial design methods developed in Aim 3 of Project 1. The candidate therapies studied in Aim 5 of Project 3 can be applied to the design methodology developed in Aim 3 of Project 1 as well as to the dynamic treatment regimes studied in Project 5. The multi-attribute utility function concept developed in Aim 5 of Project 3 can be useful in the individualized therapy discovery methodology studied in Aim 4 of Project 4. Both the missing data methodology of Project 2 and the genomic techniques developed in Project 4 could provide useful for Aims 1, 2 and 3 of Project 3. Project 3 will need Core C to assist with code for the needed simulation studies. Core B will be needed to provide some of the existing data sets for evaluation of the new methods. Core C will also be needed to prepare and disseminate the software which implements the new methods. Cross-cutting research themes are Statistical Learning for Aim 5; Model and Variable Selection for Aims 1, 3 and 5; Statistical Genetics and Genomics for Aims 1, 2 and 3; Semiparametric Inference and Empirical Processes for Aims 4 and 5; and Bayesian methods for Aims 1–3. Core A will facilitate interaction between Project 3 and the other projects and cores.

Project 4: Methods for Pharmacogenomics and Individualized Therapy Trials. There is great current interest in pharmacogenomic studies for identifying genetic determinants of inter-individual differences in the efficacy and toxicity of cancer medications and in individualized therapy trials for tailoring treatment regimens to each patient's genomic profile. The four specific aims of this project focus on developing novel and high-impact statistical and computational methods for design and analysis of such studies:

1. Construct robust and efficient statistical methods for assessing the effects of single nucleotide polymorphism (SNP) genotypes and haplotypes on drug response. We will develop statistical methods that can handle any phenotypes, including binary and continuous efficacy and toxicity measures, right-censored time-to-event outcomes, interval-censored time to disease progression, and informatively censored PSA levels and adverse events; accommodate population stratification and clinical factors correlated with genetic variables; and allow association analysis at the gene/pathway, haplotype or SNP level (even for SNPs not on the genotyping chip).

2. Develop statistical and data-mining techniques for predicting drug response based on high-dimensional and highly correlated genomic data. We will develop efficient variable selection procedures for ultra-high dimensional SNP and gene expression data under a variety of parametric and semiparametric regression models for all possible measures of drug response, allowing a hierarchical structure in selecting main effects and interactions and the inclusion of genetic variables at a group level. We will also develop machine learning techniques for classification with variable selection capabilities.

3. Investigate statistical procedures for providing low-bias estimation of effect sizes with complex and highly multivariate genetic data for follow-up and confirmation studies. We will explore a novel, conditional likelihood approach for producing low-bias estimation of effect sizes for follow-up and confirmation of effects and predictors. We will also pursue methods for a large number of simultaneous tests and penalized regression techniques for clinical outcomes.

4. Explore machine learning techniques for identifying candidate individualized therapies in both pre-clinical and clinical studies. We will provide a unified framework that combines the discovery power of data mining with the stabilizing influence of statistical inference by creating a new form of machine learning, called "latent supervised learning". We will utilize empirical process methods and advanced computational technology to develop and validate latent supervised learning for use in both pre-clinical and clinical studies for discovery of candidate individualized therapies for cancer.

Intra-Program Collaboration. All of the aims of this project have potential missing data challenges, and so the results of Project 2 will be useful here, especially Aim 4 of Project 2. The output of Aim 3.i of Project 1 will both benefit and be benefited by results from Aims 1–3 of Project 4. The results of Aim 5 of Project 3 will both contribute to developments in Aims 1–3 of Project 4 and benefit from techniques developed in Aim 4 of Project 4. The candidate therapeutic regimens derived in Project 4 can also serve as useful building blocks for the more complex dynamic treatment regimes involving multiple decision times developed in Project 5. Project 4 will need Core C to assist with code for the potentially complex simulation studies needed. Core B will be needed to provide some of the existing data sets for evaluation of the new methods. Core C will also be needed to prepare and disseminate the software that implements the new methods. Cross-cutting research themes are Statistical Learning for Aims 2 and 4; Model and Variable Selection for Aims 1–3; Statistical Genetics and Genomics also for Aims 1–3; Semiparametric Inference and Empirical Processes for all of the Aims; and Bayesian methods for Aims 1–3. Core A will facilitate interaction between Project 4 and the other projects and cores.

Project 5: Methods for Discovery and Analysis of Dynamic Treatment Regimes. In clinical practice, treatment of cancer is a dynamic process involving a series of therapeutic decisions over time. However, most cancer clinical trials focus on effects of treatments given at a single decision point in the course of the disease, e.g., the selection of a first-line chemotherapeutic option for patients with Stage IIIB/IV non-small cell lung cancer. Conclusions on the best overall strategy over the series of key decision points in the disease are consequently cobbled together from the results of many such single-decision studies, and, due in part to the possibility that the treatment given at one point in time may have delayed effects on the efficacy of future treatment, may be misleading and, indeed, possibly even harmful. This perspective has led to considerable recent interest in methodology for developing and studying *dynamic treatment regimes.* A dynamic treatment regime is a set of sequential decision rules dictating at each decision point the selection of the next treatment for a patient based on information on the patient, including measures of disease progression, biomarkers, and previous treatment, thereby individualizing each step of treatment to the patient. We propose four specific aims:

1. Develop and evaluate learning methods for optimal dynamic treatment regimes. Because of the complexity of the problem, standard statistical methods are not useful for identification of the optimal regime from data. *Reinforcement learning methods* from computer science, adapted to incorporate statistical inference, are a promising and powerful approach to this problem. In this aim, we will carry out the first, comprehensive study of reinforcement learning and other statistical learning methods in the context of cancer research.

2. Develop methods for identifying optimal dynamic treatment regimes from a restricted, feasible set. A key challenge in identification of the optimal dynamic treatment regime is that, with many decision points, treatment options, and high-dimensional patient information, the number of possible regimes can be enormous. An alternative, practical approach is to restrict the candidate regimes to a smaller, feasible set based on considerations including current clinical practice, cost, and complexity. We will develop methods for estimating mean outcome for regimes within a feasible set in order to facilitate identifying the best regime.

3. Develop and evaluate inferential methods for dynamic treatment regimes. Methods for making inference on optimal dynamic treatment regimes derived from the learning techniques in Aim 1 pose a significant challenge in that parameters in the statistical models that characterize these regimes are often constrained to lie on the boundary of the parameter space. Standard inferential approaches, including bootstrap methods, break down under these conditions, and a fundamentally new statistical framework is needed. We will address this challenge directly by using empirical process techniques to develop methods for constructing hypothesis tests and confidence intervals for optimal dynamic treatment regimes.

4. Develop methods for the design of sequentially randomized trials for dynamic treatment regimes. We will develop a new model for cancer clinical trials, *clinical reinforcement trials*, which involve sequential randomization, allow for a continuum of treatment options, and have the goal of developing optimal regimes using learning techniques. We will apply these first to non-small cell lung cancer and generalize to other cancers. A key challenge in the design of sequentially randomized studies for deducing optimal regimes is that, as the number of decision points and treatment options grows, the greater the sample size requirements can be. We will develop new approaches to evaluating the properties of these designs that will enable sufficient precision for finding optimal regimens with realistic sample sizes.

Intra-Program Collaboration. Because all of the aims of this project have potential missing data challenges, the results of Project 2 may shed some light here, especially the results from Aim 4 of Project 2. The candidate therapeutic regimens derived in Project 4 and those developed in Aim 5 of Project 3 can also serve as useful building blocks for the more complex dynamic treatment regimes here. The latent learning technology developed in Aim 4 of Project 4 will also be useful for Aim 1 of Project 5. Candidate dynamic treatment regimes developed in Project 5 could potentially be evaluated using techniques from Project 1, especially techniques from Aim 1.i. Project 5 will need Core C to assist with code for the the potentially complex simulation studies needed. Core B will be needed to provide some of the existing data sets for evaluation of the new methods. Core C will also be needed to prepare and disseminate the software which implements the new methods. Cross-cutting research themes are Statistical Learning for Aims 1 and 4; Model and Variable Selection for Aims 1–3; Statistical Genetics and Genomics also for Aims 1 and 2; Semiparametric Inference and Empirical Processes for all of the Aims; and Bayesian methods for Aim 1. Core A will facilitate the various modes of interaction needed between Project 4 and the other projects and cores.

Core A: Administrative Core. The Administrative Core will provide administrative support to the program project and be responsible for organizing the program investigators and staff into an effective and well-coordinated team to develop and implement the statistical methods for cancer clinical trials proposed in the research projects and supported by Cores B and C. The program leadership will be integrated across the three institutions with a lead PD/PI at one institution (UNC-CH) and two additional PD/PIs at the the other two institutions (NCSU and Duke). These three PD/PIs will form an Executive Committee with overall responsibility for the management and administration of the program. Each PD/PI will also be responsible for intra-institutional administration of the program. Each institution will have an additional co-PD/PI to assist the PD/PIs with both the overall and intra-institutional administration of the program project. The Executive Committee, three co-PD/PIs, and individual project leaders will form a Steering Committee providing overall scientific guidance for the program. An External Advisory Committee (EAC) of experts will provide feedback to the Steering Committee on the goals and progress of the program during an annual retreat. Communication and collaboration between project investigators will be facilitated with a program project wiki. Communication and dissemination of new results and software will be aided with a program project web page. The matrix leadership structure of Core A maximizes the scientific integration of this multi-disciplinary and trans-institutional collaboration. See Section **7.1**.

Core B: Data Compilation Core. The Data Compilation Core will develop and maintain a central resource of analysis-ready, annotated and documented data sets from cancer clinical trials and related studies to be used by the investigators in each of the individual research projects. These data sets will be used to evaluate and illuminate the methods developed in this program as well as to demonstrate the software developed in the Computational Resource and Dissemination Core (Core C). The primary sources of the data will be the clinical trials and related studies of the CALGB and data from cancer research studies conducted at the two large NCI-designated Comprehensive Cancer Centers at UNC-CH (Lineberger Comprehensive Cancer Center, LCCC) and at Duke (Duke Comprehensive Cancer Center, DCCC). This is a major advantage for the program

in that the data sets provided can be exceptionally well annotated and documented.

Core C: Computational Resource and Dissemination Core. The Computational Resource and Dissemination Core will assist with computational aspects and creation and dissemination of software implementing the new methods developed in each individual research project. Sound, tested implementations; simulation studies of performance; and demonstrations of use in applications are fundamental aspects of statistical methodological development to ensure that new techniques are reliable, accessible to, and adopted by the research community. Core C will be an essential part of the program project by achieving the following specific objectives: 1) Providing programming expertise in efficient and robust implementation of methodology; 2) Developing shared computational resources in support of project methodology; and 3) Creating and disseminating software and associated tutorials for methodology developed by the project in an accessible form to practitioners.

7 ORGANIZATION, ADMINISTRATIVE, AND PROGRAM MANAGEMENT STRUCTURE

7.1 Leadership Structure and Chain of Authority

The leadership structure for this program project consists of three interwoven components: overall program administration, intra-institutional administration, and project and core leadership. This is represented in Figure 2. In this figure, boxes delineate administrative units, dotted lines denote a supportive or advisory relationship, solid lines denote a supervisory relationship, and double arrows denote two-way support or joint supervisory roles. Dr. Kosorok at UNC-CH is lead PD/PI and will provide overall leadership for the program project. Dr. Kosorok's overall leadership responsibilities will be shared by two additional PD/PIs, Dr. Davidian at NCSU and Dr. George at Duke, who, together with Dr. Kosorok, will form the Executive Committee. Each member of the Executive Committee will be assisted by an additional co-PD/PI at his/her institution with both overall and intra-institutional leadership: Dr. Ibrahim at UNC-CH, Dr. Tsiatis at NCSU, and Dr. Jung at Duke. The PD/PI and co-PD/PI at each institution will constitute an intra-institutional Administrative Office to provide intra-institutional administrative support and leadership. The Executive Committee and the three Administrative Offices will provide administrative support to the Steering Committee, composed of the Executive Committee, the three co-PD/PIs, and individual project and core leaders: Drs. Kosorok, Davidian, George, Ibrahim, Tsiatis, Jung, Cai, and Lin. The Steering Committee will provide scientific oversight for the entire program project. This is shown in the figure by solid arrows from the Steering Committee pointing to the individual research project and Cores B and C. The leadership functions of the Executive Committee and Administrative Offices will be coordinated through the Administrative Core (Core A), led by Dr. Kosorok and the Executive Committee. This is indicated by double arrows between the Executive Committee and Core A. Core A will plan and coordinate the annual retreat and meeting with the EAC (External Advisory Committee), who will provide feedback and guidance to the Steering Committee. Core A also provides an Inter-Institutional Computing Committee to coordinate inter-institutional computing issues for the individual research projects and Cores B and C. The Data Compilation Core (Core B) and Computational Research and Dissemination Core (Core C) will support each other and all of the individual research projects. Only the top level of the individual research project and core leadership is given in Figure 2 for simplicity. Each project and core also will have co-Leaders and co-Directors as presented in Section **4**. The leadership structure within project and core is more-or-less traditional, while

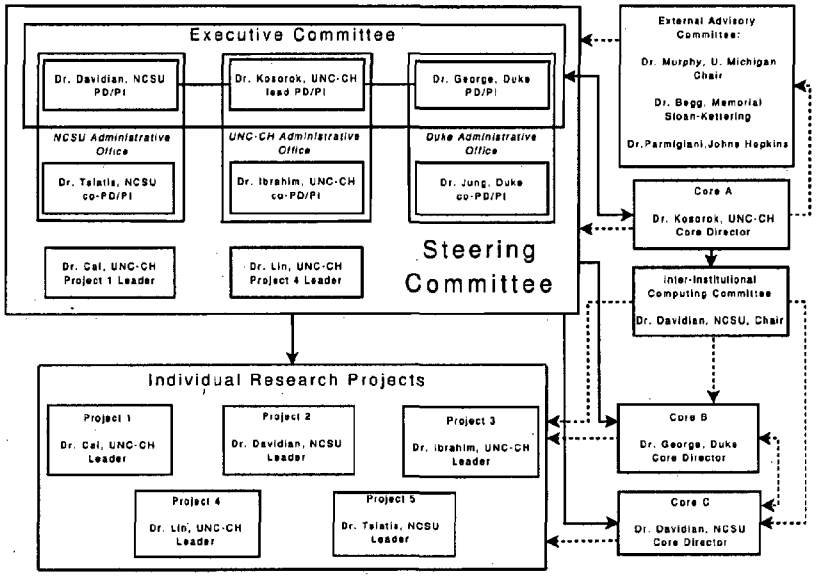


Figure 2: Program project leadership structure.

the overall leadership structure has a non-traditional matrix organization. This organization will create effective framework for scientific advances requiring multiple modes of attack and areas of expertise.

7.2 Procedures for Planning, Coordinating and Evaluating

The planning, coordination, and evaluation components include: meetings of the Executive Committee and Administrative Offices; Steering Committee meetings; project and core operations; an annual retreat, which includes the EAC (External Advisory Committee), and a program project web site and a wiki accessible to all investigators on the project.

The Executive Committee will meet monthly and additionally as needed, mostly via phone conference but also in person at least quarterly. Dr. Kosorok will chair the committee, which is responsible for the day-to-day management of the overall program. Overall program management activities include organizing regular meetings of the Executive and Steering Committees; organizing the annual retreat and meeting with the EAC; addressing inter-institutional computing issues; and preparation of annual reports, financial records and progress reports. Each institutional Administrative Office will also meet monthly and additionally as needed for day-to-day management of the intra-institutional components of the program.

The Steering Committee will provide scientific oversight for the program project, continuing assessment of project and core objectives, and articulation of program-wide objectives that involve interaction and integration across projects and cores. Dr. Kosorok will chair the Steering Committee, assisted by Drs. Davidian and George. The Steering Committee has been meeting in the process of preparing for this grant for at least six months and will continue to meet monthly throughout the program project period. Most of these monthly meetings will be via teleconference, but once a quarter the meetings will be face-to-face at the NISS facilities in RTP, where we have already met several times. Once a year, one of these face-to-face meetings at NISS will be part of the the program project annual retreat. Minutes from all of these meetings will be placed on the program project wiki for tracking and communication.

A traditional weekly meeting structure for each project and core is in general unlikely to be flexible enough to maximize the synergistic opportunities presented by the project. Thus we will modify the traditional structure by including weekly and ad-hoc scientific group meetings dictated by cross-cutting research. These weekly and ad-hoc scientific group meetings will be organized and directed as needed by project and core leaders and investigators and will be tracked and minutes recorded on the program project wiki,

An annual retreat will be held once a year at NISS. The retreat will include the Steering Committee and all project investigators and staff as well as a review by the EAC. The EAC will consists of three international experts in clinical trial methodology. Dr. Susan A. Murphy, H. E. Robbins Professor of Statistics, Professor of Psychiatry, and Research Professor in the Institute for Social Research at the University of Michigan, is the foremost world authority on dynamic treatment regimes and will serve as Chair of the committee. Drs. Colin B. Begg and Giovanni Parmigiani, both internationally recognized experts in cancer statistical methodology, will serve as additional members of the committee. This distinguished group will provide feedback to the Steering Committee on the progress and goals of the program project.

A dedicated web site for the program project will be developed and housed on a server at UNC-CH. The purpose of this web site is to be a single point of contact for all interested parties. The main page will be accessible to the general public and will include links to both an external set of pages also available to the general public as well as internal pages available only to program investigators. The external pages will include general information about the program projects, links to published papers, software developed by the program project, along with instructions and tutorials and other items for outreach. The internal pages will include a program project wiki for tracking and communicating among investigators, sharing data, developmental software and other digital information for investigators. The wiki will include pages for each project and core as well as other resources for the investigators. Each page has the capacity for editing and adding, deleting and changing additional pages as well as minutes, papers, figures and short comments used to track progress and communicate between investigators. We have been successfully using a prototype of the proposed wiki since June 2008, and we are certain it will continue to facilitate the matrix and cross-cutting aspects of the project.

7.3 Multiple PD/PI Leadership Plan

As noted previously, the proposed program project will employ a multiple PD/PI structure, the rationale for which follows naturally from the fact that the three institutions will all play significant scientific roles across

multiple research projects and cores. Accordingly, a strict hierarchy, with exclusive leadership at one of the three institutions, would be inconsistent with the extensive trans-institutional collaborative nature of the project. Rather, the proposed matrix structure, with both intra- and inter-institutional components, is more suitable. Dr. Kosorok will serve as the lead PD/PI, with Drs. Davidian and George as PD/PIs at NCSU and Duke, respectively. While in his lead role Dr. Kosorok will have some unique responsibilities, overall administrative and scientific leadership of the project as a whole will be shared by Drs. Kosorok, Davidian, and George, and no major decisions affecting the project will be made without consensus of all members. Simultaneously, each PD/PI will have responsibility for administration of project functions at his/her own institution as well as trans-project issues consistent with his/her additional roles as core and/or project leader.

The full Multiple PD/PI Leadership Plan is presented in detail in Item G.

7.4 Consultants

The only consultants on this program project are the members of the EAC (External Advisory Committee), who will be paid an honorarium for their service during the annual retreat:

- Dr. Susan A. Murphy, PhD, H. E. Robbins Professor of Statistics, Professor of Psychiatry, and Research Professor in the Institute for Social Research at the University of Michigan (Chair of the EAC).
- Dr. Colin B. Begg, PhD, Attending Biostatistician and Chair of the Department of Epidemiology and Biostatistics at the Memorial Sloan-Kettering Cancer Center in New York.
- Dr. Giovanni Parmigiani, PhD, Professor of Biostatistics at Johns Hopkins University.

7.5 Relationship with Other Units

The proposed program project is a joint venture of three institutions, UNC-CH, NCSU and Duke, with a PD/PI at each institution. The program will be administered in the Department of Biostatistics of the Gillings School of Global Public Health at UNC-CH, with Dr. Kosorok as lead PD/PI. The program subcontract to NCSU will be administered by the CQSB (Center for Quantitative Sciences in Biomedicine) in the College of Physical and Mathematical Sciences at NCSU, which has close ties to the Department of Statistics, with Dr. Davidian as PD/PI. The subcontract to Duke will be administered in the Department of Biostatistics and Bioinformatics in the School of Medicine, with Dr. George as PD/PI.

At UNC-CH, most of the investigators are housed in the Department of Biostatistics, but several investigators at UNC-CH are located in other departments. The UNC-CH Collaborative Studies Coordinating Center, housed in the Department of Biostatistics, is a large coordinating center involved in many clinical studies, most of which are non-cancer. While this center has a shared interest in clinical trials with the proposed program project, there is no overlap in scientific aims or shared resources. Many of the program investigators, with representation from all three institutions, are members of UNC-CH CICT (Center for Innovative Clinical Trials) in the Gillings School, the LCCC, and the TraCS Institute at UNC-CH (CTSA), with Dr. Ibrahim as Director of the CICT and Director of the Biostatistics Core of the LCCC and Dr. Kosorok as the Director of the Biostatistics Core of the TraCS Institute. Moreover, Drs. Davidian and Tsiatis at NCSU are also investigators on the UNC-CH TraCS Institute. The theme of clinical trial methodology is a shared emphasis for both the proposed program and the CICT, and some of the data obtained through Core B of the proposed program will be provided by the Biostatistics Core of the LCCC, however, there is no duplication of services nor scientific overlap of the program project with either the CICT or LCCC. Both the TraCS Institute and the proposed program project have some mutual interest in the development of biostatistical methodology, but there is no overlap in the aims of the program project. The Department of Statistics and Operations Research in the College of Arts and Sciences at UNC-CH has a cooperative relationship with the Department of Biostatistics in the Gillings School, and two investigators have joint appointments with that department (Drs. Ibrahim and Kosorok) and one investigator has a primary appointment there (Dr. Liu). Both departments have a strong national reputation with a long history and enjoy a productive and synergistic relationship.

All of the NCSU investigators are housed in the Department of Statistics in the College of Physical and Mathematical Sciences. Several of the NCSU investigators also have affiliations with other research entities, both on campus and at the other two campuses. Many of the program investigators, including Dr. Kosorok at

UNC-CH, are members of the CQSB, for which Dr. Davidian serves as Director, which is jointly supported by the NCSU Colleges of Physical and Mathematical Sciences and Agriculture and Life Sciences. Drs. Davidian and Tsiatis are adjunct faculty in the Department of Biostatistics and Bioinformatics at Duke, and in this capacity they spend one day per week at the DCRI (Duke Clinical Research Institute), where they have adjunct appointments on the DCRI Faculty and collaborate with DCRI biostatisticians and clinicians on cardiovascular disease research. Dr. Davidian is a member of the Executive Committee of the Center for Comparative Medicine and Translational Research (CCMTR) in the College of Veterinary Medicine, which has formal ties to the TraCS Institute. The CQSB, DCRI, and CCMTR research in biostatistical methodology does not overlap with the aims of the proposed program, but all four efforts share a common interest in methods for discovering and evaluating treatment strategies. In addition, the Department of Statistics at NCSU houses the NCSU Bioinformatics Research Center (BRC) which focuses specifically on computational and statistical methods for genetics and genomics studies and is thus related to Project 4 research; however, there is no overlap with the aims.

All of the biostatistical investigators at Duke are faculty members in the Department of Biostatistics and Bioinformatics and all are heavily involved in cancer research through the CALGB Statistical Center and the DCCC. The clinical co-investigators at Duke are all members of the DCCC with appointments in the Departments of Medicine or Surgery. These two organizations plus the Duke Translational Medicine Institute (DTMI,5 UL1 RR024128-03), established by the Duke CTSA, and the Department of Statistical Science at Duke are relevant to the project, although there is no duplication of core resources nor scientific overlap with the proposed program project aims. Some Department of Statistical Science faculty members have secondary appointments in the Department of Biostatistics and Bioinformatics, and vice-versa. Secondary appointments are also held by several faculty in the Center for Human Genetics, whose primary appointments are in the Department of Medicine, or in the Department of Community and Family Medicine. The CALGB Statistical Center located at Duke is directed by Dr. George; members include Drs. Jung, Owzar, Wang and Pang. Dr. George is also the Director of the DCCC shared resource in biostatistics as well as in the SPORE programs in breast, brain and lung cancers. The other investigators are all part of one or more of the shared resources and/or collaborate with individual members of the DCCC on other projects. There is a DTMI biostatistics core resource, directed by Dr. George and including Dr. Jung, as well as an internal biostatistics methodology research program within the DTMI, but there is no overlap with the proposed program project.

Dr. Kosorok is a member of the NISS Board of Trustees. While the mission of NISS is to identify, catalyze, and foster high-impact cross-disciplinary research involving the statistical sciences, there is no duplication of aims or core resources between the program project and NISS. Dr. Kosorok is a member of the Chairs Committee at SAMSI, and Dr. Davidian has been a research program organizer. There is no duplication of aims or core resources between the program project and SAMSI.

8 INSTITUTIONAL ENVIRONMENT AND RESOURCES

The Department of Biostatistics, located in the UNC Gillings School of Global Public Health, is one of the largest and highest ranked Biostatistics Departments in the U.S. The Department has over 35 full time faculty members; 130 graduate students pursuing MPH, MS, DrPH and PhD degrees; and 15 undergraduate students. The Department occupies a total of 26,333 square feet and boasts outstanding computer support for all students, faculty and staff and a state of the art 400 square-foot conference room which seats 20 people and has a drop-down projector, wireless capabilities and conference calling facilities. The Department is very supportive of the program project and will do all in its power to ensure its ongoing success; in particular, the Department will contribute $30,000 per year to the project for all five years of the grant. Dr. Kosorok, as lead PD/PI of the program project, is the Chair of the Department and will ensure that the needed support is provided. UNC-CH is one of the nation's foremost research universities, with top rankings in many disciplines. The UNC-CH Gillings School of Global Public Health is the top ranked public school of public health and has seven academic departments, including the Department of Biostatistics, and several centers, programs and institutes. The School also has several high-tech conference rooms, including the Blue Cross and Blue Shield of North Carolina Foundation Auditorium. Both the Gillings School and UNC-CH have pledged their support, including contributing $8,000 and $20,000, respectively, annually to the program project.

The NCSU CQSB, the administrative home for the NCSU component of the project, shares a state-of-the-art conference facility and two smaller conference rooms, each seating 10-15, people with the Center for

Research in Scientific Computing (CSRC) on the third floor of Cox Hall. The main conference facility seats 24-30 people in different configurations for conferences, seminars, and instructional events, and has the latest technology, including LCD projection equipment to display presentations on screens at the front and back of the room, Smart Board and Symposium technology, and video-conferencing capabilities. Almost all NCSU project investigators have offices in the NCSU Department of Statistics, which is one of the largest and oldest departments of statistics or biostatistics in the world, with approximately 70 undergraduate and 170 graduate students, and enjoys excellent office and meeting room facilities and excellent computer support for all of its faculty, students and staff. With UNC-CH, NCSU is one of the two flagship research institutions of the University of North Carolina system, with major colleges and schools of Agriculture and Life Sciences, Design, Education, Engineering, Humanities and Social Sciences, Management, Natural Resources, Physical and Mathematical Sciences, Textiles, and Veterinary Medicine. The College of Physical and Mathematical Sciences, which houses CQSB and the Department of Statistics, and the College of Agriculture and Life Sciences, which also supports CQSB, are very supportive of the proposed program project, and each will contribute $17,500 to the project in each of the five years of the grant. The University's Vice Chancellor for Research and Graduate Studies will contribute an additional $25,000 in each of the five years.

All of the biostatistical investigators at Duke are faculty members in the Department of Biostatistics and Bioinformatics, and the clinical co-investigators at Duke are all members of the DCCC with appointments in the Department of Medicine or the Department of Surgery at Duke. The Biostatistics and Bioinformatics Department currently consists of two divisions: Biostatistics (38 faculty members) and Computational Biology (7 faculty members) and has excellent office and meeting facilities and computational support for all of its faculty, students, and staff. The Department of Biostatistics and Bioinformatics at Duke is very supportive of this program project and will contribute $15,000 per year to the project for all five years of the grant. Duke University, founded in 1924, is a top ranked private school with many schools and colleges that are highly ranked nationally. The College of Arts and Sciences houses the nationally recognized Department of Statistical Science and the Duke University School of Medicine houses the Biostatistics and Bioinformatics Department. The Duke University School of Medicine is ranked in the top ten with schools twice its age, and is committed to socially relevant education, translational research, compassionate patient care and global healthcare solutions. There is ample meeting space for all program project investigators. The School of Medicine is very supportive of the program project and will contributed $15,000 per year in addition to the contribution from the Department of Biostatistics and Bioinformatics.

Dr. Alan F. Karr, NISS Director, has provided space to us during the past several months and has agreed to continue providing meeting space to the program project without charge. The NISS building in RTP, shared with its sister institute, SAMSI, has 28,000 square feet of state-of-the-art office and meeting space. The meeting space includes traditional conference rooms, a fully-equipped video conference room, and a lecture and conference room that supports web streaming.

Key to Acronyms: For convenience, we provide a key to some of the major acronyms used in the foregoing narrative.

CALGB	Cancer and Leukemia Group B
CICT	UNC-CH Center for Innovative Clinical Trials
CQSB	NCSU Center for Quantitative Sciences in Biomedicine
DCCC	Duke Comprehensive Cancer Center
DCRI	Duke Clinical Research Institute
DTMI	Duke Translational Medicine Institute
EAC	External Advisory Committee
ECOG	Eastern Cooperative Oncology Group
LCCC	Lineberger Comprehensive Cancer Center
NCSU	North Carolina State University
NISS	National Institute of Statistical Sciences
RTP	Research Triangle Park
SAMSI	Statistical and Mathematical Sciences Institute
TraCS	UNC-CH Translational and Clinical Sciences Institute
UNC-CH	University of North Carolina at Chapel Hill

9 PROGRAM-RELATED PUBLICATIONS

The following publications are results of the preliminary studies discussed in Section **5.3**.

Brown, E. R. and Ibrahim, J. G. (2003a). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221-228.

Brown E. R. and Ibrahim J. G. (2003b). Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59**, 686-693.

Brown E. R., Ibrahim J. G., and DeGruttola V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61**, 64-73.

Cai, J. and Zeng, D. (2004). Sample size/power calculation for case-cohort studies. *Biometrics* **60**, 1015-1024.

Cai, J. and Zeng, D. (2007). Power calculation for case-cohort studies with non-rare events. *Biometrics* **63**, 1288-1295.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2006). Posterior propriety and computation for the Cox regression model with applications to missing covariates. *Biometrika* **93**, 791–807.

Chi, Y. and Ibrahim J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62**, 432-445.

Chi, Y. and Ibrahim J. G. (2007). Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions. *Statistica Sinica* **17**, 445-462.

Cho, H., Ibrahim, J. G., Sinha, D., and Zhu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, in press. NIHMSID: 89842.

Chu, H., and Cole, S. R. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* **59**, 1331–1332.

Chu, H., and Guo, H. (2009). Letter to the Editor on "A unification of models for meta-analysis of diagnostic accuracy studies" by Harbord et al." *Biostatistics* **10**, 201–203. NIHMSID: 90001.

Gangnon, R. E. and Kosorok, M. R. (2004). Sample-size formula for clustered survival data using weighted log-rank statistics. *Biometrika* **91**, 263–275.

Ghosh, A., Zou, F., and Wright, F. A. (2008). Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *American Journal of Human Genetics.* **82**, 1064–1074.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica* **14**, 863-883.

Jeong, J. H. and Jung, S. H. (2006). Rank tests for clustered survival data when dependent subunits are randomized. *Statistics in Medicine* **25**, 361-373.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008) Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 672–680.

Johnson, B. A. and Tsiatis, A. A. (2004). Estimating mean response as a function of treatment duration in an observational study, where duration may be informatively censored. *Biometrics* **60**, 315–323.

Johnson, B. A. and Tsiatis, A. A. (2005). Semiparametric inference in observational duration-response studies, with duration possibly right-censored. *Biometrika* **92**, 605–618.

Jung, S. H. and George, S. L. (2009). Between-arm comparisons in randomized phase II trials. *Journal of Biopharmaceutical Statistics*, in press. NIHMSID: 89741.

Jung, S. H. and Jeong, J. H. (2003). Rank tests for clustered survival data. *Lifetime Data Analysis* **9**, 21-33.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* New York: Springer.

Kosorok, M. R., and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics* **35**, 957–989.

Lin, D. Y., Hu, Y., and Huang, B. E. (2008). Simple and efficient analysis of disease association with missing genotype data. *American Journal of Human Genetics* **82**, 444–452.

Lin, D. Y., and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* **101**, 89–118.

Lin, D. Y., Zeng, D., and Millikan, R. (2005). Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genetic Epidemiology* **29**, 299–312.

Lu, W. and Zhang, H. H. (2007). Variable selection for proportional odds model. *Statistics in Medicine* **26**, 3771-3781.

Lunceford, J., Davidian, M., and Tsiatis, A. A. (2002) Estimation of the survival distribution of treatment regimes in two-stage randomization designs in clinical trials. *Biometrics* **58**, 48–57.

Song, R., Kosorok, M. R., and Cai, J. (2008a). Robust covariate-adjusted log-rank statistics and corresponding sample size formula for recurrent events data. *Biometrics* **64**, 741–750.

Song, R., Kosorok, M. R., and Fine, J. P. (2008b). On asymptotically optimal tests under loss of identifiability in semiparametric models. *Annals of Statistics*, In press. NIHMSID: 75080.

Todem, D., and Fine, J.P. (2008). Global sensitivity testing for longitudinal data subject to potentially informative dropout. *Biometrics*, invited revision.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809-834.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* **27**, 4658-4677.

Wahed A. S. and Tsiatis A. A. (2004). Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics* **60**, 124–133.

Wahed A. S. and Tsiatis A. A. (2006). Semiparametric efficient estimation of survival distribution for two-stage randomization designs with censored data. *Biometrika* **93**, 163–178.

Wang, X., and Zhou, H. (2006). A semiparametric empirical likelihood method for biased sampling schemes in epidemiologic studies with auxiliary covariates. *Biometrics* **62**, 1149-1160.

Zeng, D. and Cai, J. (2005). Asymptotic results for maximum likelihood estimates in joint analysis of repeated measurements and survival time. *The Annals of Statistics* **33**, 2132-2163.

Zeng, D., Lin, D. Y., Avery, C. L., North, K. E., and Bray, M. S. (2006). Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics* **7**, 486–502.

Zhang, H. H., Liu, Y., Wu, Y., and Zhu, J. (2008). Variable selection for multicategory SVM via sup-norm regularization. *Electronic Journal of Statistics* **2**, 149–167.

Zhang, H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inference in randomized clinical trials using auxiliary covariates. *Biometrics* **64**, 707–715. PMCID:PMC2574960.

Zhang, M., Tsiatis, A. A., Davidian, M., Pieper, K., and Mahaffey, K. (2009) Inference on treatment effects from a randomized clinical trial in the presence of premature treatment discontinuation: The SYNERGY trial. Unpublished manuscript; submitted to *Biometrics*.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2008). Reinforcement learning design for cancer clinical trials. Submitted.

10 LITERATURE CITATIONS

Boos, D. D., and Stefanski, L. A. (2006). *Modern Statistical Inference: Theory and Methods*. New York: Springer.

Brown, E. R. and Ibrahim, J. G. (2003a). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221-228.

Brown E. R. and Ibrahim J. G. (2003b). Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59**, 686-693.

Brown E. R., Ibrahim J. G., and DeGruttola V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61**, 64-73.

Cai, J. and Zeng, D. (2004). Sample size/power calculation for case-cohort studies. *Biometrics* **60**, 1015-1024.

Cai, J. and Zeng, D. (2007). Power calculation for case-cohort studies with non-rare events. *Biometrics* **63**, 1288-1295.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2006). Posterior propriety and computation for the Cox regression model with applications to missing covariates. *Biometrika* **93**, 791–807.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2008). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chen, S., Watson, P., and Parmigiani, G. (2005). Accuracy of MSI testing in predicting germline mutations of MSH2 and MLH1: a case study in Bayesian meta-analysis of diagnostic tests without a gold standard. *Biostatistics* **6**, 450–464.

Chen, Q., Zeng, D., and Ibrahim, J. (2007). Sieve maximum likelihood estimation for regression models with covariates missing at random. *Journal of the American Statistical Association* **102**, 1309-1317.

Chi, Y. and Ibrahim J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62**, 432-445.

Chi, Y. and Ibrahim J. G. (2007). Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions. *Statistica Sinica* **17**, 445-462.

Cho, H., Ibrahim, J. G., Sinha, D., and Zhu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, in press. NIHMSID: 89842.

Chu, H., and Cole, S. R. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* **59**, 1331–1332.

Chu, H., and Guo, H. (2009). Letter to the Editor on "A unification of models for meta-analysis of diagnostic accuracy studies" by Harbord et al." *Biostatistics* **10**, 201–203. NIHMSID: 90001.

Fine, J. P., and Tsiatis, A. A. (2000). Testing for differences in survival with delayed ascertainment. *Biometrics* **56**, 145–153.

Food and Drug Administration (2004). Innovations or stagnation: Challenge and opportunity on the critical path to new medical products. White paper.

Gangnon, R. E. and Kosorok, M. R. (2004). Sample-size formula for clustered survival data using weighted log-rank statistics. *Biometrika* **91**, 263–275.

Ghosh, A., Zou, F., and Wright, F. A. (2008). Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *American Journal of Human Genetics*. **82**, 1064–1074.

Grossi, F., Gridelli, C., Aita, M. and De Marinis, F. (2008). Identifying an optimum treatment strategy for patients with advanced non-small cell lung cancer. *Critical Reviews in Onocology/Hematology* **67**, 16–26.

Högberg, T. (2005). Widening bottlenecks in drug discovery: Glimpses from Drug Discovery Technology Europe. *Drug Discovery Today* **10**, 820–822.

Huang, X., Stefanski, L. A., and Davidian, M. (2009). Latent-model robustness in joint models for a primary endpoint and a longitudinal process. *Biometrics*, in press. NIHMSID: 89865.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica* **14**, 863-883.

Jeong, J. H. and Jung, S. H. (2006). Rank tests for clustered survival data when dependent subunits are randomized. *Statistics in Medicine* **25**, 361-373.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008) Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 672–680.

Johnson, B. A. and Tsiatis, A. A. (2004). Estimating mean response as a function of treatment duration in an observational study, where duration may be informatively censored. *Biometrics* **60**, 315–323.

Johnson, B. A. and Tsiatis, A. A. (2005). Semiparametric inference in observational duration-response studies, with duration possibly right-censored. *Biometrika* **92**, 605–618.

Jung, S. H. and George, S. L. (2009). Between-arm comparisons in randomized phase II trials. *Journal of Biopharmaceutical Statistics*, in press. NIHMSID: 89741.

Jung, S. H. and Jeong, J. H. (2003). Rank tests for clustered survival data. *Lifetime Data Analysis* **9**, 21-33.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Kosorok, M. R., and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics* **35**, 957–989.

Lee, B. L., Kosorok, M. R., and Fine, J. P. (2005). The profile sampler. *Journal of the American Statistical Association* **100**, 960–969.

Lin, D. Y., Hu, Y., and Huang, B. E. (2008). Simple and efficient analysis of disease association with missing genotype data. *American Journal of Human Genetics* **82**, 444–452.

Lin, D. Y., and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* **101**, 89–118.

Lin, D. Y., Zeng, D., and Millikan, R. (2005). Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genetic Epidemiology* **29**, 299–312.

Lu, W. and Zhang, H. H. (2007). Variable selection for proportional odds model. *Statistics in Medicine* **26**, 3771-3781.

Lunceford, J., Davidian, M., and Tsiatis, A. A. (2002) Estimation of the survival distribution of treatment regimes in two-stage randomization designs in clinical trials. *Biometrics* **58**, 48–57.

Ma, S., Kosorok, M. R., Fine J. P. (2006). Additive risk models for survival data with high dimensional covariates. *Biometrics* **62**, 202–210.

Qu, P., Chu, H., Ibrahim, J. G., Peacock, J., Shen, X. J., Tepper, J., Sandler, R. S., and Keku, T. O. (2008). Statistical strategies to improve the efficiency of molecular studies of colorectal cancer prognosis. *British Journal of Cancer* **99**, 2001–2005.

Song, R., Kosorok, M. R., and Cai, J. (2008a). Robust covariate-adjusted log-rank statistics and corresponding sample size formula for recurrent events data. *Biometrics* **64**, 741–750.

Song, R., Kosorok, M. R., and Fine, J. P. (2008b). On asymptotically optimal tests under loss of identifiability in semiparametric models. *Annals of Statistics*, In press. NIHMSID: 75080.

Todem, D., and Fine, J.P. (2008). Global sensitivity testing for longitudinal data subject to potentially informative dropout. *Biometrics*, invited revision.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809-834.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* **27**, 4658-4677.

Wahed A. S. and Tsiatis A. A. (2004). Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics* **60**, 124–133.

Wahed A. S. and Tsiatis A. A. (2006). Semiparametric efficient estimation of survival distribution for two-stage randomization designs with censored data. *Biometrika* **93**, 163–178.

Wang, F., and Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* **17**, 193–208.

Wang, X., Pang, H., and Schwartz, T. A. (2009). Building and Validating High Throughput Lung Cancer Biomarkers. *Chance*: under revision.

Wang, W., and Yang, J. (2005). *Mining Sequential Patterns from Large Data Sets.* New York: Springer.

Wang, X., and Zhou, H. (2006). A semiparametric empirical likelihood method for biased sampling schemes in epidemiologic studies with auxiliary covariates. *Biometrics* **62**, 1149-1160.

Zeng, D. and Cai, J. (2005). Asymptotic results for maximum likelihood estimates in joint analysis of repeated measurements and survival time. *The Annals of Statistics* **33**, 2132-2163.

Zeng, D., Lin, D. Y., Avery, C. L., North, K. E., and Bray, M. S. (2006). Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics* **7**, 486–502.

Zhang, H. H., Liu, Y., Wu, Y., and Zhu, J. (2008). Variable selection for multicategory SVM via sup-norm regularization. *Electronic Journal of Statistics* **2**, 149–167.

Zhang, H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inference in randomized clinical trials using auxiliary covariates. *Biometrics* **64**, 707–715. PMCID:PMC2574960.

Zhang, M., Tsiatis, A. A., Davidian, M., Pieper, K., and Mahaffey, K. (2009) Inference on treatment effects from a randomized clinical trial in the presence of premature treatment discontinuation: The SYNERGY trial. Unpublished manuscript; submitted to *Biometrics*.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2008). Reinforcement learning design for cancer clinical trials. Submitted.

Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621–634.

11 LETTERS OF SUPPORT

- H. Shelton Earp, MD, Director of the UNC-CH Lineberger Comprehensive Cancer Center, Providing institutional support.
- Barbara K. Rimer, DrPH, Dean of the UNC-CH Gillings School of Global Public Health, Providing monetary and institutional support.
- Tony G. Waldrop, PhD, UNC-CH Vice Chancellor for Research and Economic Development, Providing monetary and institutional support.
- Raymond E. Fornes, PhD, Associate Dean for Research of the NCSU College of Physical and Mathematical Sciences, Providing monetary and institutional support.
- Terri L. Lomax, PhD, NCSU Vide Chancellor for Research and Graduate Studies, Providing monetary and institutional support.
- Steven A. Lommel, PhD, Associate Dean for Research of the NCSU College of Agriculture and Life Sciences, Providing monetary and institutional support.
- Sastry G. Pantula, PhD, Chair of the NCSU Department of Statistics, Providing institutional support.
- Nancy C. Andrews, MD, PhD, Dean of the Duke University School of Medicine, Providing monetary and institutional support.
- Elizabeth R. DeLong, PhD, Chair of the Duke Department of Biostatistics and Bioinformatics, Providing monetary and institutional support.
- H. Kim Lyerly, MD, Director of the Duke Comprehensive Cancer Center, Providing institutional support.
- Richard L. Schilsky, MD, Chair of Cancer and Leukemia Group B, Providing institutional support.
- Alan F. Karr, PhD, Director of the National Institute of Statistical Sciences, Providing meeting space and facilities.
- Colin B. Begg, PhD, Memorial Sloan-Kettering Cancer Center, Member of Program Project External Advisory Committee.
- Susan A. Murphy, PhD, University of Michigan, Chair of Program Project External Advisory Committee.
- Giovanni Parmigiani, PhD, Johns Hopkins University, Member of Program Project External Advisory Committee.

UNC
LINEBERGER COMPREHENSIVE CANCER CENTER
N.C. CANCER HOSPITAL

December 19, 2008

Michael R. Kosorok, PhD
Lead Principal Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Michael:

The strong partnership between UNC Lineberger Cancer Center and the Department of Biostatistics is decades old but the events of the last five years have taken it to another level. Recruitment has assembled an extraordinary group of statistical methodologic researchers, including Joe Ibrahim, Danyu Lin, Fred Wright, and Jason Fine, who contribute substantially to our cancer research. Under your leadership, an impressive era of scientific productivity is underway that will measurably improve clinical trials methodology.

As Director of the UNC Lineberger Comprehensive Cancer Center, I am writing to express my great enthusiasm and support for your application for an NCI Program Project (P01) entitled "Statistical Methods for Cancer Clinical Trials" at the University of North Carolina at Chapel Hill. This is certainly an important initiative for the UNC Lineberger, the Gillings School of Global Public Health, and for the University of North Carolina at Chapel Hill. With all of the recent advances in biomedicine, there remains a serious bottleneck between laboratory discoveries and their utilization in clinical practice. New clinical trials methodology is needed to keep abreast of and take advantage of molecular genetic discovery. I believe that the innovative program you have outlined will make important breakthroughs in solving this fundamental problem and have broad applicability for breast, colon and lung cancer as well as for other cancers and other diseases. I am very supportive of you and your research group utilizing existing clinical trial data sets housed in the Lineberger Comprehensive Cancer Center.

An important aspect of this program project is the collaboration with North Carolina State University and Duke University. This brings together a diverse group of investigators not only in biostatistics but also in medical oncology, health policy, pharmacogenomics, and computer science.

In summary, your program project application has my highest level of support and commitment. I will do all that I can to help you and your colleagues achieve the goals of this forward-looking P01 and, in the process, to help UNC become a leader in the field of cancer clinical trials.

Sincerely yours,

H. Shelton Earp III, MD
Director and Lineberger Professor
Professor of Medicine and Pharmacology

UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

170 ROSENAU HALL, CB 7400
135 DAUER DRIVE
CHAPEL HILL, NC 27599-7400

T 919.966.3215
F 919.966.7678
brimer@unc.edu

BARBARA K. RIMER, DrPH, MPH
Dean and Alumni Distinguished Professor

December 20, 2008

Michael R. Kosorok, PhD
Chair and Professor, Biostatistics
Lead Principal Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
UNC Gillings School of Global Public Health
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-7420

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Michael:

I write to express enthusiastic support for your National Cancer Institute Program Project (P01) application entitled "Statistical Methods for Cancer Clinical Trials" at the University of North Carolina at Chapel Hill. This is an important initiative for our School, the Lineberger Comprehensive Cancer Center and the University of North Carolina at Chapel Hill.

This is an interdisciplinary effort that brings together researchers from biostatistics, health policy, pharmacogenomics, medicine and computer science. The innovative program you have outlined in this proposal will significantly increase translation from laboratory discoveries to clinical practice which could lead to important improvements in public health – particularly in cancer research. I am especially enthusiastic about the cross-campus collaborations between Duke University, UNC and NC State University. The results of all of us working together could lead to great advances that ultimately can benefit patients.

We will contribute $8,000.00 per year toward this grant for each of the five years of the award. I wish we could do more, but we face additional rounds of budget cuts.

The program project "Statistical Methods in Cancer Clinical Trials" has my highest level of support and commitment. I pledge to do what I can to help achieve the goals you have set.

Warm regards,

Barbara K. Rimer

Barbara K. Rimer

BKR/smb

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

OFFICE OF THE VICE CHANCELLOR FOR
RESEARCH AND ECONOMIC DEVELOPMENT

312 SOUTH BUILDING
CAMPUS BOX 4000
CHAPEL HILL, NC 27599-4000

T 919-962-1319
F 919-962-1476

December 19, 2008

Michael R. Kosorok, PhD
Lead Principal Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Michael:

As Vice Chancellor for Research and Economic Development, I am writing to express very enthusiastic institutional support for your application for a National Cancer Institute Program Project (P01) entitled "Statistical Methods for Cancer Clinical Trials" at the University of North Carolina at Chapel Hill. This is certainly an important initiative for the Gillings School of Global Public Health, the Lineberger Comprehensive Cancer Center, and for the University of North Carolina at Chapel Hill. As you know, my own research background is in physiology, so I am attuned to the exciting developments in basic biomedical research knowledge and the enormous potential that exists to translate this knowledge into improvements in public health. Even with all of the recent advances in biomedicine, there remains a serious bottleneck between laboratory discoveries and their utilization in clinical practice. I believe that the innovative program you have outlined in this proposal will significantly relieve this bottleneck and lead to important improvements in public health, especially in cancer. I am also pleased that your collaborators include a diverse range of disciplines and departments across the university, not only in the Gillings School of Global Public Health but also researchers from the School of Medicine, the School of Pharmacy, and the College of Arts and Sciences.

Another important facet of this project is the collaboration with North Carolina State University and Duke University that will be both leveraged and facilitated by your program. We are very supportive of inter-university cooperation of this kind and recognize that this combination of institutions offers a uniquely powerful resource for making advances in clinical trial methods research that will have a high public health impact.

Because of the importance and value of this project for the university, we will contribute $20,000.00 per year towards this grant for each of the five years of the award.

In summary, the program project "Statistical Methods in Cancer Clinical Trials" has my highest level of support and commitment. I pledge to do whatever I can to see that we achieve the goals you have laid out and, in the process, become a leader in the field of cancer clinical trials.

Sincerely,

Tony G. Waldrop, PhD
Vice Chancellor for Research and Economic Development

North Carolina State University is a land-grant university and a constituent institution of The University of North Carolina

College of Physical and Mathematical Sciences

NC STATE UNIVERSITY

Office of the Associate Dean for Research
Campus Box 8209 / 300 Cox Hall
Raleigh, NC 27695-8209
919.515.7865 (phone)
919.515.7668 (fax)

January 9, 2009

Marie Davidian, PhD
Program Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
Center for Quantitative Sciences in Biomedicine and
Department of Statistics
North Carolina State University
Raleigh, NC 27695

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Marie:

I am writing to offer my enthusiastic endorsement and commitment to your application for a National Cancer Institute Program Project (P01) Award, entitled "Statistical Methods for Cancer Clinical Trials," which will be a joint venture between North Carolina State University, Duke University, and the University of North Carolina at Chapel Hill. This important project is consistent with the College's emphasis on health-related research and in particular with the mission of the Center for Quantitative Sciences in Biomedicine (CQSB), which our College strongly supports and which will be the administrative home for the project.

I am well aware of the critical role the quantitative sciences and statistical science in particular play in the development of new methodology for the conception, design, and analysis of clinical trials, and I am excited at the prospect that the innovative program of research proposed in this project will lead to new advances that can speed discoveries in the laboratory to clinical practice in the treatment of cancer. I am also pleased that the project involves a significant collaboration leveraging the complementary expertise at our institution, Duke University, and the University of North Carolina at Chapel Hill, which together comprise an unparalleled resource for this sort of effort.

The College is pleased to contribute $17,500 per year for each of the five years of the award to the CQSB in support of the activities of this project. The project has my highest level of support. I look forward to hearing of the progress you make on your ambitious research program, and I and the College are happy to assist you I any way we can to ensure that in the goals of the project are achieved.

Sincerely,

Raymond E. Fornes, PhD
Associate Dean for Research, College of Physical and Mathematical Sciences

NC STATE UNIVERSITY

Office of the Vice Chancellor
Campus Box 7003
103 Holladay Hall
Raleigh, NC 27695-7003

919.515.2117
919.515.7521 (fax)

January 12, 2009

Marie Davidian, PhD
Program Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
Center for Quantitative Sciences in Biomedicine and
Department of Statistics
North Carolina State University
Raleigh, NC 27695

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Marie:

As Vice Chancellor for Research and Graduate Studies at North Carolina State University, I am pleased to offer my enthusiastic support for your inter-university grant application for a National Cancer Institute Program Project (P01), entitled "Statistical Methods for Cancer Clinical Trials." This exciting initiative, which will translate advances in basic biomedical science to clinical practice and develop new ways to conceive and evaluate treatment strategies for cancer, is consistent with the University's emphasis on health-related research. It is also an ideal endeavor in which to exploit the strengths of the Center for Quantitative Sciences in Biomedicine (CQSB), which the University strongly supports and which will serve as the administrative home for the project.

I am especially pleased that this project offers yet another opportunity for trans-institutional collaboration with Duke and the University of North Carolina at Chapel Hill. The resources at our three institutions for carrying out such a transformative project are unique and abundant, and integrating them in the way that you propose is certain to result in advances in clinical trial methods research that will have high visibility and impact.

To recognize the value of this project to the University's research mission and to assist you in achieving your ambitious research objectives, we will contribute $25,000 per year in each of the five years of the award to the CQSB in support of the activities of this project. I look forward to assisting you in any way I can to ensure the success of the project.

Sincerely,

Terri L. Lomax

Terri L. Lomax, PhD
Vice Chancellor for Research and Graduate Studies

TLL/mh

North Carolina State University is a land-grant university and a constituent institution of The University of North Carolina

College of Agriculture and Life Sciences
North Carolina Agricultural Research Service

NC STATE UNIVERSITY

Campus Box 7643
Raleigh, NC 27695-7643

919.515.2717
919.515.7745 (fax)
ag_research@ncsu.edu

January 9, 2009

Marie Davidian, PhD
Program Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
Center for Quantitative Sciences in Biomedicine and
Department of Statistics
North Carolina State University
Raleigh, NC 27695

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Marie:

I am pleased to offer my unqualified support of your grant application for a National Cancer Institute Program Project (P01) entitled "Statistical Methods for Cancer Clinical Trials." This exciting project fits well with our College's focus on the life sciences and health as well as with that of the Center for Quantitative Sciences in Biomedicine (CQSB), to which our College is strongly committed. I am especially enthusiastic about the opportunity this project represents for trans-institutional collaboration, which will draw on the complementary strengths of our institution, Duke, and the University of North Carolina at Chapel Hill, and I am pleased that the CQSB is a partner in this important effort.

In support of this transformative project, the College is pleased to commit $17,500 per year for each of the five years of the project to the CQSB in support of the activities of this project. The College and I pledge to assist you in any way possible to advance the goals of the project and contribute new innovations to cancer clinical trials methodology.

Sincerely,

Steven A. Lommel, PhD
Associate Dean for Research, College of Agriculture and Life Sciences

Cc: Ray Fornes
Mike Cross
Joy Martin
Gail Hill

North Carolina State University is a land-grant university and a constituent institution of The University of North Carolina

Sastry G. Pantula
Head, Department of Statistics
Director, Institute of Statistics
Campus Box 8203 / Patterson Hall
Raleigh, NC 27695-8203

NC STATE UNIVERSITY

College of Physical and Mathematical Sciences
College of Agriculture and Life Sciences

919.515.1949
919.515.7591 (fax)
pantula@stat.ncsu.edu
www.stat.ncsu.edu

January 14,2009

Marie Davidian, PhD
Program Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trial" Program Project
Center for Quantitative Sciences in Biomedicine and
Department of Statistics, North Carolina State University
Raleigh, NC 27695

Re: PAR-09-025, National Cancer Institute Program Project (P01) Applications

Dear Marie:

I am delighted to lend my enthusiastic support to your application for a National Cancer Institute Program Project (P01) Award, entitled "Statistical Methods for Cancer Clinical Trials," which will be a joint venture between North Carolina State University, Duke University, and the University of North Carolina at Chapel Hill and which will involve a number of faculty from the Department of Statistics. This project fits well with the Department's many activities related to the health sciences, including our popular graduate and research programs in biostatistics, bioinformatics, and biomathematics; our established relationship with Duke Clinical Research Institute through our joint training and graduate internship programs; and our recently minted relationship with the Department of Biostatistics at the University of North Carolina at Chapel Hill (UNC-CH) through your membership in the Biostatistics Core of the Translational and Clinical Sciences (TraCS) Center. I am especially excited about the opportunity the project represents for expanded and deeper collaboration among faculty in our Department, the Department of Biostatistics and Bioinformatics at Duke, and the Department of Biostatistics at UNC-CH. The project will be an important resource for all of our faculty, students, and postdocs, who will be exposed to the cutting-edge research on methodology for cancer clinical trials that you propose.

I am pleased that the both Colleges of Physical and Mathematical Sciences and Agriculture and Life Sciences, in which our Department jointly resides, have committed generous support to the project. In addition to the funds they have provided, I am pleased to commit space in the new Mathematics and Statistics Building, to which the Department will move in May 2009, to house the equipment that will host the project software repository as well as office and meeting space for project activities as needed.

The Department is eager to contribute to the success of this high-profile and important project in any way we can. Please do not hesitate to contact me if you need further resources. Also, as the President-Elect of the American Statistical Association, I am thrilled to see this proposal and its benefits to our profession and for human health.

Sincerely,

Sastry G. Pantula

Sastry G. Pantula

Duke University
School of Medicine

Nancy C. Andrews, M.D., Ph.D.
Dean, Duke University School of Medicine
Vice Chancellor for Academic Affairs

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As Dean of the Duke University School of Medicine and Vice Chancellor of Academic Affairs, I am writing to express my strong support for the Duke participation in the multi-institutional Program Project (P01) application to the National Cancer Institute entitled "Statistical Methods for Cancer Clinical Trials". The overall scientific goal of this project, to develop highly innovative methods for cancer clinical trials, is especially important in speeding the introduction of effective new therapies into practice, and is in line with the strategic research plans for the School of Medicine. The involvement of two other major universities in our region, the University of North Carolina and North Carolina State University, provides an outstanding opportunity for collaborative research.

Because of the importance and value of this program, The School of Medicine will contribute $15,000 per year towards this grant for each of the five years of the award.

In summary, the program project "Statistical Methods in Cancer Clinical Trials" has my enthusiastic support and commitment. I pledge to do whatever I can to see that we achieve the goals you have laid out.

Sincerely,

Nancy Andrews, M.D., Ph.D.
Dean, School of Medicine

BOX DUMC 2927
Durham, NC 27710
TEL 919.684.2455
FAX 919.684.0208
URL medschool.duke.edu
dukemedicine.org

DUKE UNIVERSITY MEDICAL CENTER
Department of Biostatistics and Bioinformatics

Telephone: (919) 684-9447
Facsimile (919) 681-7918

January 13, 2009

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As chair of the Department of Biostatistics and Bioinformatics, I am writing to express my enthusiastic support for the P01 application entitled "Statistical Methods for Cancer Clinical Trials", in which Duke will participate jointly with the University of North Carolina – Chapel Hill and North Carolina State University. The overall goal of the research, to transform the current paradigm for drug discovery and translation to clinic, resulting in improved survival and quality of life for cancer patients, is extremely important in itself. And the opportunity for our faculty to engage in high level collaborative research in statistical methodology is consistent with the strategic plans of our department.

Because of the importance of the research, I am willing to commit $15,000 per year of the grant for use in offsetting the costs of research. In addition, I will help in whatever other ways are needed to help this program succeed.

Sincerely,

Elizabeth R. DeLong, Ph.D.
Professor and Chair

Mail: DUMC 2721 • DURHAM, NORTH CAROLINA 27710
Courier: 11th Floor Hock Plaza • 2424 Erwin Road • Durham, NC 27705

DukeMedicine

Duke Comprehensive Cancer Center
A National Cancer Institute-designated Comprehensive Cancer Center

H. Kim Lyerly, MD
George Barth Geller Professor of Research in Cancer
Director
Duke Comprehensive Cancer Center

January 13, 2009

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As Director of the Duke Comprehensive Cancer Center, I am writing to express my strong support for your Program Project (P01) application entitled "Statistical Methods for Cancer Clinical Trials". The overall scientific goal of this project, to develop highly innovative methods for cancer clinical trials, is highly relevant to the strategic plans of the DCCC. Efficient statistical methods are extremely important in accelerating the development of anti-cancer therapy and in translating results into clinical practice. Developments from your proposed research program can be quickly implemented in cancer research projects here because of your role as the director of the biostatistics unit in the DCCC.

I am enthusiastic about this program and I pledge to help in whatever I can to see that you achieve the goals you have laid out.

Sincerely,

H. Kim Lyerly, M.D.
George Barth Geller Professor of Research in Cancer
Director, Duke Comprehensive Cancer Center

BOX DUMC 2714, Durham, NC 27710
LOC 2424 Erwin Road
Hock Plaza, Suite 601
Durham, NC 27705
TEL 919.684.5613
FAX 919.684.5653
EMAIL lyerl001@mc.duke.edu
URL www.cancer.duke.edu
dukemedicine.org

Cancer and Leukemia Group B
CENTRAL OFFICE OF THE CHAIRMAN

230 W. Monroe Street, Suite 2050
Chicago, IL 60606-4703

TEL (773) 702-9171
FAX (312) 345-0117

www.calgb.org

Richard L. Schilsky, M.D.
Chairman

January 14, 2009

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As chair of the Cancer and Leukemia Group B (CALGB), I am writing to express my enthusiastic support for your P01 application entitled "Statistical Methods for Cancer Clinical Trials". Indeed, the CALGB will be a major partner in this research through the participation of several clinical co-investigators participating from Duke and UNC, through the involvement of statisticians from the CALGB Statistical Center, which you direct as Group Statistician, and through the sharing of data from selected CALGB studies to illustrate the methods that are developed. The overall goal of the research, to transform the current paradigm for drug discovery and translation to practice, resulting in improved survival and quality of life for cancer patients, is a shared goal of the CALGB. For all of these reasons, it is anticipated that the results from this program can and will be implemented directly and immediately into the design and analysis of CALGB studies, to the benefit of all.

In summary, I enthusiastically support this program and look forward to our partnership in achieving its aims.

Sincerely,

Richard L. Schilsky, M.D.
Chair, Cancer and Leukemia Group B
Professor of Medicine
University of Chicago

NISS

National Institute of Statistical Sciences
PO Box 14006, Research Triangle Park, NC 27709-4006
Tel: 919.685.9300 FAX: 919.685.9310
www.niss.org

Alan F. Karr, Director
karr@niss.org

January 5, 2009

Dr. Michael Kosorok
Dr. Marie Davidian
Dr. Stephen George
Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg, CB 7420
Chapel Hill, North Carolina 27599–7420

Dear Michael, Marie, and Steve,

I am delighted to hear that you and your colleagues at Duke University, North Carolina State University, and the University of North Carolina at Chapel Hill are collaborating on a application for a joint Program Project grant from the National Cancer Institute on "Statistical Methods for Cancer Clinical Trials."

The Research Triangle is a natural setting for trans-institutional research projects such as the one you are proposing, and, as you know, the National Institute of Statistical Sciences (NISS) and the Statistical and Applied Mathematical Sciences Institute (SAMSI) have a long history of catalyzing and facilitating such interactions. NISS would be pleased to support this important initiative by making our centrally-located facilities available to you and other project personnel for meetings during the project period, as we have already done during the months leading up to the submission of your application. These facilities include "traditional" conference rooms, a fully-equipped video conference room and a lecture room and conference that support web streaming of events.

I wish you success with this proposal, and NISS looks forward to hosting activities associated with the project.

Sincerely,

Sincerely,

Colin B. Begg, PhD
Eugene W. Kettering Chair
Department of Epidemiology & Biostatistics

December 23, 2008

Michael R. Kosorok, PhD
Marie Davidian, PhD
Stephen L. George, PhD
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg Hall
CB 7420
Chapel Hill, NC 27599-7420

Re: Statistical Methods in Cancer Clinical Trials

Dear Michael, Marie and Steve,

I am writing to confirm my willingness to serve on the External Advisory Committee for your joint P01 Program Project application entitled "Statistical Methods in Cancer Clinical Trials". Although the key ingredients of clinical trial methodology have been established for many decades, the new drug development paradigm of trying to create new agents that are specifically targeted to the characteristics of relatively small subgroups of patients, with the ultimate goal of "personalized medicine", promises to change the landscape for designing and analyzing clinical trials. At this juncture we certainly need fresh, innovative approaches to maximize the efficiency of the drug development and testing strategies in the context of this paradigm. Statistical methods must play a central role in this effort. The group of investigators you have put together to tackle these difficult issues is impressive, drawing on the considerable strengths of your three institutions. Your team encompasses several prominent experts in both statistical theory and the application of clinical trials, and so you are in a great position to enhance our knowledge in this important area. I am very happy to serve on your External Advisory Committee, and generally to help in any way I can.

With best wishes,

Colin

Colin Begg, PhD
Attending Biostatistician
Chair, Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center
307 East 63rd Street, 3rd Floor, New York, New York 10065
Telephone 646.735.8108 • FAX 646.735.0009
E-mail: beggc@mskcc.org

The University of Michigan
Department of Statistics

12/17/2008

Michael R. Kosorok, PhD, Marie Davidian, PhD, Stephen L. George, PhD
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg Hall
CB 7420
Chapel Hill, NC 27599-7420

Re: Statistical Methods in Cancer Clinical Trials

Dear Michael, Marie and Steve,

I am very happy to serve as chair of the External Advisory Committee for your joint Program Project "Statistical Methods in Cancer Clinical Trials" that you are submitting to the National Cancer Institute. This is an exiting project that will bring together the combined strengths of Duke University, North Carolina State University and the University of North Carolina at Chapel Hill.

I am keenly aware of the importance of clinical trials in the discovery of new treatments, and, as you know, have worked for many years in my own research on creating new clinical trial methods, especially in the areas of dynamic treatment regimes and reinforcement learning. I believe that the application of these new areas to cancer, as well as many of the other novel approaches proposed in your application, will likely have a large impact on public health.

As a member of the External Advisory Committee, I am looking forward to following your research progress and providing feedback on at least an annual basis. I wish you success in your application.

Sincerely,

Susan Murphy

H. E. Robbins Professor of Statistics
Research Professor, Institute for Social Research
Professor of Psychiatry

Address	*Telephone*	*Fax*	*email/URL*
Department of Statistics	734-647-3684	734-763-4676	samurphy@umich.edu
444D West Hall			http://www.stat.lsa.umich.edu/~samurphy
The University of Michigan			
Ann Arbor, MI 48109-1107			

THE SIDNEY KIMMEL
COMPREHENSIVE CANCER CENTER
AT JOHNS HOPKINS

CLINICAL TRIALS AND BIOMETRY

Oncology Biostatistics
550 North Broadway, Suite 1103
Baltimore, Maryland 21205-2013
Office (410) 955-4884 Fax (410) 955-0859
http://www.cancerbiostats.onc.jhmi.edu

Program Leader
Steven Goodman, MD, PhD
Professor, Acting Director
(410) 955-4596

Faculty
Giovanni Parmigiani, PhD
Professor, Director Bioinformatics Core
(410) 614-3426

Michael Ochs, PhD
Associate Professor
(410)955-8830

Peng Huang, PhD
Visiting Associate Professor
(410)502-0944

Jeanne Kowalski, PhD
Assistant Professor
(410) 955-4286

Leslie Cope, PhD
Assistant Professor
(410) 502-0945

Sarah Wheelan, MD, PhD
Assistant Professor
(410) 955-8841

Xiaobu Ye, MD, MS
Research Associate
(410) 614-6261

Senior Biostatisticians
Marianna Zahurak, MS
(410) 955-4219

Biostatisticians
Amanda Blackford, ScM
(410) 614-0361

Hua-Ling Tsai, MS
(410) 502-6529

Zhe Zhang, MS
(410) 502-0946

Administrative Staff
Helen Cromwell
Administrative Manager
(410) 955-4885

Alisa Moore
Administrative Coordinator
(410) 614-3432

December 18, 2008

Michael R. Kosorok, PhD
Marie Davidian, PhD
Stephen L. George, PhD
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg Hall
CB 7420
Chapel Hill, NC 27599-7420

Re: Statistical Methods in Cancer Clinical Trials

Dear Michael, Marie and Steve,

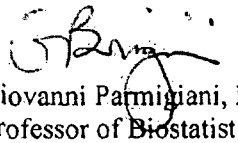
I am pleased to serve as a member of the External Advisory Committee for your joint Program Project "Statistical Methods in Cancer Clinical Trials" that you are submitting to the National Cancer Institute.

Many of your research aims involve genomics and Bayesian methods, areas of research in which I have been very active. Thus I am keenly aware of the potential these methods have in improving design and analysis of cancer clinical trials. I believe your proposed research in cancer clinical trials is fundamental and will likely have a large public health impact. This is an exciting project that will bring together the combined strengths of Duke University, North Carolina State University and the University of North Carolina at Chapel Hill.

NCI CCC

As a member of the External Advisory Committee, I am looking forward to following your research progress and providing feedback on at least an annual basis. I wish you success on your application.

Best wishes,



Giovanni Parmigiani, PhD
Professor of Biostatistics
Johns Hopkins University

MULTIPLE PD/PI LEADERSHIP PLAN

Lead PD/PI: Michael R. Kosorok, PhD
PD/PI: Marie Davidian, PhD
PD/PI: Stephen L. George, PhD

MULTIPLE PD/PI LEADERSHIP PLAN

Dr. Kosorok's duties as lead PD/PI include managing the integration of the five projects and three core facilities; leading Core A operations; scheduling regular meetings of the Executive Committee, the UNC-CH Administrative Office, the Steering Committee, and the Annual Retreat and Meeting with the External Advisory Committee. Dr. Kosorok chairs the Executive and Steering Committees and also coordinates yearly summaries of progress and non-competitive renewal materials. Other activities include preparation of annual reports, oversight of budgetary disbursements and financial records for both the program project and the UNC-CH components, and communication with NIH and other external communications that impact the program.

Dr. Kosorok is assisted in his overall leadership responsibilities and duties by Drs. Davidian and George who comprise the Executive Committee and who share overall administrative and scientific leadership and responsibility for the program project. Dr. Davidian is also the Core C leader, Project 2 leader and chair of the inter-institutional computing committee which coordinates trans-institutional computing issues. As chair of this committee, Dr. Davidian will provide overall program project leadership for inter-institutional computing issues. Dr. Davidian also has overall responsibility for the NCSU budgetary components of the program. Dr. George is also the Core B leader, and, as such, coordinates trans-institutional data sharing and data compilation issues. Dr. George also has overall responsibility for the Duke budgetary components of the program. Drs. Davidian and George also assist Dr. Kosorok with external communication, although Dr. Kosorok is the primary contact person.

Dr. George will provide overall program project leadership on human subjects issues. Dr. George will also be responsible for any human subjects approvals at Duke University. Drs. Davidian and Kosorok will be responsible for any human subjects approvals at NCSU and UNC-CH, respectively, with assistance and advice from Dr. George as needed.

Dr. Kosorok will be responsible for preparation of annual non-competitive renewal applications and dispersion of funds to the subcontracts at NCSU and Duke as well as to the project and core components at UNC-CH. Drs. Davidian and George will be responsible for assisting with the non-competitive renewal applications as well as dispersing funds to the project and core components at their respective institutions. No budgetary changes in subcontracts will be undertaken without approval from the Administrative Office at the affected institution. No budgetary changes affecting projects or cores will be undertaken without consulting with the affected project or core leaders nor without approval of the Executive Committee.

The proposed administrative structure is not a strict hierarchy but is more of a matrix with both intra-institutional leadership under the Administrative Offices and trans-institutional scientific leadership of the overall program as well as of individual projects and cores. While Dr. Kosorok is the lead PD/PI, the overall leadership and responsibility for the core is shared among all members of the Executive Committee, and no major decisions affecting the project will be made without the consensus of the entire Executive Committee. Moreover, no major changes in scientific focus or budget allocations to projects and cores will be made without input and guidance from the Steering Committee (which includes all project and core leaders). We have been successfully functioning in this manner for about six months now, and we do not anticipate there being any conflicts or difficulties that cannot be successfully resolved within this administrative structure.

The rationale for the proposed multiple PD/PI leadership structure follows naturally from the fact that the three institutions involved are all playing a significant scientific role across multiple projects and cores within the proposed program project. More importantly, the novel matrix administrative structure will greatly facilitate research by fostering and coordinating trans-institutional collaboration.

PROJECT 1

INNOVATIVE CLINICAL TRIAL DESIGN AND ANALYSIS

Project Leader: Jianwen Cai, PhD

PROJECT SUMMARY (See instructions):

Study design is a crucial first step in clinical trials. Well-designed studies are essential for successful cancer research and cancer drug development. Innovative clinical trial designs can potentially require fewer patients, save resources, and accelerate cancer drug development. The broad, long-term objective of this research project is to develop new statistical methodology to address new and challenging issues in the design and analysis of cancer clinical trials. There are 3 specific aims in this project. The first aim addresses statistical methods for the design and sample size calculation for longitudinal data and joint models for longitudinal and survival data. Statistical methods will be developed for sample size and power estimation for the overall and direct treatment effect on survival, for the effect of the longitudinal process on survival, and for settings involving multivariate longitudinal and multivariate survival processes. The second aim studies statistical methodology for the design and analysis of group randomized cancer prevention trials with survival and recurrent event outcomes. Empirical process theory will be used to study the asymptotic behavior of the test statistics and both asymptotic approximation as well as permutation test will be used to develop sample size formulas and power estimation. The third aim addresses important statistical issues in the oncology drug development pathway. There are three sub-aims. The first sub-aim is in the area of targeted designs. Methods for alternative designs, including "enrichment" designs, will be developed, and the operating characteristics and costs of these designs to fully targeted designs will be compared. Valid and efficient statistical methods for these trials will be developed by applying a semiparametric empirical likelihood approach. The second sub-aim is in the area of phase II designs. New methods for phase II and phase II/III clinical trials will be developed and their operating characteristics, costs, and predictive ability for subsequent phase III trials will be assessed. Information on both combination and non-combination therapies in phase II studies and subsequent phase III studies will be gathered to build prediction models using machine learning and other nonparametric classification methods. The third sub-aim is in the area of partially randomized designs. New semiparametric empirical likelihood methods will be developed for the design and analysis of such trials to adjust for selection bias and to improve efficiency. Our research will produce important new and efficient design and analysis tools for cancer research.

RELEVANCE (See instructions):

This research will provide valuable new design and analysis tools to cancer researchers and other biomedical researchers. These new and improved design and analysis tools will help to improve the quality and efficiency of cancer clinical trials. They will help to improve public health by enabling accurate and efficient estimation of sample size and power calculation for cancer clinical trials and by accelerating cancer drug development.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location			
Organizational Name: The University of North Carolina at Chapel Hill			
DUNS: 608195277			
Street 1: Office of Sponsored Research, CB #1350	Street 2: 104 Airport Dr., Suite 2200		
City: Chapel Hill	County: Orange	State: NC	
Province:	Country: USA	Zip/Postal Code: 27599-1350	
Project/Performance Site Congressional Districts: NC-004			

Additional Project/Performance Site Location			
Organizational Name: North Carolina State University			
DUNS: 042092122			
Street 1: Research Admin/ SPARCS	Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514		
City: Raleigh	County: Wake	State: NC	
Province:	Country: USA	Zip/Postal Code: 27695-7514	
Project/Performance Site Congressional Districts: NC-02			

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location		
Organizational Name: Duke University		
DUNS: 044387793		
Street 1: Hock Plaza	Street 2: Box 2716 Med Ct.	
City: Durham	County: Durham	State: NC
Province:	Country: USA	Zip/Postal Code: 27705
Project/Performance Site Congressional Districts: NC-004		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Cai, Jianwen	Jainwen_Cai	UNC-CH	Project 1 Leader
Blackwell, Kimberly L.		Duke University	Co-Investigator
Crawford, Jeffrey	crawf006	Duke University	Co-Investigator
George, Stephen L.	georg001	Duke University	Project Co-Leader
Goldberg, Righard M.	RICHARD_GOLDBERG	UNC-CH	Co-Investigator
Ibrahim, Joseph G	JOE_IBRAHIM	UNC-CH	Project Co-Leader
Jung, Sin-Ho	Jung0005	Duke University	Co-Investigator
Kosorok, Michael R	Michael_Kosorok	UNC-CH	Co-Investigator
Pang, Herbert	Oxbert	Duke University	Co-Investigator
Tsiatis, Anastasios	butch_tsiatis	NCSU	Co-Investigator
Wang, Xiofei	XIAOFEI.WANG	Duke University	Co-Investigator
Zeng, Donglin	Donglin_Zeng	UNC-CH	Co-Investigator
Zhang, H. Helen		NCSU	Co-Investigator
Zhou, Haibo	Haibo_Zhou	UNC-CH	Co-Investigator

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project

Human Embryonic Stem Cells ☒ **No** ☐ **Yes**

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

RESEARCH PLAN

1 INTRODUCTION TO RESUBMISSION/REVISION APPLICATION - N/A

2 SPECIFIC AIMS

Design is a crucial first step in clinical trials. Well-designed studies are essential for successful cancer research and cancer drug development. Innovative clinical trial designs can potentially require fewer patients, save resources, and accelerate cancer drug development. Although much effort has been put into analysis methods with complicated data structure, the design aspect has not kept pace. Hence developing statistical methods for innovative clinical trial design is timely and much needed.

In this project, we propose to develop new statistical methodology to address issues in the design and analysis of clinical trials. We will investigate both the analytical and the empirical behavior of the proposed methodologies. Related software will be developed. These high impact and innovative statistical methods will improve public health by enabling accurate and efficient estimation of sample size and power for studies with time-to-event and longitudinal endpoints, cluster randomized cancer prevention and therapeutic trials, and cancer drug development trials. Specifically, we have the following specific aims:

Aim 1: Develop methods for design and sample size calculation for longitudinal and joint models for longitudinal and survival data. We will consider a wide range of design issues for joint models. Specifically, we will undertake a methodological development of design issues, such as sample size and power considerations, for investigating the treatment effect on both the survival and longitudinal processes, and the effect of the longitudinal process on survival in a joint modeling setting using a two-stage modeling approach. We will develop statistical methods for sample size and power estimation for testing the overall treatment effect on survival, the effect of the longitudinal process on survival, and the direct treatment effect on survival. We will consider settings from the simple case involving a univariate longitudinal process and a univariate survival process to the very complex case involving multivariate longitudinal processes and multivariate survival processes. These high impact and innovative statistical methods will provide new and innovative methodology for clinical trials design using joint models and will provide an important set of data analysis and design tools for the practitioner. These methods will improve public health by enabling accurate and efficient estimation of sample size and power in the presence of time-to-event and longitudinal data using joint modeling approaches.

Aim 2: Develop statistical methodology for the design and analysis of group randomized cancer prevention trials with survival and recurrent event outcomes. We will consider design issues with clustered survival endpoint and clustered recurrent events endpoint from cluster-randomized trials. Specifically, we will consider sample size and power calculations for investigating the treatment effect on the survival process or recurrent event process. Test statistics for clustered recurrent event data will be considered and its asymptotic properties will be developed using modern empirical process theory. Permutation tests will also be considered when the number of clusters is small. We will also consider the situation where the elements of a cluster are randomized to different treatments. The strengths and weaknesses of different designs will be compared via theoretical investigations and simulation studies. These new methods will provide important data analysis and design tools and will improve community-based cancer research.

Aim 3: Develop statistical methodology for cancer drug development. The process for clinical development of anti-cancer agents is time-consuming and costly. Phase III clinical trials, the most costly and time-consuming type of clinical study in the development process, are required in order to demonstrate safety and efficacy in a specific setting and to receive regulatory approval. Unfortunately, only a minority of the phase III 'pivotal' trials conducted are successful in achieving regulatory approval for the tested therapy. In addition, the explosion of new agents requiring development, particularly molecularly targeted agents, has led to an increased need for efficiency in identifying and screening promising therapies and in conducting clinical trials of those therapies with a high probability of success. In this aim, we will address important statistical issues in the oncology drug development pathway, including three specific sub-aims in the areas of targeted designs, phase II designs, and partially randomized designs.

i. Develop methods for the design and analysis of clinical trials of targeted therapy. The explosion in the number of anti-cancer agents targeted to a particular biologic pathway has led to the need for a rethinking of the design of clinical trials using such agents. If the only patients who could benefit from such therapy are

those with the biologic target, if the assay for assessing the target is accurate, and if the prevalence is low, a 'targeted' design is much more efficient than an untargeted design in which all patients are eligible. However, these conditions do not always hold or may be uncertain. We will develop methods for alternative designs, including "enrichment" designs in which some, but not all, of the patients without the target are randomized, and compare the operating characteristics and costs of these designs to fully targeted designs. We will develop valid and efficient statistical methods for these trials by applying a semiparametric empirical likelihood approach.

ii. Develop designs for phase II trials that are predictive of phase III trial success. Phase II trials are an important step in the drug development process, designed to screen out unpromising therapies and to identify therapies to be tested further in phase III trials. The track record of phase II trials in predicting phase III success is not particularly goodet al., 2008; Zia et al., 2005; Joffe et al., 2004) . In addition, newer cytostatic agents produce different effects than the cytotoxic agents for which such trials were originally designed. Better approaches are needed. We will develop new methods for phase II trials, particularly randomized phase II trials, and phase II/III clinical trials, and assess their operating characteristics, costs, and predictive ability for subsequent phase III trials. We will also gather information on both combination and non-combination therapies in phase II studies and subsequent phase III studies to build prediction models using machine learning and other nonparametric classification methods.

iii. Develop methods for the design and analysis of partially randomized clinical trials. Randomization is a powerful technique in experimental design and the randomized phase III clinical trial represents the gold standard technique for comparing treatments. Unfortunately, randomization is often a serious impediment to patient accrual, particularly when there is a major difference in the types of treatments under study. One alternative to a fully randomized trial is a partially randomized trial in which some patients, those without treatment preferences, are randomized and others are assigned their treatment of choiceand Bradley 1989; Brocklehurst, 1997) . We will develop new semiparametric empirical likelihood methods for the design and analysis of such trials to adjust for selection bias and to improve efficiency. The performance of these methods will be compared with standard methods via simulation. In addition, we will work closely with collaborators on Aim 1 of Project 2 on semiparametric inverse weighting methods, a topic with potential application to partially randomized trials.

3 BACKGROUND AND SIGNIFICANCE

3.1 Specific Aim 1 - Design and Sample Size Calculation for Longitudinal and Joint Models for Longitudinal and Survival data

In many observational studies and randomized clinical trials, both time-to-event and longitudinal data are frequently simultaneously collected. For example, in many cancer clinical trials, both time to death (or relapse) and longitudinal quality of life (QOL) measures are collected. In such studies, there is a great interest in characterizing the relationship between time-to-event and the longitudinal process. For example, if a treatment has a significant effect on the longitudinal process, and the longitudinal process has a significant effect on survival time, then the longitudinal biomarkers are in the casual pathway and can potentially be used as a surrogate endpoint for the time-to-event.

Characterizing the treatment effect on the time-to-event and longitudinal response processes, and the effect of longitudinal process on survival are usually complicated by many challenges including missing and/or mismeasured longitudinal data (Tsiatis et al., 1995, Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2004) and informative censoring of survival data (Hogan and Laird, 1997). Failure to correctly deal with those challenges can lead to biased and inefficient estimators for the treatment effect on both the survival and longitudinal processes, and the effect of longitudinal response process on survival. Thus, naive approaches to inference on the treatment effect on both the survival and longitudinal processes, and the relationships between longitudinal and time-to-event data are inappropriate (Tsiatis and Davidian, 2004).

Joint models for survival and longitudinal data have recently become quite popular and have played an important role in cancer and AIDS clinical trials, where a longitudinal biologic marker such as CD4 cell count or immune response can be an important predictor of survival (Ibrahim, Chen, and Sinha, 2001). Joint modeling for survival and longitudinal data is an innovative framework that efficiently utilizes the intrinsic relationships between the time-to-event and the longitudinal processes, for example, by incorporating a trajectory function for the unobserved true longitudinal process into the hazard function of the time-to-event process (Wulfsohn and

Tsiatis, 1997; Henderson, Diggle, and Dobson, 2000; Lin et al., 2002). The others have considered joint models where the event time distribution and longitudinal data are taken to depend on a common set of latent random effects (Tsiatis and Davidian, 2004; Guo and Carlin, 2004). It has been shown theoretically and by simulations that joint modeling leads to unbiased estimators of the treatment effect on both survival and the longitudinal response process, and the effect of the longitudinal response process on survival. Furthermore, joint modeling approaches are very likely to provide more efficient estimators.

However, to the best of our knowledge, fundamental statistical design issues regarding sample size and power estimation have never been considered in the literature for joint models, potentially due to the very complex integrated joint likelihood involved in deriving sample size and power estimation formulas. In this aim, we will tackle this very important and high impact design problem. The statistical methods to be developed will be immediately applied to designing randomized clinical trials conducted in the Lineberger Comprehensive Cancer Center at the University of North Carolina at Chapel Hill (UNC-CH) and several Cancer and Leukemia Group B (CALGB) studies to answer important scientific questions, and will be disseminated to the broad community through publications in high-impact journals and development of user-friendly, publicly available software. Specifically, we will check design plans on the following current CALGB studies which involves both survival and longitudinal data collection:

1. CALGB 9221 - A randomized phase III controlled trial of subcutaneous 5-azacytidine versus observation in myelodysplastic syndromes (MDS). The quality of life (QOL) part of this study actually played a key role in FDA approval for 5-azacytidine in MDS;
2. CALGB 90401 - A randomized double-blinded placebo controlled phase III trial comparing docetaxel and prednisone with and without bevacizumab in men with hormone refractory prostate cancer;
3. CALGB 49907 - A randomized trial of adjuvant chemotherapy with standard regimens, cyclophosphamide, methotrexate and fluorouracil - (CMF) or doxorubicin and cyclophosphamide - (AC), versus capecitabine in women 65 years and older with node positive or node negative breast cancer.

3.2 Specific Aim 2 - Design and Analysis of Prevention and Therapeutic Trials

Because cancer has been difficult to treat, an important component in cancer research is to find effective ways to prevent cancer. In many cancer prevention trials, it is often of interest to examine group-administered interventions to determine if these interventions help to improve the health of the members in those groups on average. Many of the intervention strategies are designed to be delivered to groups of participants, where the groups are usually not constituted at random, for example, workers at worksites or students in classrooms. When study conditions (e.g., intervention and control) are randomly assigned to such groups, that is, the groups are used as the unit of assignment, these trials are referred to as *group-randomized* trials. They are also referred to as *cluster-randomized* trials. Because of the increased focus on disease prevention and the fact that many intervention programs are delivered in group settings, we have seen a great increase in the use of group-randomized trials.

One example for the group-randomized trial is the Community Intervention Trial for Smoking Cessation (COMMIT) (The COMMIT Research Group, 1995a, 1995b). The intervention strategies were designed to promote smoking cessation by using a wide range of community resources to approach individual smokers and to affect community attitudes and policies toward smoking. Twenty-two communities were paired up and randomized to receive the active intervention or control. In this 5-year trial, the benefit of community smoking cessation strategies were examined. Other examples of group-randomized trials include a study of the impact of vitamin A supplementation (Sommer et al., 1986), which randomized villages to study the effect of vitamin A supplementation for preschool children, and the Working Well Trial (WWT) (Abrams et al., 1994), which studied the effects of worksite interventions regarding smoking cessation and diet on employee health.

The design issues and analysis for group-randomized trials have been considered for complete continuous and categorical data by many authors. For example, Raudenbush (1997), Slymen and Hovell (1997), Murray (1998), Hayes and Bennett (1999), Kerry and Bland (2001), Lake et al. (2002), and Liu et al. (2002) considered the power analysis for group-randomized trials. Feng et al. (1996), Gail et al. (1996), Murray et al. (1996), and Braun and Feng (2001) considered analysis issues in group-randomized trials. Murray (1998), Donner and Klar (2000), and Murray et al. (2004) have provided comprehensive reviews on these issues.

In some of the prevention trials, survival times or recurrent event times are the outcome of interest. An example for survival time is the time of onset of smoking in smoking prevention trials. It is of interest to study if

intervention strategies help to delay the time of onset of smoking in such trials. An example of recurrent event is quitting smoking in smoking cessation trials. Such event could recur, since people could start to smoke again after quitting and then quit again after re-starting. Data from such trials are naturally clustered within the cluster and we refer to them as clustered survival data and clustered recurrent event data, respectively.

Another related design is the subunit randomized design. In a cancer clinical trial, we may randomize patients with multiple tumor sites between treatment arms and observe time to an event from each site. Or, we may assign different treatments among the tumor sites of each patient. The former belongs to the *cluster-randomized* trial mentioned above and the latter is called a *subunit randomization* trial. While there exists dependency only within each arm in cluster randomization trial data, there exists dependency both within and across arms in subunit randomization trial data.

A common feature of the data generated from such design is that the data within the group are correlated or clustered. Regression analysis approach has been considered by Lee et al. (1992) for clustered survival data and by Schaubel and Cai (2005a, 2005b) for clustered recurrent event data. Log-rank type of tests regression methods can be applied to compare the marginal survival distributions between arms, but their variance estimators need to be modified to account for the possible dependency in clustered survival data. Adjusted variance estimator for the log-rank test was proposed by Jung (1999) for the paired survival data and Jung (2007a) proposed an efficient simulation method to calculate the sample size based on the paired log-rank test. Although logrank-type of tests have been proposed for clustered survival data (Jung and Jeong, 2003; Jeong and Jung, 2006), design methods have been limited. Gangnon and Kosorok (2004) proposed for the weighted rank tests with general clustered data. However, the formulas do not clearly separate the contributions by joint survival distribution and censoring distribution to the final sample size. Jung (2007b) proposed a simulation-based sample size calculation method for clustered survival data. However, simulation-based method requires long computing time, does not show the direct relationship between the sample size and the input parameters, and requires specification of the full dimensional joint distributions. Therefore, it is desirable to develop closed form sample size formulas for clustered survival data from cluster randomized trials.

Among the methods for treatment comparisons of recurrent events data from independent individuals, a robust log-rank test proposed by Lawless and Nadeau (1995) is widely used. Based on the test by Lawless and Nadeau (1995), Cook (1995) and Matsui (2005) considered sample size calculations in this context via a nonhomogeneous Poisson process model. Their methods are parametric in the sense that, conditional on a frailty, the intensity of a homogeneous Poisson process is needed as an input parameter for sample size calculations. Hughes (1997) and Bernardo and Harrington (2001) considered power and sample size calculations based on a multiplicative intensity model and a marginal proportional hazards model respectively. In spite of these developments for recurrent event data from independent individuals, fundamental statistical design issues regarding sample size and power estimation have never been considered in the literature for the clustered recurrent endpoint for group-randomized trials. In this aim, we will tackle these very important and highly practical design problems. The resulting methods will provide much needed statistical tools for cancer prevention trials.

3.3 Specific Aim 3 - Statistical Methodology for Cancer Drug Development

The impetus for this aim is the need to streamline the process by which new anti-cancer agents obtain regulatory approval and thus become available for general use. The process is time-consuming, complicated and expensive (Malakoff, 2008; Fricker, 2008). There is no easy fix, but the sub-aims in this proposal address three important aspects of the process: designs for targeted therapy, phase II trials, and partially randomized trials. New methodology and improved efficiency in these areas can make a substantial impact on the process.

3.3.1 Targeted Therapy

The term 'targeted therapy' refers to drugs designed to interfere with a specific molecular target that is believed to play a critical role in tumor growth or progression, is not expressed significantly in normal cells, and is correlated with clinical outcome (Chon et al., 2006). The mechanisms of action and toxicities of targeted therapies differ substantially from those of traditional cytotoxic chemotherapy. With the rapid development of high-throughput techniques for identifying novel specific molecular targets in human cancer over the past few years, targeted cancer therapy has dramatically increased and is now used for many common malignancies, including breast, colorectal, lung, and pancreatic cancers; lymphoma; leukemia; and multiple myeloma. The

implications of this change in the types of therapy have yet to be fully incorporated into the design and analysis of cancer clinical trials.

For trials of targeted therapy, there are potential efficiency gains that can be achieved by limiting eligibility in trials to those patients who are known to have the target in question (Simon and Maitournam, 2004; Maitournam and Simon, 2005). There are important examples of situations in which such a strategy has worked well (e.g., trastuzumab in breast cancer) and in which the alternative, entering all patients, has failed (gefitinib in NSCLC). However, there are other examples (e.g., cetuximab in metastatic colorectal cancer) in which patients without the target appear to benefit as much as those with the target.

It is important to consider alternative designs and to understand the circumstances under which competing designs are most appropriate. One alternative to a targeted design is a biomarker stratified randomized (BSR) design, in which marker positivity is a stratification factor and both marker positive and negative patients are randomized to target agent versus placebo. A BSR design allows testing of whether the marker positive patients benefit from the targeted agent compared to placebo, testing an overall treatment benefit for all patients, and evaluating the performance of the predictive classifier in identifying targeted patients. Another alternative is an enrichment design, which lies somewhere between a targeted design and a BSR design. An enrichment design randomizes only a subset of marker negative patients in order to reduce cost and to improve study efficiency. Selecting which patients to randomize will depend on the biomarker prediction, other baseline patient characteristics, or an intermediate efficacy endpoint.

3.3.2 Phase II Designs

The primary purposes of a phase II trial are to identify promising treatment regimens for further testing in a subsequent phase III trial and to screen out unpromising regimens that do not warrant further testing. There is a vast literature on phase II trials. Although the earliest single-arm phase II clinical trials were also designed as single-stage trials (Gehan, 1961; Fleming, 1982; A'Hern, 2001), most phase II clinical trials are designed as multi-stage trials, ordinarily with a stopping rule for ineffectiveness at the early stages (Simon, 1987). Two-stage designs are commonly used because of simplicity and because of diminishing returns beyond two stages. One may also employ an upper boundary to stop the trial early when a significantly high efficacy is observed from stage 1 (Chang et al., 1987; Spiegelhalter et al., 1986). It is also possible to introduce decision rules allowing for an intermediate outcome (Storer, 1992). In recent years, randomized phase II trials have become increasingly employed because of the unreliability of historical data used in the design of single-arm trials and because of the use of other endpoints (e.g., progression-free survival) that require a concurrent control arm (Jung, 2008a; Jung and George, 2009; Piedbois, 2007; Redman and Crowley, 2007; Stadler, 2007; Tangen and Crowley, 2006; Lee and Feng, 2005; Rubinstein et al., 2005; Wieand, 2005; Rosner et al., 2002; Steinberg and Venzon, 2002; Simon et al., 1985).

For phase II trials with promising results, a subsequent phase III trial is the gold standard for assessing efficacy and is required for regulatory approval. But these trials are very expensive and time-consuming and there are an increasingly large number of new cancer agents and combinations of agents that need to be assessed. Most of the reported phase II trials in the medical literature are in oncology (Michaelis et al., 2007), but many phase III cancer trials have not led to improvements over standard (control) therapy, despite promising results in the phase II trials (Turrisi, III, 2005; El-Maraghi and Eisenhauer, 2008). Thus, the traditional phase I - II - III sequence of clinical trials, designed originally for cytotoxic agents, may not be the best strategy for the newer types of agents under development. Indeed, we may be doing the wrong type of phase II trials (Burton, 2007). It is increasingly important to design, conduct and analyze such trials with care (Mariani and Marubini, 2000; Ottaiano et al., 2007; Mariani and Marubini, 1996; Chang et al., 2005) and new approaches are needed (Rawlins, 2004).

In this sub-aim, we will do two things: (i) develop statistical models for predicting positive phase III trials based on the outcome of phase II trials; (ii) improve the standard two-stage design for single-arm phase II trials when the patient population is heterogeneous.

3.3.3 Partially Randomized Designs

Randomized clinical trials have been the cornerstone of cancer drug development. Randomization balances known and unknown prognostic factors among treatment groups, allows unbiased estimation of treatment effects and provides the basis for valid statistical inference. However, randomized clinical trials often have accrual

difficulties due to patient or physician preferences (i.e. a lack of equipoise), low adherence in the randomly assigned treatment group, and low generalizability due to enrollment bias. For example, CALGB 30102 is a phase III trial designed to compare the success rates of pleural effusion control by a talc/chest tube and a small PluerX catheter in lung cancer patients. PluerX is a small device used at home and a chest tube is an inpatient device requiring attention from the medical staff. This study was terminated due to the slow accrual rate after enrolling 40 patients. Patient or physician preference for one of the two treatments was obviously the primary reason for slow accrual and early termination. Similarly, CALGB 140503 is an ongoing randomized trial designed to test the non-inferiority of sublobar resection to lobectomy for small peripheral non-small cell lung cancer (NSCLC). It is also at risk of early termination due to slow accrual. The investigators are interested in alternative designs to improve accrual and to increase the generalizability of the results.

In the literature, alternative designs to the classic randomized clinical trial have been proposed to relieve or bypass randomization. One example is the randomized consent design (Zelen, 1990), in which a patient is randomized prior to the informed consent stage, the physician presents the treatment that has been selected and then asks for informed consent. However, this design has ethical, legal and scientific problems and is rarely used in practice. Another design is the non-randomized patient preference design, in which patients receive their preferred treatment and are followed prospectively for clinical outcome. This design may encourage patient accrual, but it introduces selection bias between treatment arms. Because the assignment of treatment is determined by both known and unknown factors, it is possible that the apparent effect of treatments is due primarily to these other factors in ways that cannot be fully accounted for in the analysis. There are other hybrid designs that combine features of randomization and patient preference of treatments (Lambert and Wood, 2000). One variant of the hybrid designs is the doubly randomized preference design (Rüker, 1989), where patients are randomized into a randomization arm, within which treatments are randomized; and a preference arm, within which treatments are chosen by patients. Long, Little, and Lin (2008) proposed a two-stage model for estimating the causal treatment effects and the preference effects for the doubly randomized preference design under certain assumptions. Their work also provides a framework to understand other hybrid designs.

We are particularly interested in partially randomized designs (Brewin and Bradley, 1989; King et al,. 2005). In one design of this type, the comprehensive cohort design, patients with preferences are offered their preferred treatment, while those without preferences are randomized. This design was originally proposed to increase the external validity of the randomized component of the trial. The advantage for such hybrid designs are fast accrual and inclusion of all patients with and without preference for treatment while still allowing unbiased inference of treatment effects in the set of patients receiving randomly assigned treatments. If it is truly believed that there is no systematic difference in baseline covariates in the treatment groups for the non-randomized patients, a joint efficient analysis of randomized and non-randomized patients is possible. In order to adjust for potentially unbalanced covariates in the non-randomized patients and to improve efficiency through auxiliary covariates in the randomized patients, we will estimate the constrained covariate distribution by the empirical likelihood method. A consistent and efficient estimator for treatment effect for all patients can be obtained using the constrained covariate distribution. Our method requires the assumption that the treatment selection of a patient is independent of his or her clinical outcome, given the observed covariates. We will evaluate the impact of violating this assumption on the estimation and the testing of treatment effect using nonignorable treatment assignment model. In Project 2, Aim 1, semiparametric inverse weighting methods will also be developed for data arising from the partially randomized design. We will evaluate via simulation the performance of these semiparametric methods.

4 PRELIMINARY STUDIES

4.1 Investigators

The study team for the proposed project is highly qualified. This section summarizes the relevant experience of the investigators.

Jianwen Cai of UNC-CH will serve as Project Leader and will in addition lead Aim 2. Dr. Cai has extensive experience in longitudinal data analysis and joint modeling of longitudinal and survival data (Zeng and Cai, 2005a, 2005b), study designs (Cai and Zeng, 2004, 2007), multivariate survival analysis (Cai and Prentice, 1995, 1997; Cai et al. 2005, 2007, 2008), and clinical trials (Shen and Cai, 2003), making her ideally suited to lead the project. Her work with colleagues (Schaubel and Cai, 2005a; Song, Kosorok, and Cai, 2008) forms the

basis for the research proposed in Aim 2.

Stephen George of Duke University, the Duke PD/PI for the overall Program Project, will serve as Project Co-Leader and will in addition lead Aim 3. Dr. George has extensive expertise in the methodology of clinical trials (George, 2009; Potthoff and George, 2009) and translational science (George, 2007, 2008).

Richard M. Goldberg of UNC-CH is the Chief of the Division of Hematology and Oncology and Distinguished Professor of Medicine at UNC-CH. Dr. Goldberg is also the Physician-in-Chief of the North Carolina Cancer Hospital. Dr. Goldberg has extensive experience in cancer care and cancer research. Dr. Goldberg's clinical interest includes colorectal and other GI cancers, carcinoid tumors, new drug development, and inherited predisposition to GI cancers. Dr. Goldberg's research interest focuses on four areas: 1) clinical studies in patients with GI cancers (principally colorectal cancer); 2) translational studies done on biologic specimens from patients with cancer done with laboratory collaborators; 3) development of new cancer drugs and drug combinations; and, 4) clinical trials methodology (Goldberg et al., 2004; Hoskins, Goldberg, and McLeod, 2007; Dy et al., 2007; O'Neil et al., 2007). Dr. Goldberg will provide his expertise in the subject matter and provide advice on the interpretation of statistical results to the research team.

Joe Ibrahim of UNC-CH, the UNC-CH co-PD/PI for the overall Program Project, will serve as Project Co-Leader and will in addition lead Aim 1. Dr. Ibrahim has extensive experience in statistical methodological development for joint models of longitudinal and time-to-event data (Brown and Ibrahim, 2003a, 2003b, 2005; Ibrahim, Chen, and Sinha, 2004; Brown, Ibrahim, and DeGruttola, 2005; Chi and Ibrahim, 2006, 2007).

Sin-Ho Jung of Duke University is a co-investigator. Dr. Jung's expertise includes statistical methods for clustered survival data (Jung, 1999, 2008b; Jung and Jeong, 2003) and design and analysis of phase II cancer clinical trials (Jung, 2008a; Jung, Carey, and Kim, 2001; Jung et al., 2004).

Michael Kosorok of UNC-CH, the PD/PI for the overall Program Project, is a co-investigator for this project. Dr. Kosorok's expertise includes clinical trials (Kosorok, Shi, and DeMets, 2004), time-to-event data (Eng and Kosorok, 2005), empirical processes and semiparametric inference (Kosorok, 2008).

Herbert Pang of Duke University is a co-investigator for the project. Dr. Pang's expertise and research interests include classification methods (Pang et al., 2006; Pang, Kim and Zhao, 2008), genomics (Pang and Zhao, 2008), and shrinkage-based discriminant analysis (Pang, Tong and Zhao, 2009).

Anastasios A. Tsiatis of NCSU is a co-investigator for the project. Dr. Tsiatis has extensive experience in semiparametric theory (Tsiatis, 2006), survival analysis and clinical trials (Tsiatis et al., 2008), joint modeling of longitudinal and survival data (Tsiatis and Davidian, 2004), and causal inference (Davidian et al., 2005; Tsiatis and Davidian, 2007; Cao et al., 2009).

Xiaofei Wang of Duke University is a co-investigator for the project. Dr. Wang is involved in design and analysis of cancer biomarker studies, such as CALGB 30203, 30506 and 30801, in which targeted design and biomarker stratification design have been used. He is also the faculty statistician for CALGB 30102 and 140503, which are used as examples to motivate the partially randomized design. He has expertise in semiparametric inference for data arising from outcome/auxiliary-dependent sampling (Wang and Zhou, 2006) and subsampling (Wang and Zhou, 2009; Wang, Wu, and Zhou, 2009).

Donglin Zeng of UNC-CH is a co-investigator for the project. Dr. Zeng's expertise includes semiparametric inference with event time data (Zeng and Lin, 2007), joint modelling (Zeng and Cai, 2005), missing data (Chen, Zeng, and Ibrahim, 2007), and study design (Cai and Zeng, 2004; 2007).

Hao (Helen) Zhang of NCSU is a co-investigator for the project. Dr. Zhang's expertise is in nonparametric smoothing (Zhang et al. 2004; Lin and Zhang, 2006), variable selection (Zhang, 2006; Zhang and Lu, 2007), and statistical machine learning and high dimensional data analysis (Zhang et al., 2006; Zou and Zhang, 2008).

Haibo Zhou of UNC-CH is a co-investigator for the project. Dr. Zhou is an expert on cost-effective study designs especially the outcome-dependent sampling design and inference (Zhou et al., 2002, Weaver and Zhou 2005, Zhou et al 2007). He is also an expert on measurement error problems (Zhou and Pepe, 1995; Zhou and Wang, 2000, Zhou and You, 2007), survival analysis (Jiang and Zhou, 2007), epidemiological methods and environmental statistics (Zhou and Weinberg, 1999; Zhou et al., 2009).

There have been very active intra-institutional collaborations. Drs. Kosorok and Cai have collaborated on sample size calculation for recurrent event data (Song, Kosorok, Cai, 2008); Drs. Kosorok and Zhou on outcome dependent sampling (Song, Zhou, Kosorok, 2009); Drs Cai and Zeng on joint modeling of longitudinal and survival data (Zeng and Cai, 2005a, 2005b) and sample size calculations for case-cohort studies (Cai and Zeng, 2004, 2007); Drs. Zhou and Cai on partially linear models (Cai et al., 2007, 2008) and measurement

error problems (Liu, Zhou, and Cai, 2009); Drs. Ibrahim and Zeng on missing data problem (Chen, Zeng, and Ibrahim, 2007); Drs. George and Jung on clinical trial method (Jung and George, 2009); and Drs. Wang and Pang on methods for cancer biomarker (Wang, Pang, and Schwartz, 2009).

In addition to the intra-institutional collaboration, there has also been some inter-institutional collaboration among the investigators from the three institutions. The investigators on this project have all given seminars in the other two institutions. Dr. Cai from UNC-CH had meetings with Dr. Jung from Duke to discuss issues related to multivariate survival analysis and possible collaboration. She has also had discussion with Dr. George from Duke about issues related to data monitoring. Dr. Wang from Duke and Dr. Zhou from UNC-CH have collaborated on research for methods for outcome dependent sampling. This collaboration has resulted in two publications (Wang and Zhou, 2006, 2009) and one manuscript under review (Wang, Wu, and Zhou, 2009).

4.2 Preliminary Studies

4.2.1 Specific Aim 1 - Design and Sample Size Calculation for Longitudinal and Joint Models for Longitudinal and Survival data

The study team has extensive experience in statistical methodological development for joint models of longitudinal and time-to-event data (Brown and Ibrahim, 2003a, 2003b; Ibrahim, Chen and Sinha, 2004; Tsiatis and Davidian, 2004; Brown, Ibrahim, and DeGruttola, 2005; Zeng and Cai, 2005a, 2005b; Chi and Ibrahim, 2006, 2007). these models to the design setting. For subject i, $(i = 1, \ldots, n)$, let T_i and C_i denote the event and censoring times, respectively; $S_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. Let Z_i be a treatment indicator, and $X_i(u)$ be the longitudinal process of the longitudinal markers at time $u \geq 0$ (also referred as the trajectory). In a very general case, Z_i can be a q-dimensional vector of baseline covariates including the treatment indicator. For ease of exposition, Z_i denotes the treatment indicator here. Values of $X_i(u)$ are measured intermittently at times $u = t_{ij} \leq S_i, j = 1, \ldots, m_i$, for subject i. Let $Y_i(t_{ij})$ denote the observed value of $X_i(t_{ij})$ at time t_{ij}, which is subject to measurement error. The joint modeling approach links two sub-models, one for the longitudinal process $X_i(u)$ and one for the event time T_i, e.g., by including the trajectory in the hazard function of T_i. Thus,

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta X_i(t) + \alpha Z_i\} \quad (1)$$

Although other types of joint models for $X_i(u)$ have been proposed (Henderson et al., 2000; Wang and Taylor 2001), here we focus on a general polynomial model,

$$X_i(u) = \theta_{0i} + \theta_{1i}u + \theta_{2i}u^2 + \cdots + \theta_{pi}u^p + \gamma Z_i \quad (2)$$

where $\boldsymbol{\theta_i} = \{\theta_{0i}, \theta_{1i}, \ldots, \theta_{pi}\}^T$ is identically and independently distributed as a multivariate normal distribution with mean $\boldsymbol{\mu_\theta}$ and covariance matrix $\boldsymbol{\Sigma_\theta}$. The parameter γ is a fixed treatment effect. Here, we only consider the main effect of γZ_i in equation (2) to estimate the treatment effect on longitudinal process averaged over time, which is, in general, a reasonable assumption. Alternatively, one can include the interaction term between Z_i and u in equation (2) to allow time dependent linear treatment effect on the longitudinal process (in this case, we do not need to include the main effect of γZ_i since the treatment effect is usually considered to be zero at baseline), which may be questionable. However, at the design stage, some functional assumptions on the treatment effect on the longitudinal process are needed to develop methods for sample size and power estimation. The polynomial term of $\sum_{k=0}^{p}\theta_{ki}u^k$ can be easily generalized to nonparametric forms such as B-splines. The model in equation (2) can be easily extended to include other covariates at baseline. The observed longitudinal measurements are modeled as $Y_i(t_{ij}) = X_i(t_{ij}) + e_{ij}$, where the e_{ij} are identically and independently distributed as $N(0, \sigma_e^2)$, and $\text{Cov}(e_{ij}, e_{ij'}) = 0$, for $j \neq j'$. The observed data likelihood for subject i is,

$$L_i = \int_{-\infty}^{\infty}\left[\prod_{j=1}^{m_j} f(Y_{ij}|\boldsymbol{\theta_i}, \gamma, \sigma_e^2)\right] f(\theta_i|\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta}) f(S_i, \Delta_i|\boldsymbol{\theta_i}, \beta, \gamma, \alpha) d\boldsymbol{\theta_i} \quad (3)$$

Figure 1 shows the underlying causal diagram for the models that we will consider here for design considerations. We will focus on design issues for testing the overall treatment effect on survival $\beta\gamma + \alpha$; the effect of the longitudinal process on survival β; and the direct treatment effect on survival α. We will not consider design issues for testing the direct treatment effect on the longitudinal process γ in this aim since it has been already well studied (Diggle et al,. 2002). As a first step, let us consider the two-step inferential approach developed by Tsiatis et al. (1995) based on a first-order approximation, $E[f(X(t), \beta|\bar{Y}(t), S \geq t)] \approx f[E(X(t)|\bar{Y}, S \geq t, \beta)]$

where $\bar{Y}(t)$ denotes the observed longitudinal data up to time t. Under this approximation, we can replace $\{\theta_{0i}, \theta_{1i}, \ldots, \theta_{pi}\}^T$ in the Cox model with the empirical estimates $\{\hat{\theta}_{0i}, \hat{\theta}_{1i}, \ldots, \hat{\theta}_{pi}\}^T$ described by Laird and Ware (1982). The Cox partial likelihood (Cox 1975) can then be applied to obtain parameter estimates and inferences without the need to use the complex integrated full joint likelihood as in equation (3). Despite several drawbacks (Wulfsohn and Tsiatis 1997), the two-stage modeling approach has two major advantages: (1) The separate likelihood is simple to maximize and standard statistical software for longitudinal and survival data can be used directly; (2) It can provide nearly unbiased estimators (i.e., the biases associated with this simple approach are negligible) and correct biases caused by missing and/or mismeasured longitudinal data, and/or informative censoring of survival data.

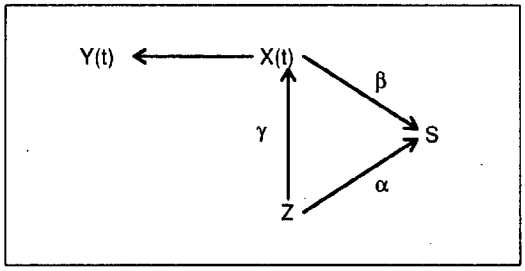


Figure 1: Causal Diagram

Bias of estimating the direct treatment effect on survival if ignoring the longitudinal trajectory. When a treatment has an effect on the longitudinal process (i.e., $\gamma \neq 0$ in equation (2)) and the longitudinal process is associated with survival (i.e., $\beta \neq 0$ in equation (2)), the overall treatment effect on the time-to-event is $(\beta\gamma + \alpha)$. Thus, it is obvious that ignoring the longitudinal process in the proportional hazard model would result in a biased estimator of the treatment effect on survival. When the longitudinal process is not associated with the treatment (i.e, $\gamma = 0$ in equation (2)), it is not easy to understand that ignoring the longitudinal trajectory in the proportional hazard model would result in an attenuated estimator of the hazard ratio for the treatment effect on survival (i.e., bias towards the null). This attenuation is known in the econometrics literature as the attenuation due to unobserved heterogeneity (Horowitz 1999; Abbring and Van den Berg, 2007). Suppose that the true trajectory is known and it is a linear function of time, $X_i(u) = \theta_{0i} + \theta_{1i}u$, where $Z_i = \{0, 1\}$ is the treatment indicator. If event time follows an exponential distribution, the likelihood for subject i is

$$L_i = \{\lambda_0 \exp[\beta(\theta_{0i} + \theta_{1i}S_i) + \alpha Z_i]\}^{\Delta_i} \exp\left\{\frac{-\lambda_0 \exp(\beta\theta_{0i} + \alpha Z_i)[\exp(\beta\theta_{1i}S_i) - 1]}{\beta\theta_{1i}}\right\}. \quad (4)$$

The maximum likelihood estimate of the hazard ratio for the treatment effect on survival takes the form

$$\exp(\hat{\alpha}) = \frac{\left\{\sum_{i=1}^{N} W_i I(Z_i = 0)\right\}\left\{\sum_{i=1}^{N} Z_i I(\Delta_i = 1)\right\}}{\left\{\sum_{i=1}^{N} W_i I(Z_i = 1)\right\}\left\{\sum_{i=1}^{N} I(\Delta_i = 1) - \sum_{i=1}^{N} Z_i I(\Delta_i = 1)\right\}} \quad (5)$$

where $W_i = \frac{\exp(\beta\theta_{0i})\{\exp(\beta\theta_{1i}S_i) - 1\}}{\beta\theta_{1i}}$, which shows that the hazard ratio is dependent on β. Table 1 below shows estimates of the hazard ratio using different models based on some preliminary simulation studies using different β's.

4.2.2 Specific Aim 2 - Design and Analysis of Prevention and Therapeutic Trials

The investigators in this project have done extensive work in methodological development for survival data, especially multivariate survival data such as correlated failure times and recurrent event times. With respect to correlated failure time data, Cai and Prentice (1995, 1997) considered the marginal proportional hazards model with distinguishable and common baseline hazards, respectively, and propose a weighted method to improve the efficiency of the estimator, which took the correlation of subjects within the same cluster into consideration. They showed that the efficiency is improved with the weighted estimator when the correlation is high and the

Table 1: Effect of β on the Estimation of Direct Treatment Effect on Survival (α)

	$\lambda_i(t) = \lambda_0(t)\exp(\alpha Z_i)$		$\lambda_i(t) = \lambda_0(t)\exp\{\beta(\theta_{0i} + \theta_{1i})t + \alpha Z_i\}$	
β	$\exp(\hat{\alpha})$ under exponential model	$\exp(\hat{\alpha})$ under Cox model	$\exp(\hat{\alpha})$ under exponential model	$\exp(\hat{\alpha})$ under Cox model
0	0.67	0.67	0.67	0.66
0.5	0.75	0.72	0.67	0.67
1	0.80	0.79	0.67	0.67

[a] Note: $\exp(\hat{\alpha})$ is the average value based on 1000 simulated trials, each with 200 subjects per arm. Minimum follow-up time is set to be 0.75 year (9 months), and maximum follow-up time is set to be 2 years. The hazard for the time to event follows equation (2) with constant baseline hazard $\lambda_0 = 0.85$, and the true direct treatment effect on survival $\alpha = -0.4$ (i.e., HR = 0.67).

censoring percentage is small. regression parameter varies with time, Cai et al. (2007b) studied the varying coefficient model for multivariate failure times. Yin and Cai (2004) considered additive hazards model for multivariate failure time data. With respect to recurrent event data, Pepe and Cai (1993) proposed a rate model conditioning on the covariates and the numbers of recurrences prior to time t to analyze the recurrent events. The additive rates model was also considered by Schaubel et al. (2006). Multiple type recurrent event data were analyzed by Cai and Schaubel (2004). Schaubel and Cai (2005a) considered the proportional rates models for clustered recurrent event data. Statistical methods were developed to analyze multivariate failure time data under case-control and case-cohort studies (Cai, Qaqish, and Zhou, 2001; Kong, Cai, and Sen, 2004; 2006; Kang and Cai, 2008; Kong and Cai, 2008).

The study team also has extensive experience in methods for study design. Shen and Cai (2003) proposed statistical methodology to calculate the sample size and power for survival data in clinical trials. Cai and Zeng (2004; 2007) used log-rank type statistics to calculate the sample size and power under case-cohort study with rare events and non-rare events, respectively. Jung (1999) proposed to use the two-sample weighted rank tests for comparing the marginal distributions of paired survival data. He modified the variance estimators of the rank tests to account for possible dependence between paired survival variables. Jung (2007a) proposed a simulation-based sample size method and Jung (2008b) proposed a closed form sample size formula based on these tests for paired survival data. Jung and Jeong (2003) proposed the weighted rank test for clustered survival data under cluster randomization and Jeong and Jung (2006) under subunit randomization. Jung (2007b) proposed a simulation-based sample size method based on the log-rank test by Jung and Jeong (2003). Gangnon and Kosorok (2004) proposed closed form sample size formulas for the weighted rank tests with general clustered data. Recently, Song, Kosorok, and Cai (2008) proposed a sample size calculation formula for recurrent event data using robust covariate-adjusted logrank statistics.

In summary, the study team has extensive experience in working in the areas of methods for dealing with correlated failure time data and statistical issues in study designs. Their work and experience form the basis of this aim and make the team to be highly qualified for the proposed work.

4.2.3 Specific Aim 3 - Statistical Methodology for Cancer Drug Development

4.2.3.1 Targeted Therapy

A recent review paper was published on statistical methods in translational cancer research (George, 2008), in which considerations for targeted designs were discussed. In particular, it was argued that the efficiency of a targeted design in a trial with a time-to-event endpoint may be measured by the time to reaching a final answer, not solely in terms of the number of patients accrued.

We developed a semiparametric inference procedure for data from studies conducted with a two-component sampling scheme where both a simple random sample and multiple outcome- or outcome/auxiliary-dependent samples are observed (Wang and Zhou, 2006). This sampling scheme allows the investigators to oversample certain subpopulations believed to have more information about the regression model while still gaining insights about the underlying population through the simple random sample. The proposed method applies to both binary and multicategorical outcome data and allows an arbitrary link function in the framework of generalized linear models. Simulation studies showed that the proposed estimator has good small sample properties. The techniques are relevant to the hybrid targeted design that we propose to investigate in this project. In fact,

the example used in the paper was an ongoing study to assess the association between the mutation level of epidermal growth factor receptor (EGFR) and the antitumor response to EGFR-targeted therapy among non-small cell lung cancer patients. In a related but different design, we (Wang and Zhou, 2009) developed an estimated likelihood method for data arising from outcome/auxiliary-dependent subsampling with a kernel smoother to utilize the information from continuous auxiliary variables.

4.2.3.2 Phase II Designs

We defined a family of designs, called admissible designs (Jung et al., 2004; Jung et al., 2001), generalizing the popular minimax and optimal designs (Simon, 1989). A widely used graphical computer program was developed to facilitate the identification of all admissible designs. When there exist no or very small historical control data, a single-arm phase II trial based on a hypothetical or unreliable parameter estimates can lead to wrong conclusions. In this setting, we have proposed optimal and minimax designs for randomized phase II trials (Jung, 2008a).

In multi-stage phase II clinical trials, the ordinary maximum likelihood estimator (MLE) is biased. We derived uniform minimum variance unbiased estimators (UMVUE) for such trials (Jung and Kim, 2004) and demonstrated that, unlike the MLE-based confidence interval (CI), the UMVUE-based CI does not require specification of the critical value at the stopping stage. Due to this property, we can calculate an exact UMVUE-based CI adjusting for the multiple stages even when the number of patients at the stopping stage is different from the one by the original design. Further, in publishing phase II study results, investigators often report whether the experimental drug is accepted or not, but do not report how strong the evidence is to support the final decision. We proposed (Jung et al., 2006) to calculate a p-value to this end. We show that the UMVUE-based p-value is unbiased and does not require specification of the critical value at the stopping stage.

If the primary endpoint of a phase II trial is a censored variable, such as time to progression, the two most popular analysis methods are (1) MLE based on exponential distribution and (2) nonparametric method for median time. The first requires a parametric assumption and both are based on large sample approximation. However, due to the small sample size of phase II trials, these methods do not control the type I error. We proposed using a dichotomous endpoint based on a fixed time point and controlling the type I error using normal distributions (Owzar and Jung, 2008).

When there are multiple therapies available for testing in a phase II trial, one approach is to evaluate each therapy separately in a single-arm phase II trial. When multiple arms are accepted through the separate evaluations, we proposed a testing method to compare multiple arms accepted by the individual multistage randomized phase II trials (Jung and George, 2009).

4.2.3.3 Partially Randomized Designs

The constrained empirical likelihood is a nonparametric approach for the use of side information in the form of a known statistical functional. See Owen (2001) for a comprehensive review on the subject. Qin and Lawless (1994) studied on the case where the number of estimating equations exceeds the number of parameters, i.e. the overdetermined case. The constrained empirical likelihood method has been successfully used in incorporating auxiliary information about the underlying population (Wang et al., 2008) and in the biased sampling problem (Wang and Zhou, 2006). In a partially randomized design, we would like to force balance of covariates distribution for the non-randomized patients in order to reduce bias. We also add the balance of the covariate distribution in the randomized patients as a side condition to improve efficiency. The idea of reducing bias by balancing covariate distribution is related to using a propensity score as a covariate in regression analysis for causal inference on treatment effect (Lunceford and Davidian, 2004). Tsiatis et al. (2008) developed a new semiparametric method for covariate adjustment that separates modeling of covariate relationships from estimation of the treatment effect.

5 RESEARCH DESIGN AND METHODS

5.1 Specific Aim 1 - Design and Sample Size Calculation for Longitudinal and Joint Models for Longitudinal and Survival data

5.1.1 Plan

In this aim, we will undertake a methodologic development of design issues for investigating the treatment effect on the longitudinal and survival processes, and the effect of the longitudinal process on survival in a joint

modeling setting based on the two-stage approach described above. Extensive simulations will be conducted to investigate the tentatively proposed methods for sample size and power estimation. There are four scientific goals for this aim. First, we will develop statistical methods for sample size and power estimation for the overall treatment effect on survival $\beta\gamma + \alpha$. Second, we will develop statistical methods for sample size and power estimation for the effect of the longitudinal process on survival β. Third, we will develop statistical methods for sample size and power estimation for the direct treatment effect on survival α. Last, we will extend those methods to address complex design issues in settings involving multivariate longitudinal and multivariate survival processes. The methods to be developed will be disseminated to the broad biomedical research community through publications in high-impact journals and development of user-friendly, publicly available software.

5.1.2 Statistical Methods for Sample Size and Power Estimation for the Overall Treatment Effect on Survival ($\beta\gamma + \alpha$)

When the longitudinal trajectory is a linear function of time, we have derived a sample size and power formula for testing the overall treatment effect, which generalizes the sample size formula developed by Schoenfeld (1983). The number of events required for a one-sided level α^* test with power $1 - \beta^*$ can be estimated by the following formula when the hazard follows (1) and the longitudinal trajectory is linear function of time:

$$D = \frac{(z_{\beta^*} + z_{1-\beta^*})^2}{p_1(1-p_1)(\beta\gamma + \alpha)^2}, \tag{6}$$

where p_1 is the proportion of patients assigned to treatment 1 ($Z_i = 1$). Properties of the random effects in the longitudinal trajectory do not play a role in the sample size and power estimation for the overall treatment effect on survival at the design stage. However, correct assumptions must be made with regard to the overall treatment effect ($\beta\gamma + \alpha$). Limited simulation studies, presented in Table 2 with different β's based on 1000 simulated trials with 100 subjects per simulation per arm show that the Schoenfeld's formula works approximately well in joint modeling approaches when the primary objective is to investigate the overall treatment effect on survival when the longitudinal trajectory is a linear function of time. In this aim, we will generalize the derived formula to a very general setting where the longitudinal process is modeled by a general polynomial model as in equation (2) and conduct a large number of simulations to investigate the performance of Schoenfeld's sample size formula as in equation (6).

Table 2: Validation of Schoenfeld's Sample Size Formula for Testing the Overall Treatment Effect on Survival $\beta\gamma + \alpha$ by Simulations

				Power to Estimate Overall Treatment Effect on Survival $\beta\gamma + \alpha$	
β	$Var(\theta_{0i})$	$Var(\theta_{1i})$	$Cov(\theta_{0i}, \theta_{1i})$	Empirical	Calculated
0.15	0.5	0.9	0	50.5	50.0
0.15	0.8	1	0	49.3	50.0
0.15	0.8	1	0.5	47.0	50.0
0.2	1.2	0.7	0	51.4	52.4
0.2	0.7	1.2	0	51.9	52.4
0.2	0.7	1.2	0.2	49.8	52.4
0.2	0.7	1.2	-0.2	49.6	52.4

[a] Note: The trajectory and variance-covariance of $(\theta_{0i}, \theta_{1i})$ were assumed known. Minimum follow-up time is 0.75 year (9 months), and maximum follow-up time is 2 years. Event time simulated with exponential distribution with $\lambda_0 = 0.85$, $\alpha = -0.3$, $\gamma = -0.1$, $E(\theta_{0i}) = 0$, and $E(\theta_{1i}) = 3$.

5.1.3 Statistical Methods for Sample Size and Power Estimation for the Effect of Longitudinal Process on Survival (β)

When the longitudinal trajectory follows a general polynomial function of time as in equation (2), we can derive the number of events required for a one-sided significance level α^* with power $1 - \beta^*$ by the following formula:

$$D = \frac{(z_{\beta^*} + z_{1-\alpha^*})^2}{\sigma_s^2 \beta^2} \tag{7}$$

Table 3: Validation of Tentatively Derived Formula (7) for Testing the Effect of Longitudinal Process on Survival (β) by Simulations

				Statistical Power to Estimate the Effect of Longitudinal Process on Survival β	
β	$\text{Var}(\theta_{0i})$	$\text{Var}(\theta_{1i})$	$\text{Cov}(\theta_{0i}, \theta_{1i})$	Empirical	Calculated
0.15	0.5	0.9	0	41.6	41.7
0.15	0.8	1.0	0	52.9	54.0
0.15	0.8	1.0	0.5	66.1	69.1
0.20	1.2	0.7	0	87.1	86.8
0.20	0.7	1.2	0	75.9	78.7
0.20	0.7	1.2	0.2	82.7	84.6
0.20	0.7	1.2	-0.2	69.8	70.9

[a] Note: The trajectory and variance-covariance of $(\theta_{0i}, \theta_{1i})$ were assumed known. Minimum follow-up time is 0.75 year (9 months), and maximum follow-up time is 2 years. Event time simulated with exponential distribution with $\lambda_0 = 0.85$, $\alpha = -0.3$, $\gamma = -0.1$, $E(\theta_{0i}) = 0$, and $E(\theta_{1i}) = 3$.

where $\sigma_s^2 = \mathbf{S}_{\bar{t}}{}^T \hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}} \mathbf{S}_{\bar{t}}$, $\mathbf{S}_{\bar{t}} = \left\{1,\ E[I(T \le \bar{t})T],\ E[I(T \le \bar{t})T^2],\ \ldots,\ E[I(T \le \bar{t})T^p]\right\}^T$, and $\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}$ is the covariance matrix of $\{\hat{\theta}_{0i}, \hat{\theta}_{1i}, \ldots, \hat{\theta}_{pi}\}^T$, and $\bar{t}$ is the average follow-up time on all subjects. Thus, the power for estimating the effect of the longitudinal process on survival (β) depends on: (a) The expected log hazard ratio associated with a unit change in the longitudinal trajectory, i.e., the expected size of β. As the expected β increases, the required sample size decreases; (b) The covariance matrix $\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}$. A larger variance and positive covariance leads to a smaller required sample size, while smaller variances and larger negative covariances require larger sample sizes; and (c) the truncated moments of survival time, T. To obtain the truncated moments of T, we need to assume a distribution function for T. Suppose that T follows an exponential distribution with parameter η, then

$$E[I(T \le \bar{t})T^p] = \int_0^{\bar{t}} T^p \eta \exp(-\eta T) dT = \frac{1}{\eta^p}\gamma(p+1, \bar{t}), \tag{8}$$

where $\gamma(p+1, \bar{t})$ is a lower incomplete gamma function. The exponential parameter η can be estimated if we know the median survival or the median event time of the study population, $\eta = -\log(0.5)/T_M$, where T_M is the median survival time. When the trajectory is a linear function of time,

$$\sigma_s^2 = \text{var}(\hat{\theta}_{0i}) + E[I(T \le \bar{t})T^2]\text{var}(\hat{\theta}_{1i}) + 2E[I(T \le \bar{t})T]\text{cov}(\hat{\theta}_{0i}, \hat{\theta}_{1i}). \tag{9}$$

Both $E[I(T \le \bar{t})T^2]$ and $E[I(t \le \bar{t}T]$ have closed-form solutions:

$$E[I(T \le \bar{t})T^2] = \int_0^{\bar{t}} T^2 \eta \exp(-\eta T) dT = \frac{2}{\eta^2} - \exp(-\eta\bar{t})(\bar{t}^2 + \frac{2\bar{t}}{\eta} + \frac{2}{\eta^2}), \text{ and}$$

$$E[I(T \le \bar{t})T] = \int_0^{\bar{t}} T\eta \exp(-\eta T) dT = \frac{1}{\eta} - \exp(-\eta\bar{t})(\bar{t} + \frac{1}{\eta}).$$

In this aim, we will develop methods to obtain the truncated moments of T using a more general parametric distribution for survival time, such as the generalized gamma distribution which contains all four of the most common types of hazard functions: monotonically increasing and decreasing, as well as bathtub and arc-shaped hazards (Cox et al., 2007). If we have a known longitudinal trajectory, $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ can be used directly in $\sigma_s^2 = \mathbf{S}_{\bar{t}}{}^T \hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}} \mathbf{S}_{\bar{t}}$ replacing $\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}$. When $\boldsymbol{\theta}_i$ is unknown with known or unknown $\mathbf{\Sigma}_{\boldsymbol{\theta}}$, the trajectory is characterized by the empirical Bayes estimates of $\boldsymbol{\theta}_i$. $\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}$ is associated with the number of data collection points (m_i) and the correlation between Y_{ij} and $Y_{ik}(j \ne k)$ of the longitudinal data. Table 3 shows that the power estimated by formula (7) agrees with the empirical power in the simulated data with a linear trajectory based on 1000 simulated trials with 100 subjects per arm.

Efficient Design with Optimal Data Collection Strategy. Although we may estimate the sample size and power by assuming a known longitudinal trajectory with known parameters as shown above, we need to

investigate how the longitudinal data measurements (i.e., the number of measurements, the intervals between measurements etc.) affect $\hat{\Sigma}_{\theta}$ and the sample size and power estimation at the design stage. This would allow us to efficiently design a data collection strategy to maximize the statistical power of a study. Let $\mathbf{R_i} = \begin{pmatrix} 1 & t_{i1} & \dots & {t_{i1}}^p \\ 1 & t_{i2} & \dots & {t_{i2}}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{im_i} & \dots & {t_{im_i}}^p \end{pmatrix}$ be a $m_i \times (1+p)$ matrix, and $\mathbf{Z_i} = 1_{m_i}\, Z_i$, $\text{Var}(\mathbf{Y_i}) = \mathbf{V_i} = \mathbf{I_{m_i}}\sigma_e^2 + \mathbf{R_i}\mathbf{\Sigma_\theta}\mathbf{R_i}^T$ and $\mathbf{W_i} = \mathbf{V_i}^{-1}$, then $\hat{\theta}_i$ and $\hat{\Sigma}_{\theta}$ can be expressed as the following (Laird and Ware 1982):

$$\hat{\theta}_i = \mathbf{\Sigma_\theta}\mathbf{R_i}^T\mathbf{W_i}(\mathbf{Y_i} - \hat{\gamma}\mathbf{Z_i}) \tag{10}$$

$$\text{Var}(\hat{\theta}_i) = \hat{\Sigma}_{\theta} = \mathbf{\Sigma_\theta}\mathbf{R_i}^T\left\{\mathbf{W_i} - \mathbf{W_i}\mathbf{Z_i}\left(\prod_i^N \mathbf{Z_i}^T\mathbf{W_i}\mathbf{Z_i}\right)^{-1}\mathbf{Z_i}^T\mathbf{W_i}\right\}\mathbf{R_i}\mathbf{\Sigma_\theta}. \tag{11}$$

We will extend the derived formulas as in equation (2) to develop sample size and power estimation formulas for studies with longitudinal data collected at irregular time points using the relationships listed above for testing the effect of the longitudinal process on survival (β). It thus will allow users to pick a data collection strategy that is optimal. We will conduct a large number of simulations to investigate the performance of the newly proposed sample size formula as in equation (7) and the to-be-derived sample size and power estimation formula when the longitudinal data are collected at irregular time points.

5.1.4 Statistical Methods for Sample Size and Power Estimation for the Direct Treatment Effect on Survival (α)

When the longitudinal trajectory is a linear function of time, the derived sample size and power formula for testing the direct treatment effect generalizes the sample size formula developed by Schoenfeld (1983). The number of events required for a one-sided level α^* test with $1-\beta^*$ power can be estimated by the following formula when the hazard follows (1) and the longitudinal trajectory is linear function of time:

$$D = \frac{(z_{\beta^*} + z_{1-\beta^*})^2}{p_1(1-p_1)\alpha^2}, \tag{12}$$

where p_1 is the proportion of patients assigned to treatment 1 ($Z_i = 1$). Properties of the random effects in the trajectory do not play a role in the sample size and power estimation for the direct treatment effect on survival (α) at the design stage. We will generalize the derived formula above to a very general setting when the longitudinal process is modeled by a general polynomial model as in equation (2) and conduct a large number of simulations to investigate the performance of Schoenfeld's sample size formula as in equation (12).

5.1.5 Statistical Methods for Sample Size and Power Estimation for Other Key Issues in Joint Modeling of Longitudinal and Survival Data

In this aim, we will derive sample size and power calculation formulas for the very general setting where studies collect multivariate longitudinal and multivariate survival data.

We will extend the methods proposed in Sections (**5.1.2**) (**5.1.3**) and (**5.1.4**) to test the overall treatment effect on survival ($\beta\gamma + \alpha$), the effect of the longitudinal process on survival (β) and the direct treatment effect on survival (α) when the longitudinal data are collected at irregular time points. In the very general setting for the joint model for multivariate longitudinal and multivariate survival data (Chi and Ibrahim, 2007), let $Y_{ik}(t_{ij})$ be an assessment of the kth indicator of the longitudinal data process for the ith patient at time t_{ij} and $X_{ik}(t_{ij})$ be the corresponding unobserved trajectory representing its true value, where $k = 1, \dots, K$, $j = 1, \dots, m_i$, and $i = 1, \dots, n$. The multivariate longitudinal model for $Y_{ik}(t_{ij})$ is given by

$$Y_{ik}(t_{ij}) = X_{ik}(t_{ij}) + \epsilon_{ijk}, \tag{13}$$

where ϵ_{ijk} represents a measurement error. Again, we consider a general $p-$dimensional polynomial model for the longitudinal process. The trajectory function for the k^{th} longitudinal process is modeled as

$$X_{ik}(u) = \theta_{0ki} + \theta_{1ki}u + \theta_{2ki}u^2 + \cdots + \theta_{pki}u^p + \gamma_k Z_i, \tag{14}$$

where $\boldsymbol{\theta_{ik}} = \{\theta_{0ki}, \theta_{1ki}, \ldots, \theta_{pki}\}^T$. The parameter γ_k is a fixed treatment effect on the k^{th} longitudinal process. We assume

$$\boldsymbol{\epsilon_{ij\cdot}} \overset{i.i.d.}{\sim} N_K(0, \boldsymbol{\Psi_k}), \quad \boldsymbol{\theta_{ik}} \overset{ind}{\sim} N(\boldsymbol{\mu_{k\theta}}, \boldsymbol{\Sigma_{k\theta}}),$$

and $\boldsymbol{\epsilon_{ij\cdot}}$ is independent of $\boldsymbol{\theta_{ik}}$, where $\boldsymbol{\epsilon_{ij\cdot}} = \{\epsilon_{ij1}, \ldots, \epsilon_{ijK}\}^T$. Therefore, the measurement errors of the longitudinal process observed at the same time may be related to each other, but are independent among observations assessed at different times. The structure of $\boldsymbol{\Psi_k}$ characterizes the association between the longitudinal process indicators measured at the same time, and is assumed to be common across time and patients. To incorporate information from both the longitudinal trajectories $X_k(t)$, $k = 1, \ldots, K$, and the baseline covariates Z (including treatment indicator) in our multivariate survival model, we let the m^{th} hazard function, i.e., $\lambda_m(t)$, depend on all covariates through a proportional hazard model as

$$\lambda_m(t) = \lambda_{0m}(t) \exp\left[\sum_{k=1}^{q} \beta_{mk} X_k(t) + \alpha_m Z\right], \tag{15}$$

for $m = 1, 2, \ldots, M$. To tackle the design issues of this very important and challenging statistical question, we will extend our proposed methodology discussed above to:

1. develop statistical methods for sample size and power estimation for studies with multivariate longitudinal data and univariate survival data.

2. develop statistical methods for sample size and power estimation for studies with multivariate longitudinal data and multivariate survival data.

We will conduct a large number of simulations to investigate the performance of our sample size and power estimation formulas for the above cases using a two-step approach (Tsiatis et al., 1995). We will compare the results to the single-step joint modeling approach by directly maximizing the joint integrated likelihood. If there is a substantial difference, we will develop sample size and power estimation formulas using a single-step joint modeling approach, which involves complex multidimensional integration over random effects and is quite mathematically and computationally challenging. However, given the advantages of the two step approach, we would expect that the sample size and power estimation formulas should perform reasonably well.

5.1.6 Evaluation by Simulation Study

Carefully designed simulation scenarios will be used to evaluate the proposed sample size and power estimation formulas in joint models. We will utilize Core C for assistance in developing code for these simulation studies. One special challenge for joint models — and especially for multivariate longitudinal and multivariate survival data — is that the analysis of simulated studies can be very computationally expensive. For multivariate longitudinal data, we will generate it from multivariate normal or log normal distributions. For the generalization of multivariate survival data, we will consider multivariate exponential distributions (Marshall and Olkina, 1967) and multivariate Weibull distributions (Hougaard, 1986; Nadarajah and Kotz, 2006) with appropriate parameters as the input for baseline hazard functions and then use equation (15) to incorporate the effect of longitudinal processes on survival. We will consider both uninformative uniform censoring and informative censoring. We will carefully evaluate both the accuracy of theoretical predictions as well as performance of the proposed methods under a broad range of simulation scenarios.

5.2 Specific Aim 2 - Design and Analysis of Prevention and Therapeutic Trials

Our objective in Specific Aim 2 is to develop sample size formula and power estimation for cluster randomized trials. We will derive closed form sample size formula for clustered survival data from cluster randomized trials and subunit randomized trials, propose test statistic and investigate its asymptotic properties for clustered recurrent event data, and derive sample size formula for clustered recurrent event data.

5.2.1 Clustered Survival Time Under Cluster Randomization

5.2.1.1 Review of the Log-Rank Test for Clustered Survival Time

Suppose that n_k clusters are randomized to arm $k(= 1, 2)$. For cluster $i(= 1, ..., n_k)$ in arm k, let m_{ki} denote the number of units, called cluster size, and $(T_{kij}, j = 1, ..., m_{ki})$ their survival times. We assume that units

within each cluster are exchangeable in the sense that these units have a common marginal survivor function $S_k(t) = P(T_{kij} \geq t)$, cumulative hazard function $\Lambda_k(t) = -\log S_k(t)$, and joint survivor function $S_k(t_1, t_2) = P(T_{kij} \geq t_1, T_{kij'} \geq t_2)$ for $1 \leq j \neq j' \leq m_{ki}$.

In conjunction with survival time T_{kij}, let C_{kij} be the censoring time. From units in cluster i of group k, we observe $\{(X_{kij}, \Delta_{kij}), j = 1, ..., m_{ki}\}$, where $X_{kij} = T_{kij} \wedge C_{kij}$, $\Delta_{kij} = I(T_{kij} \leq C_{kij})$ and $a \wedge b = \min(a, b)$. We assume that $(C_{kij}, j = 1, ..., m_{ki})$ are independent of $(T_{kij}, j = 1, ..., m_{ki})$.

In order to test if two treatment arms have the same treatment effects, i.e.

$$H_0 : \Lambda_1(t) = \Lambda_2(t) \text{ for all } t \geq 0 \qquad \text{vs.} \qquad H_a : \Lambda_1(t) \neq \Lambda_2(t) \text{ for some } t \geq 0,$$

we consider the class of rank statistics

$$W = \sqrt{n} \int_0^\infty H(t)\{d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t)\},$$

where $n = n_1 + n_2$, $\hat{\Lambda}_k(t) = \int_0^t Y_k^{-1}(s) dN_k(s)$ is the Nelson's estimator based on the clustered survival data, $N(t) = N_1(t) + N_2(t)$, $N_k(t) = \sum_{i=1}^{n_k} N_{ki}(t)$, $N_{ki}(t) = \sum_{j=1}^{m_{ki}} N_{kij}(t)$, $N_{kij}(t) = \Delta_{kij} I(X_{kij} \leq t)$, $Y(t) = Y_1(t) + Y_2(t)$, $Y_k(t) = \sum_{i=1}^{n_k} Y_{ki}(t)$, $Y_{ki}(t) = \sum_{j=1}^{m_{ki}} Y_{kij}(t)$ and $Y_{kij}(t) = I(X_{kij} \geq t)$. Here, H is a function of bounded variation that converges in probability to function h. For the log-rank test, we use $H(t) = n^{-1} Y_1(t) Y_2(t) / Y(t)$.

Jung and Jeong (2003) show that, under H_0, W is asymptotically normal with mean 0 and and variance σ^2 that can be consistently estimated under H_0 by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n_1} \hat{\varepsilon}_{1i}^2 + \frac{1}{n} \sum_{i=1}^{n_2} \hat{\varepsilon}_{2i}^2,$$

where $\hat{\varepsilon}_{ki} = \sum_{j=1}^{m_{ki}} \int_0^\infty \frac{Y_{3-k}(t)}{Y(t)} d\widehat{M}_{kij}(t)$, $\widehat{M}_{kij}(t) = N_{kij}(t) - \int_0^t Y_{kij}(s) d\hat{\Lambda}(s)$, and $\hat{\Lambda}(t) = \int_0^t Y^{-1}(s) dN(s)$.

We can reject H_0 when the absolute value of $W/\hat{\sigma}$ is larger than $z_{1-\alpha/2}$, the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

5.2.1.2 Sample Size Formula

Now we want to derive the required number of clusters $n(= n_1 + n_2)$ for a new study. The key component of a sample size formula is to describe σ^2 as a function of joint survival distributions and a censoring distribution. Although we do not have to specify the marginal and joint distribution functions in analysis, we need them in sample size calculation. We assume common censoring for subunits within each cluster, under which we have $C_{kij} = C_{ki}$. This assumption can be easily loosened. Let $G(t) = P(C_{ki} \geq t)$ denote the survivor function of the common censoring times within each cluster which will be specified by the accrual and additional follow-up periods.

In order to simplify the discussions, we assume constant cluster size $m_{ki} = m$, but the results can be extended to variable cluster size cases, see Jung et al. (2001) in clustered binary data case. Also, we limit our discussion to the log-rank test, but all the results can be modified for the general weighted rank-tests.

Let $a_k = n_k/n$ denote the allocation proportion for arm k, and $y_k(t) = S_k(t)G(t)$ the limit of $Y_k(t)/(mn_k)$. Let $M_k(t) = \sum_{i=1}^{n_k} \sum_{j=1}^{m} M_{kij}(t)$ and $M_{kij}(t) = N_{kij}(t) - \int_0^t Y_{kij}(s) d\Lambda_k(s)$. We can show that for large n

$$W \approx n^{-1/2} \left(\sum_{i=1}^{n_1} \sum_{j=1}^{m} \epsilon_{1ij} - \sum_{i=1}^{n_2} \sum_{j=1}^{m} \epsilon_{2ij} \right) + m a_1 a_2 \omega \sqrt{n} \tag{16}$$

where $\epsilon_{kij} = \int_0^\infty \frac{a_{3-k} S_{3-k}(t)}{a_1 S_1(t) + a_2 S_2(t)} dM_{kij}(t)$ and $\omega = \int_0^\infty \frac{S_1(t) S_2(t) G(t)}{a_1 S_1(t) + a_2 S_2(t)} \{d\Lambda_1(t) - d\Lambda_2(t)\}$.

Let $f_k(t_1, t_2)$ denote the bivariate probability density function of the survival times of subunits in arm k. Also let $\lambda_k(t) = d\Lambda_k(t)/dt$, $\lambda_k(t_1, t_2) = f_k(t_1, t_2)/S_k(t_1, t_2)$, $s_{k(1)}(t_1, t_2) = dS_k(t_1, t_2)/dt_1$, and $\lambda_{k(1|2)}(t_1|t_2) = -s_{k(1)}(t_1, t_2)/S_k(t_1, t_2)$. We define $s_{k(2)}(t_1, t_2)$ and $\lambda_{k(2|1)}(t_2|t_1)$ similarly. Then from (16), under H_a, we will show that W is approximately normal with mean $m a_1 a_2 \omega \sqrt{n}$ and variance

$$\sigma^2 = m[a_1\{\sigma_1^2 + (m-1)c_1\} + a_2\{\sigma_2^2 + (m-1)c_2\}],$$

where

$$\sigma_k^2 = \lim_{n\to\infty} \frac{1}{mn_k} \sum_{i=1}^{n_k} \sum_{j=1}^{m} \text{var}(\epsilon_{kij}) = a_k a_{3-k}^2 \int_0^\infty \frac{S_k(t) S_{3-k}(t)^2 G(t)}{\{a_1 S_1(t) + a_2 S_2(t)\}^2} d\Lambda_k(t)$$

$$c_k = \lim_{n\to\infty} \frac{2}{n_k m(m-1)} \sum_{i=1}^{n_k} \sum\sum_{1\le j<j'\le m} \text{Cov}(\epsilon_{kij}, \epsilon_{kij'})$$

$$= a_{3-k}^2 \int_0^\infty \int_0^\infty \frac{S_{3-k}(t_1) S_{3-k}(t_2) S_k(t_1, t_2) G(t_1 \vee t_2)}{\{a_1 S_1(t_1) + a_2 S_2(t_1)\}\{a_1 S_1(t_2) + a_2 S_2(t_2)\}} dA_k(t_1, t_2)$$

$$dA_k(t_1, t_2) = \{\lambda_k(t_1, t_2) - \lambda_{k(1|2)}(t_1|t_2)\lambda_k(t_2) - \lambda_{k(2|1)}(t_2|t_1)\lambda_k(t_1) + \lambda_k(t_1)\lambda_k(t_2)\} dt_1 dt_2.$$

Hence we can obtain the required sample size for power $1-\beta$ by

$$n = \frac{\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{(m a_1 a_2 \omega)^2}. \tag{17}$$

As expected, we can see that n decreases as a_1 gets close to $1/2$ and m increases. It can be also shown that σ^2 and n decrease as the dependency in survival times within each cluster increases. If subunits within each cluster are independent, the sample size formula reduces to a standard formula for the log-rank test with independent observations, e.g., Schoenfeld (1983). For specified marginal and bivariate joint survival distributions and accrual and additional follow-up periods, we can calculate ω and σ^2 as shown above. The sample size formula requires specification of the joint survival distributions only up to the second dimension. The sample size calculation procedure may be summarized as follows.

Sample Size Calculation

1. Specify input parameters
   (a) Type I error probability α and power $1-\beta$
   (b) Marginal and bivariate joint distributions under H_a, e.g. exponential distributions $S_k(t)$ with hazard rates λ_k and joint exponential distribution $S_k(t_1, t_2)$ defined by the marginal distribution and a copula
   (c) Accrual period a (or, accrual rate r) and additional follow-up period b which determine a censoring distribution of $U(b, a+b)$
   (d) Cluster size m (or, the distribution of m_{ki})
   (e) Allocation proportions a_1, a_2
2. Calculate ω and σ^2 (i.e., σ_k^2 and c_k for $k = 1, 2$) using numerical integrations
3. Calculate n by (17)

5.2.1.3 Design and Analysis with Small Number of Clusters and Large Cluster Size

In some studies, the number of clusters are fixed but the number of participants can be very large. In such case, studying the asymptotic behavior of the test statistic W as $m \to \infty$ will be desirable. Under the situation when the number of clusters is finite but the cluster size m goes to infinity, we will use the empirical process for dependent data to establish the asymptotic property of the testing statistic W. Particularly, we assume that for each k and i, $(X_{kij}, \Delta_{kij}), j = 1, ..., m$ is a stationary sequence but we allow different covariances for different clusters and arm k. We will apply the large sample results from Dehling et al. (2002) to show that the process $(N_{ki}(t), Y_{ki}(t), k = 1, 2, i = 1, ..., n_k)$ converges in distribution to a Gaussian process. The asymptotic covariance functions of the limiting process will be obtained using the results in Dehling et al. (2002). The asymptotic distribution of W will be derived using the functional delta method. Sample size formula will be developed based on the asymptotic approximation.

5.2.2 Clustered Recurrent Events data Under cluster Randomization

5.2.2.1 Log-rank Test for Clustered Recurrent Events Data

Suppose that subject $j(j = 1, \cdots, n_{ki})$ in cluster $i(i = 1, \cdots, n_k)$ is randomized to treatment group $k(k = 1, 2)$ and its lth recurrent event time is denoted by T_{kijl}. The censoring time is C_{kij}. The underlying counting process is $N^0_{kij}(t) = \sum_{l=1}^{\infty} I(T_{kijl} \leq t)$ and the observed counting process is $N_{kij}(t) = \sum_{l=1}^{\infty} I(T_{kijl} \leq t, C_{kij} \leq t)$. Let $Y_{kij} = I(C_{kij} \leq t)$ be the at-risk process. Let $\mu_k(t) = E(N^0_{kij}(t))$ be the cumulative mean function across the clusters under treatment group k.

In order to test if two treatment groups have the same effect on the recurrences of the event, we consider the following hypothesis:

$$H_0 : \mu_1(t) = \mu_2(t) \text{ for all } t \geq 0 \text{ vs. } H_a : \mu_1(t) > \mu_2(t) \text{ for all } t \geq 0.$$

The logrank-type statistics we consider take the following form:

$$L_n = \frac{1}{\sqrt{n}} \int_0^{\infty} \frac{Y_1(t)Y_2(t)}{Y_1(t) + Y_2(t)} \left\{ \frac{dN_1(t)}{Y_1(t)} - \frac{dN_2(t)}{Y_2(t)} \right\},$$

where $n = n_1 + n_2, N_k(t) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_{ki}} N_{kij}(t)$ and $Y_k(t) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_{ki}} Y_{kij}(t)$.

We will show that, as $n \to \infty$, under H_0, L_n asymptotically follows a zero-mean normal distribution with variance σ^2 that can be consistently estimated by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n_1} \left(\sum_{j=1}^{n_{ki}} \hat{\varepsilon}_{1ij} \right)^2 + \frac{1}{n} \sum_{i=1}^{n_2} \left(\sum_{j=1}^{n_{ki}} \hat{\varepsilon}_{2ij} \right)^2,$$

where $\hat{\varepsilon}_{kij} = \int_0^{\infty} \frac{Y_{3-k}(t)}{Y(t)} d\widehat{M}_{kij}(t)$ with $\widehat{M}_{kij}(t) = N_{kij}(t) - \int_0^t Y_{kij}(s) d\hat{\mu}(s)$, $\hat{\mu}(t) = \int_0^t Y^{-1}(s) dN(s)$, $N(t) = N_1(t) + N_2(t)$, and $Y(t) = Y_1(t) + Y_2(t)$. We reject H_0 if $L_n/\hat{\sigma} > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of the standard normal distribution.

5.2.2.2 Sample Size Calculation

To calculate the required sample size for a given power we need to investigate the asymptotic distribution of L_n under the alternative hypothesis H_a as $n \to \infty$. Hereafter, we consider the case

$$H_a : \ \mu_2(t) = \mu_1(t) \exp(-\phi(t)/2) < \mu_1(t),$$

where $\phi(t) = O(n^{-1/2})$ and $\phi(t) > 0$ for all t.

We assume constant cluster size $n_{ki} = m$, $n_k/n \to \rho_k$, and $\frac{1}{mn_k} \sum_{i=1}^{n_k} \sum_{j=1}^{m} Y_{kij}(t) \to \pi_k(t)$. From the modern theory of empirical processes, we will show that under H_a :

$$\begin{aligned} L_n &= \frac{1}{\sqrt{n}} \int_0^{\infty} \frac{\rho_2 \pi_2(t)}{\rho_1 \pi_1(t) + \rho_2 \pi_2(t)} dM_1(t) - \frac{1}{\sqrt{n}} \int_0^{\infty} \frac{\rho_1 \pi_1(t)}{\rho_1 \pi_1(t) + \rho_2 \pi_2(t)} dM_2(t) \\ &\quad - \sqrt{n}/2 \rho_1 \rho_2 m \int_0^{\infty} \frac{\pi_1(t)\pi_2(t)}{\rho_1 \pi_1(t) + \rho_2 \pi_2(t)} d(\mu_1(t)\phi(t)) + o_p(1) \\ &\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n_1} \sum_{j=1}^{m} \varepsilon_{1ij} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n_2} \sum_{j=1}^{m} \varepsilon_{2ij} - \sqrt{n}/2 \rho_1 \rho_2 m \omega + o_p(1), \end{aligned}$$

where $\varepsilon_{kij} = \int_0^{\infty} \frac{\rho_{3-k} \pi_{3-k}(t)}{\rho_1 \pi_1(t) + \rho_2 \pi_2(t)} dM_{kij}(t)$, $M_{kij}(t) = N_{kij}(t) - \int_0^t Y_{kij}(u) d\mu_k(u)$, and $\omega = \int_0^{\infty} \frac{\pi_1(t)\pi_2(t)}{\rho_1 \pi_1(t) + \rho_2 \pi_2(t)} d(\mu_1(t)\phi(t))$.

We will show that, as $n \to \infty$, L_n asymptotically converges to a normal distribution with mean $-\sqrt{n}/2 \rho_1 \rho_2 m \omega$ and variance

$$\begin{aligned} \sigma_a^2 &= m \sum_{k=1}^{2} \left(\rho_k \lim_{n \to \infty} \frac{1}{mn_k} \sum_{i=1}^{n_k} \sum_{j=1}^{m} \text{Var}(\varepsilon_{kij}) \right) \\ &\quad + m(m-1) \sum_{k=1}^{2} \left(\rho_k \lim_{n \to \infty} \frac{1}{m(m-1)n_k} \sum_{i=1}^{n_k} \sum_{1 \leq j < j' \leq m} \text{Cov}(\varepsilon_{kij}, \varepsilon_{kij'}) \right). \end{aligned}$$

Hence, the required sample size for power $1-\beta$ is given by

$$n = \frac{4\sigma_a^2(z_{1-\alpha}+z_{1-\beta})^2}{(\rho_1\rho_2 m\omega)^2}.$$

5.2.2.3 Simulation Study

We will conduct extensive simulation studies to evaluate the small sample performance of our proposed test. We will utilize Core C for assistance in designing and implementing these simulation studies. The $(l+1)$th recurrent event time $T_{kij,l}$ will be generated from the recursive formula

$$T_{kij,l+1} = T_{kij,l} - \log(1-U_{kij,l+1})(Q_i R_j m_{0k})^{-1}, \quad l = 0,1,2,\cdots$$

where $T_{kij,0}=0$, $U_{kij,l+1}$ is generated from uniform distribution on $(0,1)$, R_j from Gamma$(\sigma_R^{-2},\sigma_R^{-2})$, Q_i from Gamma$(\sigma_Q^{-2},\sigma_Q^{-2})$, and m_{0k} is the underlying recurrent rate for treatment group k. Note that R_j induces positive correlation among subjects within a cluster and Q_i induces correlation among the within-subject event times. The magnitude of the within-cluster and within-subject correlation increases with increasing σ_R^2 and σ_Q^2, respectively. We will consider σ_R^2=0, 0.25, 0.5 and σ_Q^2=0, 0.5, 1.0, since, in most practical situations, the within-subject correlation will be greater than between-subject correlation. Let $m_{0k}=k/2$.

We will consider the situation for the number of clusters to be 20, 50, and 100, and the cluster size to be 2, 5, 10, 20, 50, 100. We will investigate the empirical type I error rate and empirical power for the proposed test statistics for various combination of number of clusters, cluster size, within-cluster correlation, and within-subject correlation.

5.2.2.4 Permutation Test

The aforementioned inference relies on asymptotic approximation as $n\to\infty$. However, in situations when the number of clusters are very small, the asymptotic approximation might not work well. We will consider permutation test in the case of very small number of clusters. Suppose there are $n=n_1+n_2$ clusters and randomly select n_1 of n clusters receiving treatment group 1, so there are $R\equiv\frac{(n_1+n_2)!}{n_1!n_2!}$ different allocations. Let L_n^r be the corresponding logrank statistics under the rth permutation and the order statistics are denoted by

$$L_n^{(1)} \le \cdots \le L_n^{(R)}.$$

We still consider the following hypothesis:

$$H_0 : \mu_1(t) = \mu_2(t) \texttt{ for all } t\ge 0 \texttt{ vs. } H_a : \mu_1(t) > \mu_2(t) \texttt{ for all } t \ge 0.$$

H_0 is rejected if the originally observed L_n is greater than the $\widetilde{z}_{1-\alpha}\equiv L_n^{(100(1-\alpha))}$.

In order to compute the sample size for given power $1-\beta$, we consider the following alternative hypothesis:

$$H_a: \ \mu_2(t) = \mu_1(t)\exp(-\phi(t)/2) < \mu_1(t),$$

where $\phi(t)=O(n^{-1/2})$ and $\phi(t)>0$ for all t.

We will show that under H_a :

$$\begin{aligned} L_n &= \frac{1}{\sqrt{n}}\int_0^\infty \frac{Y_2(t)}{Y_1(t)+Y_2(t)}dM_1(t) - \frac{1}{\sqrt{n}}\int_0^\infty \frac{Y_1(t)}{Y_1(t)+Y_2(t)}dM_2(t) \\ &\quad -\frac{1}{\sqrt{n}}\int_0^\infty \frac{Y_1(t)Y_2(t)}{Y_1(t)+Y_2(t)}d(\mu_1(t)-\mu_2(t)) \\ &\equiv \frac{1}{\sqrt{n}}\int_0^\infty \frac{Y_2(t)}{Y_1(t)+Y_2(t)}dM_1(t) - \frac{1}{\sqrt{n}}\int_0^\infty \frac{Y_1(t)}{Y_1(t)+Y_2(t)}dM_2(t) - \sqrt{n}\widetilde{\omega}, \end{aligned}$$

where

$$\widetilde{\omega} = \frac{1}{n}\int_0^\infty \frac{Y_1(t)Y_2(t)}{Y_1(t)+Y_2(t)}d(\mu_1(t)-\mu_2(t)).$$

The required sample size n for power $1-\beta$ is then

$$n = \frac{(\widetilde{z}_{1-\alpha} + z_{1-\beta})^2}{\widetilde{\omega}^2}.$$

We will conduct simulation studies under the situations described in Section 5.2.2.3 with the number of clusters being 2, 5, 10 to investigate the empirical type I error rate and power. We will utilize Core C for assistance in designing and implementing these simulation studies. We will also investigate the effect of various combination of the number of clusters, cluster size, within-cluster correlation, and within-subject correlation on the power of the test.

5.2.3 Clustered Survival Data Under Subunit Randomization

5.2.3.1 Review of the Log-rank Test

Suppose that cluster $i(=1,...,n)$ has m_i subunits, of which m_{ik} are assigned to arm $k(=1,2)$, i.e. $m_{i1}+m_{i2}=m_i$. Let $(T_{ik1},...,T_{ikm_{ik}})$ be survival times for subunits in treatment k. Since subunits within a cluster share common characteristics, their survival times $T_{i11},...,T_{i1m_{i1}},T_{i21},...,T_{i2m_{i2}}$ tend to be positively correlated.

We assume that within treatment group k, $(T_{ikj}, 1\le i\le n, 1\le j\le m_{ik})$ are marginally identically distributed with cumulative hazard function $\Lambda_k(t)$. We want to test the same hypotheses H_0 vs. H_a as in the cluster randomization case.

For subunits in cluster i, let $C_{i11},...,C_{i1m_{i1}},C_{i21},...,C_{i2m_{i2}}$ be censoring times. We assume that the censoring times are independent of survival times within each cluster. The resulting clustered survival data consist of

$$\{(X_{ikj},\Delta_{ikj}), 1\le i\le n; k=1,2; 1\le j\le m_{ik}\},$$

where $X_{ikj}=\min(T_{ikj},C_{ikj})$ and $\Delta_{ikj}=I(T_{ikj}\le C_{ikj})$.

The log-rank statistic is

$$W=\sqrt{n}\{\int_0^\infty \frac{Y_2(t)}{Y(t)}d\hat{\Lambda}_1(t)-\int_0^\infty \frac{Y_1(t)}{Y(t)}d\hat{\Lambda}_2(t)\},$$

where $\hat{\Lambda}_k(t)=\int_0^t Y_k^{-1}(s)dN_k(s)$, $N_k(t)=\sum_{i=1}^n N_{ik}(t)$, $N_{ik}(t)=\sum_{j=1}^{m_{ik}} N_{ikj}(t)$, $N_{ikj}(t)=\Delta_{ikj}I(X_{ikj}\le t)$, $Y(t)=Y_1(t)+Y_2(t)$, $Y_k(t)=\sum_{i=1}^n Y_{ik}(t)$, $Y_{ik}(t)=\sum_{j=1}^{m_{ik}} Y_{ikj}(t)$, $Y_{ikj}(t)=I(X_{ikj}\ge t)$. Let $N(t)=N_1(t)+N_2(t)$ and $Y(t)=Y_1(t)+Y_2(t)$.

By Jeong and Jung (2006), under H_0, W is asymptotically normal with mean 0 and variance σ^2 that can be consistently estimated by $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$, where $\hat{\epsilon}_i=\int_0^\infty \frac{Y_2(t)}{Y(t)}d\widehat{M}_{i1}(t)-\int_0^\infty \frac{Y_1(t)}{Y(t)}d\widehat{M}_{i2}(t)$, $d\widehat{M}_{ik}(t)=dN_{ik}(t)-Y_{ik}(t)d\widehat{\Lambda}(t)$, and $\widehat{\Lambda}(t)=\int_0^t Y(t)^{-1}dN(t)$. We reject H_0 when the absolute value of $W/\hat{\sigma}$ is larger than $z_{1-\alpha/2}$, the $100(1-\alpha/2)$ percentile of the standard normal distribution.

5.2.3.2 Sample Size Formula

We assume common censoring for subunits within each cluster, under which we have $C_{kij}=C_i$. This assumption can be easily loosened. Let $G(t)=P(C_i\ge t)$ denote the survivor function of the common censoring times within each cluster which will be specified by the accrual and additional follow-up periods.

In order to simplify the discussions, we assume constant cluster size m of which $m_{ik}=m_k$ are assigned to arm k ($m_1+m_2=m$). All the results will be extended to variable cluster size cases. Also, we limit our discussion to the log-rank test. All the results can be modified for the general weighted rank-tests.

Let $a_k=m_k/m$ denote the allocation proportion for arm k, and $y_k(t)=S_k(t)G(t)$ the limit of $Y_k(t)/(m_kn)$. Let $f_k(t_1,t_2)$ denote the bivariate probability density function of the survival times of subunits in arm k. Also let $\lambda_k(t)=d\Lambda_k(t)/dt$, $\lambda_k(t_1,t_2)=f_k(t_1,t_2)/S_k(t_1,t_2)$, $s_{k(1)}(t_1,t_2)=dS_k(t_1,t_2)/dt_1$, and $\lambda_{k(1|2)}(t_1|t_2)=-s_{k(1)}(t_1,t_2)/S_k(t_1,t_2)$. We define $s_{k(2)}(t_1,t_2)$ and $\lambda_{k(2|1)}(t_2|t_1)$ similarly. Define

$$dA_k(t_1,t_2)=\{\lambda_k(t_1,t_2)-\lambda_{k(1|2)}(t_1|t_2)\lambda_k(t_2)-\lambda_{k(2|1)}(t_2|t_1)\lambda_k(t_1)+\lambda_k(t_1)\lambda_k(t_2)\}dt_1dt_2.$$

Let $f_{12}(t_1,t_2)$ and $S_{12}(t_1,t_2)$ denote the bivariate probability density function and survivor function, respectively, of the survival times of two subunits, one in arm 1 and the other in arm 2. Also let $\lambda_{12}(t_1,t_2)=f_{12}(t_1,t_2)/S_{12}(t_1,t_2)$, $s_{12(1)}(t_1,t_2)=dS_{12}(t_1,t_2)/dt_1$, and $\lambda_{12(1|2)}(t_1|t_2)=-s_{12(1)}(t_1,t_2)/S_{12}(t_1,t_2)$. We define $s_{12(2)}(t_1,t_2)$ and $\lambda_{12(2|1)}(t_2|t_1)$ similarly. Define

$$dA_{12}(t_1,t_2) = \{\lambda_{12}(t_1,t_2) - \lambda_{12(1|2)}(t_1|t_2)\lambda_1(t_2) - \lambda_{12(2|1)}(t_2|t_1)\lambda_2(t_1) + \lambda_1(t_1)\lambda_2(t_2)\}dt_1dt_2.$$

We will show that, for large n, W under H_a is approximately normal with mean $ma_1a_2\omega\sqrt{n}$ and variance

$$\sigma^2 = m_1\sigma_1^2 + m_2\sigma_2^2 + m_1(m_1-1)c_1 + m_2(m_2-1)c_2 - m_1m_2c_{12},$$

where

$$\omega = \int_0^\infty \frac{S_1(t)S_2(t)G(t)}{a_1S_1(t)+a_2S_2(t)}\{d\Lambda_1(t) - d\Lambda_2(t)\}$$

$$\sigma_k^2 = a_{3-k}^2\int_0^\infty = \frac{S_{3-k}(t)^2S_k(t)G(t)}{\{a_1S_1(t)+a_2S_2(t)\}^2}d\Lambda_k(t)$$

$$c_k = a_{3-k}^2\int_0^\infty\int_0^\infty \frac{S_{3-k}(t_1)S_{3-k}(t_2)S_k(t_1,t_2)G(t_1\vee t_2)}{\{a_1S_1(t_1)+a_2S_2(t_1)\}\{a_1S_1(t_2)+a_2S_2(t_2)\}}dA_k(t_1,t_2)$$

$$c_{12} = a_1a_2\int_0^\infty\int_0^\infty \frac{S_2(t_1)S_1(t_2)S_{12}(t_1,t_2)G(t_1\vee t_2)}{\{a_1S_1(t_1)+a_2S_2(t_1)\}\{a_1S_1(t_2)+a_2S_2(t_2)\}}dA_{12}(t_1,t_2).$$

Hence we obtain the required sample size for power $1-\beta$ by

$$n = \frac{\sigma^2(z_{1-\alpha/2}+z_{1-\beta})^2}{(ma_1a_2\omega)^2}. \tag{18}$$

If subunits within each cluster are independent, the sample size formula reduces to a standard formula for the log-rank test with independent observations, e.g., Schoenfeld (1983).

5.2.3.3 Simulation Plan

Extensive simulations will be conducted to evaluate the performance of the proposed sample size formulas. We will utilize Core C for assistance in designing and implementing these simulation studies. We will calculate a sample size n for a given input parameter setting; generate a large number (say 5000) of simulation data sets with size n each; apply the log-rank test with the specified α; and calculate the empirical power. If the empirical power is close to the nominal power $1-\beta$, then we can claim our sample size is accurate. We will conduct the simulations at a wide range of input parameter settings under both cluster randomization and subunit randomization.

5.3 Specific Aim 3 - Statistical Methodology for Cancer Drug Development

5.3.1 Targeted Therapy

Issues such as sample size, length of study, and cost will be investigated in various scenarios, which include assumptions about the size of the treatment effect, prevalence of marker positives, and patient allocation in subgroups.

In comparing a targeted design to a traditional design, the factors affecting the relative efficiency include the relative treatment effect ($\theta = \delta_0/\delta_1$) in those patients in $R-$ compared to those in $R+$, the prevalence $(1-\gamma)$ of $R+$ patients, the accuracy of the assay (through ω, the positive predictive value of the assay), and the costs of the assay and treatment. Specifically, the relative number of patients required on the traditional design (n) relative to the number required on the targeted design (n_T) is:

$$n/n_T \cong \left[\frac{(1-\omega)\theta+\omega}{\gamma\theta+(1-\gamma)}\right]^2$$

That is, the relative efficiency of the targeted design in terms of the required number of patients is high if θ is small (i.e., relatively little treatment benefit in the $R-$ patients), if $1-\gamma$ is small (i.e., few $R+$ patients in the population), and if the positive predictive value (ω) of the assay is high. If the targeted therapy does not benefit $R-$ patients at all (i.e., $\theta = 0$), the required number of patients for the traditional design is more than 20 times higher than a targeted design. These numbers help define situations in which targeted therapies may not work well in unscreened populations.

The picture is somewhat less favorable for a targeted design if we consider the time required for accrual of the requisite number of patients and the cost of screening. For example, the relative average time to accrue the

requisite number of patients is $(n/n_T)(1-\gamma)$. If the prevalence of $R+$ is low, fewer patients are required in a targeted design, but it takes longer to accrue these patients and more patients have to be screened to identify those suitable for the trial.

These considerations suggest that if we can be reasonably certain that the benefit of the targeted therapy for the $R-$ patients will be low and the assay is known to be highly accurate, a targeted design will be preferable to a traditional design. However, if we are not certain of these items, various "hybrid" or alternative designs should be considered (Wang et al., 2007; Song and Chi, 2007; Sargent et al., 2005). One such design that we will consider is an enrichment design. Like the BSR design, an enrichment design will randomize both marker positive and marker negative patients. But it only randomizes a subset of all marker negative patients in order to reduce cost and to improve study efficiency. In order to maximize the efficiency gain, the process of selecting which patient to randomize may depend on the biomarker prediction and other baseline patient characteristics. The efficiency gain due to an enrichment design could be significant when marker negatives are predominant in the unselected patient population and when there exists auxiliary variables to identify those informative patients.

For the enrichment design, standard statistical methods that fail to take into account the biased subsampling scheme would lead to an overestimate of the overall treatment effect, the interaction effect between treatment and the biomarker performance. We will develop methods for enrichment designs with the objectives of evaluating treatment effects (overall, subgroups), the interaction of treatment with biomarker, and the performance of the biomarker (e.g. Se, Sp, ROC curve, PPV and NPV). We are particularly interested in the semiparametric empirical likelihood approach (e.g. Wang and Zhou, 2006). The method has the potential to utilize all clinical and correlative sciences data collected for all registered patients (randomized or not, complete data or not). The method can be developed for continuous, binary, or time-to-event endpoints. The method involves estimating the empirical distribution of the unselected patient population, which is then incorporated as weights into the standard statistical methods to obtain unbiased estimates. The method allows one to carry out a *principled analysis* in which the treatment effect estimate is constructed independently of estimating the empirical distribution of the unselected population. Enrichment designs will be compared to target designs and BSR designs with respect to sample size and other characteristics in various scenarios. To fix the idea, we briefly consider the problem of ROC curve estimation under an enrichment design, in which all patients with positive markers and a random subset of the patients with negative markers will be selected to be randomized and to be followed for clinical outcome. Let Y be the marker value, D the clinical outcome, X the vector of covariates that affects the marker predictive accuracy, that include treatment arm. Assume that the ROC curve characterizing how well the marker Y predicts the clinical outcome D follows a binormal form: $Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 D X_D + \sigma(D)\epsilon$, where $\epsilon \sim N(0,1)$ and $\sigma(D) = \sigma_1 I[D=1] + \sigma_0 I[D=0]$. X_D could be X or a subset of X. Further, assume that Y falls into one of the K mutually exclusive intervals. When $K=2$, it could divide Y into a negative region $Y < a$ and a positive region $Y \geq a$. The combined data consists of three components: (1) a marker positive and outcome observed component of size n_1, (2) a marker negative and outcome observed component of size n_2, and (3) a marker negative and outcome unobserved component of size n_3. The combined likelihood can be shown as

$$L(\beta) = \left[\prod_{i=1}^{n_1+n_2} f_\beta(y_i|d_i,x_i)\right]\left[\prod_{i=1}^{n_1+n_2} g(d_i,x_i)\right]\left[\pi_1^{n_1}(1-\pi_1)^{n_2+n_3}\right]$$

where $\pi_1 \equiv P(y \geq a) = \int P(y \geq a|d,x)dG(d,x)$. The unknown cumulative distribution $G(d,x)$ is difficult to specify parametrically and its miss-specification could lead to biased parameter estimation. We will adopt an empirical likelihood based semiparametric method to estimate β and the induced covariate-specific and marginal ROC curves. Based on the empirical likelihood theory, to estimate $p_i \equiv g(d_i,x_i)$, it is sufficient to search the discrete probability space defined by the observed values of $\{d,x\}$. For fixed β, we search for $\{p_i\}$, $i = 1,\cdots,n_1+n_2$, that maximize $\log L(\beta, g(.))$ under the constraints using a Lagrange multiplier argument. After substitution of $\{p_i\}$ into the likelihood function we have a profiled empirical likelihood function $l_p(\beta)$. Once obtaining $\hat{\beta}$, one can estimate the induced covariate-specific ROC curve and its marginal counterpart according to the binormal ROC model.

5.3.2 Phase II Designs

We will expand on our previous work on the design of phase II clinical trials, with particular attention to the issue of making phase II trial results more predictive of the results in phase III results. Our first approach will be

to utilize existing data to build predictive models.

In order to better understand why certain phase II trials fail to predict phase III studies, a comprehensive study is needed. Chan et al. (2008), which focuses on pharmaceutical-sponsored trials, used a logistic regression model to identify predictive factors. Zia et al. (2005), which focuses on solid malignancies, used both logistic regression models and generalized estimating equations but did not identify factors that significantly predicted positive phase III. Moreover, none of the published methods used non-parametric or machine learning classifications. These state-of-art classifiers are able to train models that are sensitive to outliers in the data, resulting in better predictions. Moreover, they are able to cope with non-linearity in the data.

Our approach in building a predictive model is as follows. A pilot study will be conducted to evaluate the feasibility of identifying factors of phase II studies predictive of positive phase III. Data of various factors of phase II studies related to positive phase III trials will be collected. A positive phase III study is defined as a study if the stated primary end points were met. If primary endpoints were not stated, then it is considered to be positive if the experimental treatment is statistically significantly better in terms of overall survival than the standard therapy available. These data come from approximately 80 phase III trials and their corresponding phase II trials conducted by the CALGB that completed recruitment of patients between 1985 and 2000. Completed trials will include all trials that either met the target accrual or had been closed at an interim analysis. Trials that are terminated early will be excluded from the analysis. This will allow us to refine the variables to be collected. After the completion of the pilot study, we will request data and study protocols from phase III trials conducted between 1985 and 2000 from other National Clinical Trials Cooperative Groups. This will be our main training and test samples for building models predictive of positive phase III. Core B will act as a main hub for gathering and compiling data from different National Clinical Trials Cooperative Groups. Data will be stored in a database with technical support from Core B.

For preliminary analysis, machine learning and other nonparametric classification methods on all variables will be performed in addition to univariate and bivariate analyses. The variables considered include trial design elements, accrual, sample size, primary endpoints of interest, and whether it is randomized or not. The model built from the training set will then be tested for internal validity using cross-validation or bootstrapping methods. As classification models trained on unbalanced data sets will tend to favor the larger set, i.e. the negative phase III, we will develop new methods to account for this statistical issue. We will consider and compare resampling approaches, weighting schemes and area under the ROC curve methods to assess model performance. Feature selection techniques will be employed to rank the factors of phase II studies that are good at predicting positive phase III trials. The identification of these important factors will help design group randomized prevention trials, studies with multinomial endpoints in Aim 2 as well as for targeted trials in Aim 3. We will make use of Core C to prepare user-friendly software in SAS/R of the predictive tools for public dissemination.

As one example of the type of methodological issues we will investigate, consider a phase II trial for which the patient population is heterogeneous, say high- and low-risk. A standard design to account for the heterogeneity of the patient population is a single-arm trial based on a projected prevalence for each sub-population. When study is completed, however, the realized prevalence may be very different from the projected one. In this case, the fixed rejection value for a chosen standard phase II design may be either too strict (i.e., increasing the false rejection probability of the experimental therapy) if the trial accrues more high-risk patients than expected or too liberal (i.e., increasing the false acceptance probability of the experimental therapy) if the trial accrues more low-risk patients than expected.

In order to address this problem, we will develop an adaptive single-arm design. For cohort $j(= 1, 2)$, let p_j denote the response rate (RR) of the therapy. Suppose that we want to reject the new therapy if the RR for sub-population $j(= 1, 2)$ is p_{0j} or lower. We specify the hypotheses as

$$H_0 : p_1 = p_{01}, p_2 = p_{02} \text{ vs. } H_a : p_1 > p_{01}, p_2 > p_{02}.$$

Suppose we plan to accrue n patients in a single stage design. Let m_j denote the number of patients from sub-population j. Also, let X_j denote the number of responders among m_j patients from sub-population j. Then, given a type I error rate α and an observed m_j, we can choose a rejection value $a = a(m_1)$ from

$$P(X_1 + X_2 \geq a | p_{01}, p_{02}, m_1) \leq \alpha.$$

The above probability is calculated by assuming that X_j are independent Bin(m_j, p_{0j}) random variables. The

power conditioning on m_j can be similarly calculated for specified RRs $p_{aj}(> p_{0j})$ under H_a. By this adaptive design, the critical value will change depending on the observed number of patients m_j from subpopulation j.

This method can be easily extended to two-stage designs. The sample size for each stage is determined by a standard design, such as Simon's minimax or optimal design, based on the projected prevalence of each subpopulation, but the rejection value is adjusted depending on the observed prevalence from the trial. Conditioning on the observed prevalence, the developed design will control the type I error probability under the desired level, and the power around the prespecifed power. In contrast, the conditional type I error of a standard design based on a projected prevalence with a fixed rejection value will wildly fluctuate around the prespecified type I error depending on the observed prevalence. Furthermore, the marginal type I error and power of the conventional design can be heavily biased if the projected prevalence is different from the true prevalence.

The rejection value of a phase II trial is chosen for a fixed sample size. Because the realized sample size of a trial may be slightly different from the planned sample size, the chosen rejection value is not valid. We will also develop a p-value calculation conditioning on the final sample size and the observed prevalence, and conduct statistical testing by comparing the conditional p-value with the prespecified type I error probability using a similar method developed by Jung et al. (2006). Whatever the final sample size of the phase II trial, we can reject or accept the experimental therapy by comparing the calculated conditional p-value with a pre-specified α level.

5.3.3 Partially Randomized Designs

Some details of the empirical likelihood approach are given below. Without loss of generality, let Y be a continuous clinical endpoint, X is a covariate to be adjusted for efficiency gain, Z is a covariate to be adjusted for its potential effect on patient's treatment preference. In the partially randomized trial, patients can viewed as simple random samples drawn from four strata: (1) randomized, treatment; (2) randomized, placebo; (3) non-randomized, treatment; and (4) non-randomized, placebo. Let the index $h = 0, 1$ represent non-randomized and randomized patients, the index $k = 0, 1$ represent placebo and treatment. Let n_{hk} be the size of the hkth stratum. For the ith subject in the hkth stratum, $W_{hki} = \{Y_{hki}, X_{hki}, Z_{hki}\}$ is the observed data following a distribution function F_{hk}. The empirical likelihood approach has been discussed for a covariance adjustment problem (e.g. Owen, 2001) and a biased sampling problem (e.g. Wang and Zhou, 2006). In this setting, we will maximize the log empirical likelihood for W, $l = \sum_{h=0}^{1}\sum_{k=0}^{1} \log p_{hki}$, where $p_{hki} = pr(Z_{hk} = z_{hki})$ subject to the normalizing constraints $\sum_{i=1}^{n_{hk}} p_{hki} = 1, \quad h = 0, 1, \quad k = 0, 1$. The maximization is also subject to the randomization constraints for the randomized patients $\sum_{i=1}^{n_{11}} p_{11i}x_{11i} = \sum_{i=1}^{n_{10}} p_{10i}x_{10i}$, corresponding to the mean equivalence of X between treatment and placebo for randomized patients due to randomization, i.e., $\Delta_x = E(X_{11}) - E(X_{10}) = 0$. Furthermore, the maximization is subject to the profiling constraints for non-randomized patients $\sum_{i=1}^{n_{01}} p_{01i}z_{01i} = \sum_{i=1}^{n_{00}} p_{00i}z_{00i}$, corresponding to the constraint that we enforce the equivalence of covariate mean between treatment and placebo for non-randomized patients in order to correct any covariate unbalance between these two subgroups of patients. Notice that for non-randomized patients, a similar constraint $\sum_{i=1}^{n_{01}} p_{01i}x_{01i} = \sum_{i=1}^{n_{00}} p_{00i}x_{00i}$ can be enforced as well if it is necessary. The empirical likelihood distribution p_{hki} for h, k can be estimated using the following Lagrange multiplier argument

$$H = \sum_{h=0}^{1}\sum_{k=0}^{1} \log p_{hki} + \sum_{h}\sum_{k} \alpha_{hk}\left(\sum_{i=1}^{n_{hk}} p_{hki} - 1\right) + \lambda\left(\sum_{i=1}^{n_{11}} p_{11i}x_{11i} - \sum_{i=1}^{n_{10}} p_{10i}x_{10i}\right) + \psi\left(\sum_{i=1}^{n_{01}} p_{01i}z_{01i} - \sum_{i=1}^{n_{00}} p_{00i}z_{00i}\right)$$

where $\eta = (\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11}, \lambda, \psi)$ are Lagrange multipliers. Set $\frac{\partial H}{\partial p_{hki}} = 0$ and $\sum_{i=1}^{n_{hk}} p_{hki}\frac{\partial H}{\partial p_{hki}} = 0$. It can shown that p_{hki} can be estimated by solving a set of equations on η. The overall treatment effect is defined as $\Delta_y = \mu_{y.1} - \mu_{y.0} = (\pi_{11}\mu_{y11} + \pi_{01}\mu_{y01}) - (\pi_{10}\mu_{y10} + \pi_{00}\mu_{y00})$ where $\hat{\pi}_{hk} = n_{hk}/n_k$ is the proportion of the hth level of randomized patients in those who receive the kth level of treatment. The adjusted estimator for Δ_y is

$$\begin{aligned}\hat{\Delta}_y &= \hat{\mu}_{y.1} - \hat{\mu}_{y.0} = (\hat{\pi}_{11}\hat{\mu}_{y11} + \hat{\pi}_{01}\hat{\mu}_{y01}) - (\hat{\pi}_{10}\hat{\mu}_{y10} + \hat{\pi}_{00}\hat{\mu}_{y00}) \\ &= \left(\frac{n_{11}}{n_{.1}}\sum_{i=1}^{n_{11}} \hat{p}_{11i}y_{11i} + \frac{n_{01}}{n_{.1}}\sum_{i=1}^{n_{01}} \hat{p}_{01i}y_{01i}\right) - \left(\frac{n_{10}}{n_{.0}}\sum_{i=1}^{n_{10}} \hat{p}_{10i}y_{10i} + \frac{n_{00}}{n_{.0}}\sum_{i=1}^{n_{00}} \hat{p}_{00i}y_{00i}\right)\end{aligned}$$

The asymptotic variance estimator for Δ_y can be derived by applying the empirical likelihood theory. The variance estimate as well as the confidence interval can also be obtained by a resampling method. Notice that the

empirical distribution estimate $\{p_{hki}\}$ would be able to be used with nonparametric methods for estimation and hypothesis testing, such as Wilcoxon rank-sum test, in this setting. In the above, we assume the potential bias introduced by the unbalanced distribution Z between the two preference groups can be effectively eliminated by forcing equal means. Additional constraints on 2nd (or higher) moments) can be added if forcing equal means fails to remove the bias.

5.4 Timetable

The theoretical development and the computer programming will go hand in hand. The following is the time table for each aim.

5.4.1 Specific Aim 1 - Design and Sample Size Calculation for Longitudinal and Joint Models for Longitudinal and Survival data

In Year 1, we will develop methods for sample size and power estimation for the overall treatment effect on survival In Year 2, we will develop methods for the effect of the longitudinal process on survival in Year 3, we will develop methods for the direct treatment effect on survival. In Year 4, we will extend these methods to address complex design issues in settings involving multivariate longitudinal and multivariate survival processes. In Year 5, we will develop user-friendly software utilizing Core C

5.4.2 Specific Aim 2 - Design and Analysis of Prevention and Therapeutic Trials

In Year 1, we will develop statistical methods and sample size formula for clustered survival data, In Year 2, we will conduct simulation studies with coding assistance from Core C and data analysis utilizing Core B for clustered survival data; write and submit manuscripts. In Year 3, we will develop statistical methods and sample size formula for clustered recurrent events data. In Year 4, we will conduct simulation studies with coding assistance from Core C and data analysis utilizing Core B for clustered recurrent events data; write and submit manuscripts. In Year 5, we will develop user-friendly software utilizing Core C

5.4.3 Specific Aim 3 - Statistical Methodology for Cancer Drug Development

5.4.3.1 Targeted Therapy

In Year 1, we will develop semiparametric methods for the enrichment designs. In Year 2, we will evaluate the performance of the methods through simulation with coding assistance from Core C. In Year 3, we will write and submit research papers. In Year 4, we will extend the methods to more general cases. In Year 5, we will develop user-friendly software utilizing Core C.

5.4.3.2 Phase II Designs

In Year 1 we will, define the variables to be collected for the predictive model and check for their availability and develop single-stage phase II design for heterogeneous population. In Year 2, we will gather data for phase II pilot study from CALGB through Core B and extend phase II design for heterogeneous population to two-stage. In Year 3, we will build predictive models on data gathered through Core B and use the preliminary results to setup the framework for the design of randomized phase II and seamless phase II/III trials. We will also evaluate single-stage and two-stage designs through simulation with coding assistance from Core C. In Year 4, we will gather data through Core B for final training and testing samples from other Cancer Cooperative Groups. We will also prepare a manuscript for heterogeneous population phase II designs . In Year 5, we will build a predictive model on all data gathered through Core B and refine the designs of phase II trials.

5.4.3.3 Partially Randomized Designs

In Year 1, we will develop the proposed method to the partially randomized designs. In Year 2, we will evaluate the semiparametric methods proposed here and in Aim 1 of Project 2 through simulation with coding assistance from Core C. In Year 3, we will extend the proposed method to survival endpoints In Year 4, we will extend the proposed method to more general cases. In Year 5, we will develop user-friendly software utilizing Core C.

6 INCLUSION ENROLLMENT REPORT

N/A

7 BIBLIOGRAPHY AND REFERENCES CITED

Abbring, J. H. and Van den Berg, G. J. (2007). The unobserved heterogeneity distribution in duration analysis. *Biometrika* **94**, 87-99.

Abrams, D. B., Boutwell, W. B., Grizzle, J., Heimendinger, J., Sorensen, G., and Varnes, J. (1994). Cancer control at the workplace: the Working Well Trial. *Preventive Medicine* **23**, 15-27.

A'Hern, R. P. (2001). Sample size tables for exact single-stage phase II designs. *Statistics in Medicine* **20**, 859-866.

Bernardo, M. V. P. and Harrington, D. P. (2001). Sample size calculations for the two-sample problem using the multiplicative intensity model. *Statistics in Medicine* **20**, 557-579.

Braun, T. and Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association* **96**, 1424-1432.

Brewin, C. R. and Bradley, C. (1989). Patient preferences and randomised clinical trials. *British Medical Journal* **299**, 313-315.

Brown, E. R. and Ibrahim, J. G. (2003a). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221-228.

Brown E. R. and Ibrahim J. G. (2003b). Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59**, 686-693.

Brown E. R., Ibrahim J. G., and DeGruttola V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61**, 64-73.

Burton, A. (2007). Are we doing the right phase III trials? *The Lancet Oncology* **8**, 193-193.

Cai, J., Fan, J., Jiang, J., and Zhou, H. (2007). Partially linear hazard regression for multivariate failure time data. *Journal of the American Statistical Association* **102**, 538-551.

Cai, J., Fan, J., Jiang, J., and Zhou, H. (2008). Partially linear hazard regression with varying coefficients for multivariate survival data. *Journal of the Royal Statistical Society, Series B* **70**, 141-158.

Cai, J., Fan, J., Li, R., and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303-316.

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151-164.

Cai, J. and Prentice, R. L. (1997). Regression estimation using multivariate failure time data and a common baseline hazard function model. *Lifetime Data Analysis* **3**, 197-213.

Cai, J., Qaqish, B., and Zhou, H. (2001). Marginal analysis for cluster-based case-control studies. *Sankhya, Series B* **63**, 326-337.

Cai, J. and Schaubel, D. E. (2004). Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Analysis* **10**, 121-138.

Cai, J. and Zeng, D. (2004). Sample size/power calculation for case-cohort studies. *Biometrics* **60**, 1015-1024.

Cai, J. and Zeng, D. (2007). Power calculation for case-cohort studies with non-rare events. *Biometrics* **63**, 1288-1295.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, in press. NIHMSID: 90130.

Chan, J. K., Ueda, S. M., Sugiyama, V. E., Stave, C. D., Shin, J. Y., Monk, B. J., Sikic, B. I., Osann, K., and Kapp, D. S. (2008). Analysis of phase II studies on targeted agents and subsequent phase III trials: what are the predictors for success? *Journal of Clinical Oncology* **26**, 1511-1518.

Chang, M. N., Therneau, T. M., Wieand, H. S., and Cha, S. S. (1987). Designs for group sequential phase II clinical trials. *Biometrics* **43**, 865-874.

Chang, S. M., Reynolds, S. L., Butowski, N., Lamborn, K. R., Buckner, J. C., Kaplan, R. S., and Bigner, D. D. (2005). GNOSIS: guidelines for neuro-oncology: standards for investigational studies–reporting of phase 1 and phase 2 clinical trials. *Neuro-Oncology* **7**, 425-434.

Chen, Q., Zeng, D., and Ibrahim, J. (2007). Sieve maximum likelihood estimation for regression models with covariates missing at random. *Journal of the American Statistical Association* **102**, 1309-1317.

Chi, Y. and Ibrahim J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62**, 432-445.

Chi, Y. and Ibrahim J. G. (2007). Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions. *Statistica Sinica* **17**, 445-462.

Chon, H. S., Hu, W., and Kavanagh, J. J. (2006). Targeted therapies in gynecologic cancers. *Current Cancer Drug Targets* **6**, 333-363.

Cook, R. J. (1995). The design and analysis of randomized trials with recurrent events. *Statistics in Medicine* **14**, 2081-2098.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.

Cox, C., Chu, H., Schneider, M., and Muoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution (Tutorial in Biostatistics). *Statistics in Medicine* **26**, 4352-4374.

Davidian, M., Tsiatis, A.A., and Leon, S. (2005) Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with discussion and rejoinder). *Statistical Science* **20**, 261–301. PMCID:PMC2600547.

Dehling, H., Mikosch, T., and Srensen, M. (Eds.) (2002). *Empirical Process Techniques for Dependent Data.* A Birkhuser book.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data.* New York: Oxford University Press.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research.* London: Arnold.

Dy, G. K., Krook, J. E., Green, E. M., Sargent, D. J., Morton, R. F., Fuchs, C. S., Ramanathan, R. K., Williamson, S. K., Findlay, B. P., Pockaj, B. A., Sticca, R. P., Alberts, S. R., Pitot IV, H. C., and Goldberg, R. M. (2007). Impact of complete response to chemotherapy on overall survival in advanced colorectal cancer : results from intergroup N9741. *Journal of Clinical Oncology* **25**, 3469-3474. PMID: 17687151.

El-Maraghi, R. H. and Eisenhauer, E. A. (2008). Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. *Journal of Clinical Oncology* **26**, 1346-1354.

Eng, K. H, and Kosorok, M. R. (2005). A sample size formula for the supremum log-rank statistic. *Biometrics* **61**, 86–91.

Feng, Z., McLerran, D., and Grizzle, J. (1996). A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine* **15**, 1793-1806.

Fleming, T. R. (1982). One sample multiple testing procedure for phase II clinical trials. *Biometrics* **38**, 143-151.

Fricker, J. (2008). Time for reform in the drug-development process. *The Lancet Oncology* **9**, 1125-1126.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15**, 1069-1092.

Gangnon, R. E. and Kosorok, M. R. (2004). Sample-size formula for clustered survival data using weighted log-rank statistics. *Biometrika* **91**, 263–275.

Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* **13**, 346-353.

George, S. L. (2007). Response rate as an endpoint in clinical trials. *Journal of the National Cancer Institute* **99**, 98-99.

George, S. L. (2008). Statistical issues in translational cancer research. *Clinical Cancer Research* **14**, 5954-5958.

George, S. L. (2009). Design of phase III clinical trials. In *Fundamentals of Oncology Clinical Trials*, K. Kelly and S. Halabi (eds). New York: Demos Medical Publishing, in press.

Goldberg, R. M., Sargent, D. J., Morton, R. F., Fuchs, C. S., Ramanthan, R. K., Williamson, S. K., Findlay, B. P., Pitot, H. C., and Alberts, S. A. (2004). A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *Journal of Clinical Oncology* **22**, 23-30. PMID:-14665611.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician* **58**, 16-24.

Hayes, R. J. and Bennett, S. (1999). Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* **28**, 319-326.

Henderson, R., Diggle, P. J., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* **1**, 465-480.

Hogan, J. W. and Laird, N. W. (1997). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine* **16**, 239-257.

Horowitz, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* **67**, 1001-1028.

Hoskins, J. M. , Goldberg, R. M., and McLeod, H. L. (2007). UGT1A1*28 and irinotecan-induced neutropenia: dose matters. *Journal of the National Cancer Institute* **99**, 1290-1295. PMID: 17728214.

Hughes, M. D. (1997). Power considerations for clinical trials using multivariate time-to-event data. *Statistics in Medicine* **16**, 865-882.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica* **14**, 863-883.

Jeong, J. H. and Jung, S. H. (2006). Rank tests for clustered survival data when dependent subunits are randomized. *Statistics in Medicine* **25**, 361-373.

Jiang, J. and Zhou, H. (2007). Additive hazards regression with auxiliary covariates. *Biometrika* **94**, 359-369.

Jung, S. H. (1999). Rank tests for matched survival data. *Lifetime Data Analysis* **5**, 67-79.

Jung, S. H. (2007a) Sample size calculation for paired survival data: a simulation method. *Journal of Statistical Computing and Simulation* **78**, 85-92.

Jung, S. H. (2007b). Sample size calculation for weighted rank tests comparing survival distributions under cluster randomization: a simulation method. *Journal of Biopharmaceutical Statistics* **17**, 839-849.

Jung, S. H. (2008a). Randomized phase II trials with a prospective control. *Statistics in Medicine* **27**, 568-583.

Jung, S. H. (2008b). Sample size calculation for the weighted rank statistics with paired survival data. *Statistics in Medicine* **27**, 3350-3365.

Jung, S. H., Carey, M., and Kim, K. M. (2001). Graphical search for two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **22**, 367-372.

Jung, S. H. and George, S. L. (2009). Between-arm comparisons in randomized phase II trials. *Journal of Biopharmaceutical Statistics*, in press. NIHMSID: 89741.

Jung, S. H. and Jeong, J. H. (2003). Rank tests for clustered survival data. *Lifetime Data Analysis* **9**, 21-33.

Jung, S. H., Kang, S. H., and Ahn, C. (2001). Sample size calculations for clustered binary data. *Statistics in Medicine* **20**, 1971-1982.

Jung, S. H. and Kim, K. M. (2004). On the estimation of the binomial probability in multistage clinical trials. *Statistics in Medicine* **23**, 881-896.

Jung, S. H., Lee, T., Kim, K. M., and George, S. L. (2004). Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine* **23**, 561-569.

Jung, S. H., Owzar, K., George, S. L., and Lee, T. (2006). P-value calculation for multistage phase II cancer clinical trials. *Journal of Biopharmaceutical Statistics* **16**, 765-775.

Kang, S. and Cai, J. (2008). Marginal hazards regression for retrospective studies within cohort with possibly correlated failure time data. *Biometrics*, in press, doi:10.1111/j.1541-0420.2008.01077.x

Kerry, S. M. and Bland, J. M. (2001). Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in Medicine* **20**, 377-390.

King, M., Nazareth, I., Lampe, F., Bower, P., Chandler, M., Morou, M., Sibbald, B., and Lai, R. (2005). Impact of participant and physician intervention preferences on randomized trials: a systematic review. *Journal of American Medical Association* **293**, 1089-1099.

Kong, L. and Cai, J. (2008). Case-cohort analysis with accelerated failure time model. *Biometrics*, in press, doi:10.1111/j.1541-0420.2008.01055.x

Kong, L., Cai, J., and Sen, P. K. (2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika* **91**, 305-319.

Kong, L., Cai, J., and Sen, P. K. (2006). Asymptotic results for fitting semiparametric transformation models to failure time data from case-cohort studies. *Statistica Sinica* **16**, 135-151.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Kosorok, M. R., Shi, Y., and DeMets, D. L. (2004). Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics* **60**, 134–145.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

Lake, S., Kaumann, E., Klar, N., and Betensky, R. (2002). Sample size re-estimation in cluster randomization trials. *Statistics in Medicine* **21**, 1337-1350.

Lambert, M. F. and Wood, J. (2000). Incorporating patient preferences in randomized trials. *Journal of Clinical Epidemiology* **53**, 163-166.

Lawless, J. F. and Nadeau, J. C. (1995). Nonparametric estimation of cumulative mean functions for recurrent events. *Technometrics* **37**, 158-168.

Lee, J. J. and Feng, L. (2005). Randomized phase II designs in cancer clinical trials: current status and future directions. *Journal of Clinical Oncology* **23**, 4450-4457.

Lee, E. W., Wei, L. J., and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J. P. Klein and P. K. Goel (eds), 237-247. Boston: Kluwer Academic.

Lin, H., Turnbull, B. W., McCulloch, E. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97**, 53-65.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in smoothing spline analysis of variance models. *The Annals of Statistics* **34**, 2272-2297.

Liu, A., Shih, W. J., and Gehan, E. (2002). Sample size and power determination in clustered repeated measurements. *Statistics in Medicine* **21**, 1787-1801.

Liu, Y., Zhou, H., and Cai, J. (2009) Estimated pseudo-partial-likelihood method for correlated failure time data with auxiliary covariates. *Biometrics*, in press. NIHMSID: 90113.

Long, Q., Little, R. J., and Lin, X. (2008). Causal inference in hybrid intervention trials involving treatment choice. *Journal of the American Statistical Association* **103**, 474-484.

Lunceford, J. K., and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23**, 2937-2960.

Maitournam, A. and Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in Medicine* **24**, 329-339.

Malakoff, D. (2008). Clinical trials and tribulations. Spiraling costs threaten gridlock. *Science* **322**, 210-213.

Mariani, L. and Marubini, E. (1996). Design and analysis of phase II cancer trials: a review of statistical methods and guidelines for medical researchers. *International Statistical Review* **64**, 61-88.

Mariani,L. and Marubini, E. (2000). Content and quality of currently published phase II cancer trials. *Journal of Clinical Oncology* **18**, 429-436.

Matsui, S. (2005). Sample size calculations for comparative clinical trials with over-dispersed Poisson process data. *Statistics in Medicine* **24**, 1339-1356.

Michaelis, L. C., Ratain, M. J., Michaelis, L. C., and Ratain, M. J. (2007). Phase II trials published in 2002: a cross-specialty comparison showing significant design differences between oncology trials and other medical specialties. *Clinical Cancer Research* **13**, 2400-2405.

Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press.

Murray, D. M., Hannan, P. J., and Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials: is it ever possible to avoid Cornfield's penalties? *Evaluation Review* **20**, 313-337.

Murray, D. M., Varnell, S. P., and Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological development. *American Journal of Public Health* **94**, 423-432.

O'Neil, B. H., Allen, R., Spigel, D., Stinchcombe, T., Moore, D., Berlin, J., and Goldberg, R. M. (2007). High incidence of cetuximab-related infusion reactions in Tennessee and North Carolina; association with atopic history. *Journal of Clinical Oncology* **25**, 3644-3648. PMID: 17704414.

Ottaiano, A., Facchini, G., Nasti, G., Budillon, A., and Caraglia, M. (2007). Phase II studies of anticancer chemotherapy: indirect evidence of poor quality. *Annals of Oncology* **18**, 403.

Owen, A. B. (2001). *Empirical Likelihood*. London: Chapman and Hall/CRC.

Owzar, K. and Jung, S. H. (2008). Designing phase II studies in cancer with time-to-event endpoints. *Clinical Trials* **5**, 209-221.

Pang, H., Kim, I., and Zhao, H. (2008). Pathway-based methods for analyzing microarray data. In *Analysis of Microarray Data: Network Based Approaches*, F. Emmert-Streib and M. Dehmer (eds), 355-384. Weinheim: Wiley-VCH.

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., Floyd, E., and Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics* **22**, 2028-2036.

Pang, H., Tong, T., and Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics*, accepted. NIHMSID: 90151.

Pang, H. and Zhao, H. (2008). Building pathway clusters from random forests classification using class votes. *BMC Bioinformatics* **9**, 87-87.

Pepe, M. and Cai, J. (1993). Some graphical displays and marginal regression for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**, 811-820.

Piedbois, P. (2007). Randomised phase II trials in oncology. *Bulletin du Cancer* (Paris) **94**, 953-956.

Potthoff, R. F. and George, S. L. (2009). Flexible phase I clinical trials: allowing for non-binary toxicity response and removal of other common limitations. *Statistics in Biopharmaceutical Research*, in press. NIHMSID: 90142.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300-325.

Rawlins, M. D. (2004). Cutting the cost of drug development? *Nature Reviews Drug Discovery* **3**, 360-364.

Raudenbush, S. W. (1997). Statistical analysis and optimal design in cluster randomized trials. *Psychological Methods* **2**, 173-185.

Redman, M. and Crowley, J. (2007). Small randomized trials (editorial). *Journal of Thoracic Oncology* **2**, 1-2.

Rosner, G. L., Stadler, W., and Ratain, M. J. (2002). Randomized discontinuation design: application to cytostatic antineoplastic agents. *Journal of Clinical Oncology* **20**, 4478-4484.

Rubinstein, L. V., Korn, E. L., Freidlin, B., Hunsberger, S., Ivy, S. P., and Smith, M. A. (2005). Design issues of randomized phase II trials and a proposal for phase II screening trials. *Journal of Clinical Oncology* **23**, 7199-7206.

Rücker, G. (1989). A two-stage trial design for testing treatment, self-selection, and treatment preference effects. *Statistics in Medicine* **8**, 477-485.

Sargent, D. J., Conley, B. A., Allegra, C., and Collette, L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* **23**, 2020-2027.

Schaubel, D. E. and Cai, J. (2005a). Semiparametric methods for clustered recurrent event data. *Lifetime Data Analysis* **11**, 405-425.

Schaubel, D. E. and Cai, J. (2005b). Analysis of clustered recurrent event data with application to hospitalization rates among renal failure patients. *Biostatistics* **6**, 404-419.

Schaubel, D. E., Zeng, D., and Cai, J. (2006). A semiparametric additive rates model for recurrent event data. *Lifetime Data Analysis* **12**, 389-406.

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* **39**, 499-503.

Shen, Y. and Cai, J. (2003). Sample size re-estimation for clinical trials with censored survival data. *Journal of the American Statistical Association* **98**, 418-426.

Simon, R. (1987). How large should a phase II trial of a new drug be? *Cancer Treatment Reports* **71**, 1079-1085.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1-10.

Simon, R., Wittes, R. E., and Ellenberg, S. S. (1985). Randomized phase II clinical trials. *Cancer Treatment Reports* **69**, 1375-1381.

Simon, R. and Maitournam, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* **10**, 6759-6763.

Slymen, D. J. and Hovell, M. F. (1997). Cluster versus individual randomization in adolescent tobacco and alcohol studies: illustrations for design decisions. *International Journal of Epidemiology* **26**, 765-771.

Sommer, A., Djunaedi, E., Loeden, A. A., Tarwotjo, I., Jr., West, K. P., and Tilden, R. (1986). Impact of vitamin A supplementation on childhood mortality, a randomized clinical trials. *Lancet* **1**, 1169-1173.

Song, Y. and Chi, G. Y. H. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* **26**, 3535-3549.

Song, R., Kosorok, M. R., and Cai, J. (2008). Robust covariate-adjusted log-rank statistics and corresponding sample size formula for recurrent events data. *Biometrics* **64**, 741-750.

Song, R., Zhou, H., and Kosorok, M. R. (2009). On semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, in press. NIHMSID: 75064.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials* **7**, 8-17.

Stadler, W. M. (2007). The randomized discontinuation trial: a phase II design to assess growth-inhibitory agents. *Molecular Cancer Therapeutics* **6**, 1180-1185.

Steinberg, S. M. and Venzon, D. J. (2002). Early selection in a randomized phase II clinical trial. *Statistics in Medicine* **21**, 1711-1726.

Storer, B. E. (1992). A class of phase II designs with three possible outcomes. *Biometrics* **48**, 55-60.

Tangen, C. M. and Crowley, J. J. (2006). Phase II trials using time-to-event endpoints. In *Handbook of Statistics in Clinical Oncology*, J. Crowley and D. P. Ankerst (eds), 143-154. New York: Chapman and Hall.

The COMMIT Research Group. (1995a). Community intervention trial for smoking cessation (COMMIT): I. Cohort results from a four-year community intervention. *American Journal of Public Health* **85**, 183-192.

The COMMIT Research Group. (1995b). Community intervention trial for smoking cessation (COMMIT): II. Changes in adult cigarette smoking prevalence. *American Journal of Public Health* **85**, 193-200.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809-834.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27-37.

Tsiatis, A.A. and Davidian, M. (2007) Discussion of "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," by J.D.Y. Kang and J.L. Schafer. *Statistical Science* **22**, 569–573. PMCID: PMC2397555.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* **27**, 4658-4677.

Turrisi III, A. T. (2005). Creeping phase II-ism and the medical pharmaceutical complex: weapons of mass distraction in the war against lung cancer. *Journal of Clinical Oncology* **23**, 4827-4829.

Wang, S. J., O'Neill, R. T., and Hung, H. M. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* **6**, 227-244.

Wang, X.F., Pang H., and Schwartz T.A. (2009). Building and Validating High Throughput Lung Cancer Biomarkers. *Chance*: under revision.

Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96**, 895-905.

Wang, X., Wu, Y., and Zhou, H. (2009). Outcome and auxiliary-dependent subsampling and statistical inference for cancer biomarker studies. *Journal of Biopharmaceutical Statistics*: under review.

Wang, X. and Zhou, H. (2006). A semiparametric empirical likelihood method for biased sampling schemes in epidemiologic studies with auxiliary covariates. *Biometrics* **62**, 1149-1160.

Wang, X. and Zhou, H. (2009). Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. *Biometrics*, in press. NIHMSID: 90132.

Weaver, M. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent subsampling. *Journal of the American Statistics Association* **100**, 459-469.

Wieand, H. S. (2005). Randomized phase II trials: what does randomization gain? *Journal of Clinical Oncology* **23**, 1794-1795.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330-339.

Yin, G. and Cai, J. (2004). Additive hazards model with multivariate failure time data. *Biometrika* **91**, 801-818.

Zelen, M. (1990). Randomized consent designs for clinical trials: an update. *Statistics in Medicine* **9**, 645-656.

Zeng, D. and Cai, J. (2005a). Asymptotic results for maximum likelihood estimates in joint analysis of repeated measurements and survival time. *The Annals of Statistics* **33**, 2132-2163.

Zeng, D. and Cai, J. (2005b). Simultaneous modeling of survival and longitudinal data with an application to repeated quality of life measures. *Lifetime Data Analysis* **11**, 151-174.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). *Journal of the Royal Statistical Society, Series B* **69**, 507-564.

Zhang, H. H. (2006). Variable selection for support vector machines via smoothing spline ANOVA. *Statistica Sinica* **16**, 659-674.

Zhang, H. H., Ahn, J., Lin, X., and Park, C. (2006). Gene selection using support vector machines with nonconvex penalty. *Bioinformatics* **22**, 88-95.

Zhang, H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94**, 691-703.

Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R., and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *Journal of American Statistical Association* **99**, 659-672.

Zhou, H., Brighton, L., Hazucha, M., and Carson, J. L. (2009). Persistent upregulation of nasal epithelial ciliary beat frequency in smokers and non-smokers exposed to environmental tobacco smoke. *Inhalation Toxicology*, in press. NIHMSID: 90135.

Zhou, H., Chen, J., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T., and Longnecker, M. P. (2007). Outcome dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemology* **18**, 461-468.

Zhou, H. and Wang C. Y.(2000). Failure time regression analysis with measurement error in covariates. *Journal of the Royal Statistical Society, Series B* **62**, 657-665.

Zhou, H., Weaver, M. A., Qin, J. and Wang, M. C. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling design with a continuous outcome. *Biometrics* **58**, 413-421.

Zhou, H. and Weinberg, C. R. (1999). A semiparametric mixture model for time to conception data. *Journal of Statistical Planning and Inference* **75**, 453-462.

Zhou, H. and You, J. H. (2007). Statistical inference for a semiparametric measurement error regression models with heteroscedastic errors. *Statistica Sinica* **137**, 2263-2276.

Zia, M. I., Siu, L. L., Pond, G. R., and Chen, E. X. (2005). Comparison of outcomes of phase II studies and subsequent randomized control studies using identical chemotherapeutic regimens. *Journal of Clinical Oncology* **23**, 6982-6991.

Zou, H. and Zhang, H. (2008). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, to appear.

8 PROTECTION OF HUMAN SUBJECTS

Although the proposed research indirectly involves human subjects through the preparation, in Core B, of de-identified data sets from identifiable patient data sources, the investigators on Project 1 will have access only to the de-identified data. Thus, the investigators on Project 1 will have no access to any identifiable patient information.

9 INCLUSION OF WOMEN AND MINORITIES

The methods we develop will be applicable to studies with both women and minorities and also to studies which examine treatment differences adjusted for gender, ethnicity and race. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, many of the existing data sets to be studied and provided by Core B include women and minorities, although we will not be generating any new data involving human subjects.

10 TARGETED/PLANNED ENROLLMENT TABLE

N/A

11 INCLUSION OF CHILDREN

The methods we develop will be applicable to studies with children and also to studies which examine treatment differences adjusted for age. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, some of the existing data sets to be studied and provided by Core B may include children, although we will not be generating any new data involving human subjects.

12 VERTEBRATE ANIMALS

N/A

13 SELECT AGENT RESEARCH

N/A

14 MULTIPLE PD/PI LEADERSHIP PLAN

N/A

15 CONSORTIUM/CONTRACTUAL ARRANGEMENTS

If the present application is funded, the University of North Carolina at Chapel Hill will execute subcontracts with the consortium institutions (Duke University and North Carolina State University). These inter-institutional agreements will be written consistent with the NIH consortium agreement policy.

16 LETTERS OF SUPPORT - None

17 RESOURCE SHARING PLAN(S)

(a) Data sharing plan: The data-related resources generated by the proposed research consists of new statistical methodology, software packages for implementation of the methodology, and tutorials for the software. The statistical methodology will be shared through peer reviewed publications and national meetings and

through other standard means. All accepted publications will be deposited in PubMed Central in accordance with the NIH Public Access Policy. Summaries of the methodology, the software and tutorials will be shared through a public web site managed by Core A, while Core C will assist in preparation of the software and tutorials for dissemination. This project will use de-identified data prepared by Core B to test the methods and to create demonstrations of use of the methods to be included in tutorials. This project will not be involved in sharing of these data; this function will be addressed by Core B.

(b) Sharing model organisms: N/A

(c) GWAS: N/A

PROJECT 2

METHODS FOR MISSING AND AUXILIARY DATA IN CLINICAL TRIALS

Project Leader: Marie Davidian, PhD

PROJECT SUMMARY (See instructions):

Randomized clinical trials are and will continue to be the key vehicle for evaluation of new and existing cancer therapies. This revolutionary era of advances in the biological sciences is leading to the discovery of novel biomarkers and complex genetic and genomic information that may be highly associated with various clinical outcomes, offering the tantalizing opportunity to exploit this information to both improve the precision of the analyses of trials and to develop models of longitudinal disease progression that may reveal important insights. A recurrent challenge is that missing data and subject drop-out are commonplace, presenting complications for analyses of these trials. Through a series of aims addressing these issues, this project proposes research that will have a significant impact on the quality and strength of inferences possible from current cancer clinical trials. That it is possible to improve efficiency of primary analyses of clinical trials by exploiting prognostic baseline auxiliary information is well known; however, such analyses are controversial because of the temptation to choose the analysis that leads to the most dramatic treatment effect. In the first aim, new methods for such "covariate adjustment" will be studied that circumvent this issue and can improve over existing approaches. In the second aim, these methods will be extended so that they may be used in the common case where outcomes are missing due to drop-out. Efficient methods for longitudinal analysis of measures such as quality of life and biomarkers in the presence of drop-out will also be developed. Understanding the relationship between such longitudinal measures and clinical outcomes such as time to recurrence or survival time is of key importance. The third aim focuses on development of methods for assessing the correctness of so-called joint statistical models used for this purpose and for assessing the influence of particular observations on the fit of the model, where the data used to develop the model may be missing. Finally, taking appropriate account of missing data sometimes requires unverifiable assumptions about why the data are missing, which are incorporated in models that thus cannot be checked based on the data. The fourth aim is devoted to development of a new statistical framework for assessing how sensitive conclusions are to the modeling assumptions made.

RELEVANCE (See instructions):

Randomized clinical trials in cancer research are the most important mechanism for the evaluation of new and existing therapies. Statistical methods will be developed that will improve the precision of the analyses of these trials and provide tools for drawing valid conclusions when some of the data intended to be collected are missing, e.g., if some subjects drop out of the trial, offering cancer researchers an expanded set of tools that will greatly improve the quality and strength of analyses of current cancer clinical trials.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location			
Organizational Name: The University of North Carolina at Chapel Hill			
DUNS: 608195277			
Street 1: Office of Sponsored Research, CB #1350		Street 2: 104 Airport Dr., Suite 2200	
City: Chapel Hill	County: Orange		State: NC
Province:	Country: USA		Zip/Postal Code: 27599-1350
Project/Performance Site Congressional Districts: NC-004			

Additional Project/Performance Site Location			
Organizational Name: North Carolina State University			
DUNS: 042092122			
Street 1: Research Admin/ SPARCS		Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514	
City: Raleigh	County: Wake		State: NC
Province:	Country: USA		Zip/Postal Code: 27695-7514
Project/Performance Site Congressional Districts: NC-02			

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location		
Organizational Name: Duke University		
DUNS: 044387793		
Street 1: Hock Plaza	Street 2: Box 2716 Med Ct.	
City: Durham	County: Durham	State: NC
Province:	Country: USA	Zip/Postal Code: 27705
Project/Performance Site Congressional Districts: NC-004		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Davidian, Marie	davidian2	NC State University	Project 2 Leader
Bondell, Howard D.		NC State University	Co-Investigator
Boos, Dennis D.	dennis_boos	NC State University	Co-Investigator
Cai, Jianwen	Jianwen_Cai	UNC-CH	Co-Investigator
Fine, Jason P.	Jasonp3p	UNC-CH	Co-Investigator
Ibrahim, Joseph G.	JOE_IBRAHIM	UNC-CH	Project Co-Leader
Jung, Sin-Ho	Jung0005	Duke University	Co-Investigator
Spector, Neil L.		Duke University	Co-Investigator
Stefanski, Leonard A.		NC State University	Co-Investigator
Tsiatis, Anastasios	butch_tsiatis	NC State University	Project Co-Leader
Zhang, H. Helen		NC State University	Co-Investigator

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project

Human Embryonic Stem Cells ☒ No ☐ Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

RESEARCH PLAN

1 INTRODUCTION TO RESUBMISSION/REVISION APPLICATION - N/A

2 SPECIFIC AIMS

Randomized clinical trials are the primary mechanism by which new cancer therapies are tested for efficacy and evaluated for regulatory approval. The advent of novel biomarkers and emerging genomic technologies that may yield important new baseline predictors of primary clinical outcomes, the increasing emphasis on analyses of longitudinal progression of markers such as measures of quality of life, and the routine complications of missing information and subject drop-out, present both challenges and opportunities for the interpretation of these studies. We propose four specific aims focused on new methodological advances to exploit prognostic auxiliary information and to provide frameworks for analyses in the presence of missing data that will affect notably the strength and impact of inferences possible from current cancer clinical trials:

Aim 1: To develop methods to improve efficiency of inferences in randomized cancer clinical trials using auxiliary covariates. Auxiliary baseline information is routinely collected on trial participants in addition to clinical endpoints, and it is well recognized that relationships between these data and outcomes may be exploited to enhance precision of primary and secondary analyses, increasing power to detect important effects. However, "adjusting" for auxiliary covariates has engendered considerable controversy because of the temptation under the usual regression approach to inspect different model fits and choose that leading to the most dramatic estimated treatment effect. We have recently proposed promising new methods that separate regression modeling from effect estimation, obviating this concern and raising the possibility for greater acceptance of more efficient adjusted analyses, but several issues must be resolved prior to their widespread adoption. We will develop new approaches to model selection and inference for these methods ensuring their reliable use, including for the case where key auxiliary information is missing for some subjects. We will also extend the methods to analysis of more complex trial designs, such as the partially randomized surgical trials studied in Aim 3 of Project 1.

Aim 2: To develop methods for primary and longitudinal analyses in the presence of drop-out. A routine feature of cancer trials is drop-out, where subjects are lost prior to the end of follow-up, so that data intended to be collected are missing subsequent to the time of drop-out. In many settings, sufficient information may have been collected on participants prior to drop-out to justify the assumption that the missingness is "at random," i.e., the drop-out mechanism may be "explained" by the data observed up to the time of drop-out. We will extend the adjustment methods for primary analyses studied in Aim 1 to this setting. Many cancer trials also involve analyses of longitudinal measures such as quality of life, which are complicated by drop-out. When the missing at random assumption is plausible, popular analysis methods require a correct statistical model for either the intended, full data or for the drop-out mechanism, and may yield biased inferences if the assumed such model is incorrect. Recent interest has thus focused on "doubly robust" methods, which use models for both but require only one or the other to be correct, thus providing protection against incorrect modeling assumptions. We will develop new, efficient, doubly robust methods for longitudinal analysis under these conditions.

Aim 3: To develop diagnostic measures for joint models for longitudinal and survival data in the presence of nonignorably missing data. Cancer trials may involve studies of the association between longitudinal markers and clinical outcomes such as relapse-free survival or death, and a popular framework for analysis is that of joint models for the longitudinal data and time-to-event outcome. Because of their complexity, these frameworks must rely on correct models for the full data to ensure valid inferences; thus, it is critical that reliable diagnostic methods are available for assessing model misspecification and goodness-of-fit and for identifying data that may have disproportionate influence on the results. Because drop-out and other forms of missing data are possible, such diagnostic measures must be applicable in this setting, including when the missingness is "nonignorable," where the missing at random assumption is not realistic. We will develop new diagnostic techniques for a variety of these models when there may be nonignorable missing outcome and/or covariate data that will provide the analyst with powerful and heretofore unavailable tools for joint model assessment.

Aim 4: To develop inference methods for sensitivity analyses of missing data. A major challenge when intended data are missing is that it is impossible to evaluate whether or not the missing at random assumption is justified based on the observed data, nor can models for nonignorable missingness mechanisms be verified. Thus, analyses may be predicated on incorrect such models, leading to misleading inferences. A popular

strategy in practice is to undertake a sensitivity analysis in which one inspects how inferences vary across multiple competing such postulated models. However, interpretation is problematic, as it may not be clear how to synthesize formally the results across models. We will develop rigorous inferential methods that explicitly acknowledge the non-identifiability of the missingness model as well as the need to explore simultaneously a range of plausible models in order to formalize evaluation of sensitivity of inferences.

3 BACKGROUND AND SIGNIFICANCE

A unifying theme of this project is that the vast resources devoted to and the commitment of the volunteer participants involved in current cancer clinical trials demand continued development of state-of-the-art statistical methods that make the most efficient and reliable use of the observed data collected in these studies. We now elaborate on the considerations and issues underlying each of our four aims.

3.1 Aim 1: Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates

The primary objective of a randomized clinical trial is to make comparisons among two or more treatments, where the endpoint forming the basis for comparison may be continuous, binary, categorical, or, as in the case of many cancer trials, a censored time-to-event. The standard primary analysis focuses on comparisons averaged across the patient population and is based on data collected on outcome and randomized treatment assignment only. However, if baseline auxiliary covariates are associated with the outcome, it is well known that the precision of this analysis may be improved by "adjusting" for these relationships (Pocock et al., 2002), and there is an extensive literature on the use of regression methods and other techniques for such covariate adjustment (e.g., Senn, 1989; Tangen and Koch, 1999; Lesaffre and Senn, 2003; Grouin, Day, and Lewis, 2004). For example, in Eastern Cooperative Oncology Group (ECOG) trial E1694 (Kirkwood, Ibrahim, et al., 2001), a two-arm, phase III clinical trial comparing a vaccine to high-dose interferon (HDI) in high-risk melanoma patients on the basis of relapse-free and overall survival, baseline covariates include standard possible correlates of outcome such as ulceration of the primary tumor and Breslow thickness. The potential for efficiency gains is of particular interest in this era where advances in molecular technology are leading to new, highly prognostic measures such as the lung metagene score in non-small cell lung cancer (NSCLC) used in Cancer and Leukemia Group B (CALGB) 30506. The existence of key predictive baseline covariates in these cancers as well as leukemia and breast, colorectal, and prostate cancer suggests that adjusted primary analyses may enjoy improved precision sufficient to detect important effects for which evidence from the standard unadjusted analysis may be ambiguous.

Despite this potential, there has been considerable debate regarding standard adjusted analyses due to concerns that the treatment effect and covariates are inextricably linked in a common regression model, tempting the analyst to inspect effect estimates across models and focus on the models and covariates that "best accentuate the estimate and/or statistical significance" of the estimate (Pocock et al., 2002). Thus, trialists and regulatory authorities have been reluctant to endorse these analyses and require that, when undertaken, they involve only a few covariates in prespecified regression models (e.g., Raab, Day, and Sales, 2000; Grouin et al., 2004). A consequence is that critical opportunities to enhance power to reveal real, important treatment effects may be lost due to unfortunate a priori such choices that cannot be later revised even if warranted.

Recently, by applying semiparametric theory (Tsiatis, 2006) to this general problem, we have developed new methods for covariate adjustment under unrestrictive assumptions that instead *separate* modeling of outcome-covariate relationships from estimation of the treatment effect (Tsiatis et al., 2008; Zhang, Tsiatis, and Davidian, 2008; Lu and Tsiatis, 2008). These methods involve regression modeling of the outcome-covariate relationship separately within each treatment group and have been shown to lead to impressive efficiency gains over existing methods. They thus support a "principled analysis" in which modeling exercises for each treatment may be conducted by distinct teams of analysts, who may be given access only to the treatment-specific data if complete transparency is desired, and the effect estimate is then constructed independently, circumventing the possibility of its inspection during the modeling process. This eliminates the need to impose a priori restrictions on model forms and covariates used, allowing, in principle, analysts license to deploy model selection strategies yielding the most predictive models for outcome within each treatment, which, according to the theory underlying the methods (Zhang et al., 2008), will in turn extract the maximum possible efficiency gains.

Although our methods offer great promise for widespread acceptance of covariate adjusted analyses and hence opportunities for more efficient and powerful inferences from current cancer clinical trials, considerable

work is needed to provide guidance on their effective use before they can be recommended for routine adoption. In this aim, we will investigate several fundamental issues in their development and implementation.

The key, appealing feature of the semiparametric approach is that separation of modeling and effect estimation encourages, and indeed demands, development of the best predictive treatment-specific models possible given the available data. There is a vast literature on model selection methods, including traditional techniques such as forward, backward, and stepwise selection based on fixed tuning constants such as "α to enter," and recent advances where tuning constants are selected adaptively; an admittedly incomplete list includes penalized methods such as the Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996), the adaptive LASSO (Zou, 2006; Wang and Leng, 2007; Zhang and Lu, 2007), the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001), and the False Selection Rate (FSR) methods developed by members of our team (Wu, Boos, and Stefanski, 2007; Boos, Stefanski and Wu, 2008). The performance of these methods in terms of prediction error and identifying "important" covariates has been widely studied; however, little is known of their relative merits for developing treatment-specific models in the context of our semiparametric covariate adjustment methods. In our first sub-aim, we will carry out a comprehensive study of this issue.

A related challenge is taking faithful account of the uncertainty associated with the use of model selection methods on inferences on treatment effect. Theoretically, in large samples, that model selection methods have been employed should have no impact on the reliability of inferences, which are predicated on the assumption that the models are known a priori. However, in finite sample sizes, particularly for small and moderately sized trials, the impact may be nonnegligible. In our second sub-aim, we will determine the sample sizes where a "correction" for this phenomenon is required and develop corresponding techniques based on bootstrap methodology and second order theory that ensure the validity of inferences.

Standard model selection methods assume that the full slate of baseline auxiliary covariates is available for all subjects. A common complication in many trials is that some baseline variables are missing for some subjects; e.g., in E1694, information on tumor ulceration and Breslow thickness is missing for roughly 20% of the subjects (see Section **3.3**). In our third sub-aim, we will study approaches to adapting treatment-specific model selection for the semiparametric covariate adjustment methods to handle this complication.

Our semiparametric adjustment methods have been developed for trials carried out according to a standard randomization scheme. As reviewed in Aim 3 of Project 1, surgical and other trials may be complicated by slow accrual due to reluctance of some subjects to be randomized, and alternative partial randomization designs have been proposed for this setting; e.g., in the comprehensive cohort design, subjects with preferences are offered their preferred treatments while those without are randomized. The selection bias inherent in the preference group must be taken into account in the analysis. In our final sub-aim, under appropriate assumptions we make explicit, we propose semiparametric covariate adjustment methods for inferences on treatment effect that exploit auxiliary baseline covariates to both improve efficiency and adjust for selection bias.

3.2 Aim 2: Methods for Primary and Longitudinal Analyses in the Presence of Drop-out

Subject drop-out is common in cancer clinical trials, leading to a so-called monotone missingness pattern where data are observed on the subject only until the time of drop-out and are missing thereafter. As noted in Section **2**, if the analyst is willing to believe that information observed prior to drop-out is sufficiently rich so that the probability of drop-out at any time may be explained as a function solely of this observed information, then the assumption that the unobserved data are missing at random (MAR) may be justified. Of course, whether or not a MAR mechanism holds cannot be determined from the observed data. When MAR is not deemed realistic, the mechanism must be modeled based on variables that are not observed, so the validity of the models cannot be directly checked, and sensitivity to the model choice must be evaluated; in Aim 4, we propose general, rigorous methods for this purpose. The MAR assumption will be plausible in many trials where high quality information is recorded; even if it is not realistic, it provides a key benchmark for any missing data analysis, and, accordingly, inferential methods under MAR are of considerable interest.

Primary analyses focused on treatment comparisons as in Aim 1 are often complicated by missingness of the outcome due to drop-out. The methods to improve efficiency of these inferences in Aim 1 based on "adjusting" for auxiliary baseline covariates assume that the outcome is observed for all subjects. When drop-out renders the outcome missing for some subjects, if one is willing to make the MAR assumption that the missingness may be explained by baseline and intervening auxiliary information up to the time of drop-out, then, using the theory of semiparametrics (Tsiatis, 2006), it is possible to extend the methods in Aim 1 to exploit this auxiliary

information to both enhance precision and take appropriate account of the missingness due to drop-out. In our first sub-aim, we will develop a framework for inference on general treatment effects in this setting.

Studies of longitudinal markers, such as a quality of life (QOL) measure or immune response in an investigation of a cancer vaccine, are a feature of many cancer clinical trials. For example, one of the objectives of E1694 was to investigate whether or not there is a treatment difference in QOL score, which was ascertained at baseline and several pre-specified evaluation visits thereafter; see Section **3.3**. When some subjects drop out, longitudinal marker and other information recorded at each visit are available on a subject only up to the time of drop-out. Under the MAR assumption, inference based on likelihood methods has been advocated on grounds of efficiency and, in the current context, the fact that it does not require specification of the drop-out mechanism but only of the model assumed to govern the full data that were intended to be collected. However, if this model is incorrectly specified, biased inferences may result. Likewise, Robins, Rotnitzky, and Zhao (1995) proposed methods based on weighting of observed data by the inverse probability of not dropping out, which require modeling the drop-out mechanism (as a function of the observed data) rather than the full data but which may yield biased inferences if the drop-out model is incorrect. These issues have led to the recent focus on "doubly robust" methods (e.g., Bang and Robins, 2005). Doubly robust estimators require specification of both a full data model and a missingness model; however, they have the desirable property that they will be consistent and asymptotically normal if only one or the other of these models is correctly specified. It has thus been argued that this affords protection against misspecification not enjoyed by other methods, allowing the analyst two chances to "get it right." Nonetheless, there has been vigorous debate over the performance of some doubly robust estimators, framed in a simple non-longitudinal setting (Tan, 2006, 2007; Kang and Schafer, 2007; Tsiatis and Davidian, 2007), namely, that they can perform poorly when both models are only "slightly" incorrect. Recently, in this simple context, we have developed new doubly robust estimators that are relatively more efficient and exhibit superior robustness to slight modeling mishaps (Cao, Tsiatis, and Davidian, 2009) than existing competitors. This inspires our second sub-aim, to extend these ideas to the more complicated longitudinal setting with MAR drop-out. This nontrivial extension will lead to improved performance relative to competing doubly robust estimators of Robins et al. (1995) and Bang and Robins (2005), providing analysts of cancer trials with a robust and efficient option for inference from longitudinal studies under MAR drop-out..

3.3 Aim 3: Diagnostic Measures for Longitudinal and Joint Models in the Presence of Missing Data

As discussed in Section **3.2**, an objective in many cancer trials is to carry out analyses focused on treatment differences with respect to a longitudinal marker such as a QOL measure; a further goal is to examine the association between a longitudinal marker and a primary clinical endpoint, typically a time-to-event. This is the case in several recent CALGB phase III trials in which the Aim Leader, Dr. Ibrahim, is involved, including CALGB 9221, comparing subcutaneous 5-azacytidine vs. observation in myelodysplastic syndromes (MDS); analysis of QOL in this study played a key role in FDA approval for 5-azacytidine in MDS. Investigating the association between QOL and survival, and in particular whether or not higher QOL is associated with increased survival, also played an important part in other trials, including CALGB 90401, a double-blinded, placebo controlled trial comparing docetaxel and prednisone with and without bevacizumab in men with hormone refractory prostate cancer; and CALGB 49907, a trial of adjuvant chemotherapy with standard regimens, cyclophosphamide, methotrexate and fluorouracil (CMF) or doxorubicin and cyclophosphamide (AC), vs. capecitabine in women 65 years and older with node positive or node negative breast cancer. ECOG E1694 (Kirkwood, Ibrahim et al., 2001), mentioned in Sections **3.1** and **3.2**, was based on the premise that the GM2 ganglioside is a well-defined melanoma antigen, and anti-GM2 antibodies have been associated with improved prognosis (Livingston et al., 1994); thus, the study focused on evaluating vaccination with GM2 conjugated to keyhole limpet hemocyanin and administered with QS-21 (GMK) for 96 weeks against HDI, the current standard adjuvant therapy, for one year in high-risk melanoma patients. A subset of the subjects also participated in the QOL study and were asked to complete QOL questionnaires at baseline, 1 month, and 6 months, and at approximately 6-month intervals thereafter; resulting longitudinal QOL measures including the total QOL score from the Functional Assessment of Cancer Therapy, Biologic Response Modifiers/CIS-Retinoic Acid+Pain Symptoms form (FACT BRM/CRA+PS). The trial demonstrated a significant benefit of HDI versus GMK for both relapse-free and overall survival, and characterizing the association between QOL and immune response within the vaccine arm and between QOL and survival in high-risk melanoma patients assigned to highly toxic HDI therapy were important subsequent objectives.

These studies exemplify settings in which joint models for longitudinal and survival data (Hogan and Laird,

1997; Henderson, Diggle, and Dobson, 2000; Ibrahim, Chen, and Sinha, 2001; Brown and Ibrahim, 2003; Tsiatis and Davidian, 2004; Diggle, Farewell, and Henderson, 2007) are an appropriate analytic framework. These models, implemented via likelihood techniques, have become popular in cancer clinical trials for this purpose, especially those involving QOL or vaccine studies. However, because they require linking postulated models for both longitudinal marker progression and a possibly censored time-to-event endpoint, they are of necessity complex. Most of the associated literature, such as that cited above, focuses on inference within a specified joint model, but there is little work on diagnostic tools for assessing validity of model assumptions (Diggle et al., 2007), violation of which could lead to misleading conclusions, nor for investigating goodness-of-fit and whether or not data from certain subjects or groups of subjects exercise undue influence on the fit. As joint models enjoy increasing application in cancer studies, development of such tools is essential.

As noted previously, a further complication is missing data due to drop-out and other reasons. In E1694, response to QOL questionnaires at each visit was not compulsory, and some subjects refused to or otherwise did not complete the forms; 42.6% of subjects had at least one QOL score missing, and the extent of missing QOL scores became more pronounced at each visit, with 10.8%, 16.6%, 22.0%, and 26.7% of subjects missing these at baseline, 1 month, 6 months, and 1 year. Moreover, as discussed in Section **3.1**, key auxiliary baseline covariates, ulceration of the tumor and Breslow thickness, were also missing for some subjects. Overall, close to 60% of cases had some form of missing outcome and/or covariate. Further complicating matters, the MAR assumption may not be realistic; e.g., a QOL score may be missing because the subject either had an excellent or, more likely, poor quality of life at the time, as would have been reflected in the missing QOL score. Similarly, missingness of baseline covariates might also be nonignorable; if insufficient tissue was available after diagnostic pathology, then submission of pathologic materials was not required, so that missingness of data concerning the primary tumor is most likely related to the amount or size of the resection.

E1694 demonstrates that diagnostic tools for assessing the relevance of the component longitudinal and survival models in joint models must also incorporate assumed models for possibly nonignorably missing data. Approaches to model assessment and influence analysis for complete data settings, such as local influence and case-deletion measures, goodness of fit statistics, and schemes for perturbing model assumptions (e.g., Cook and Weisberg, 1986; Cook, 1986) are available, but simply deleting cases with missing data or imputing missing data on an ad hoc basis in order to apply these to joint models is not appropriate. Very different cases might be identified as influential when analysis is limited to only subjects with complete data versus an analysis that uses all the cases, and the strength of evidence against model assumptions may be diminished if observed data from all subjects are not considered. These issues can be especially problematic when the fraction of missing data is moderate to high. A further challenge with any measure of case influence with unbalanced longitudinal data is that "size matters" (Critchley et al., 2001); i.e., case-deletion and perturbation schemes are sensitive to the size of the data cluster for a subject. Difficulty also arises in development of goodness-of-fit statistics when many complete and incomplete covariates are present (Lin, Wei, and Ying, 2002; Zhu, 2005). Although there is work on sensitivity analyses for any complex model (Rotnitzky, Robins, and Scharfstein, 1998; Verbeke and Molenberghs, 2000; Troxel, Ma, and Heitjan, 2004; Copas and Eguchi, 2005), there is a lack of a rigorous approach for selecting an appropriate scheme to perturb complex joint models for longitudinal and survival data and for assessing the global and local influences of such perturbations.

We will develop diagnostic measures for joint models that address these challenges, providing tools to assist cancer trial analysts not only in choosing an appropriate joint model for addressing a particular study but for mining longitudinal studies from cancer clinical trials for new insights on relationships more generally, which we will demonstrate through their application to the three recent CALGB studies above.

3.4 Aim 4: Inferences for Sensitivity Analyses of Missing Data

As we have emphasized, drop-out and missing outcomes and covariates are commonplace in cancer clinical trials, as in ECOG E1694, discussed in Section **3.1–3.3**; the issues are similar in E1684, another phase III trial in high-risk melanoma patients randomized to either HDI or follow-up without intervention (Kirkwood et al. 1996; Chen and Ibrahim, 2006). In this trial, a primary outcome is occurrence of relapse, and roughly 20% of subjects had missing covariates, including Breslow thickness and number of affected nodes. In these settings, whether or not such missingness is ignorable may be speculative, as described for E1694, leading the analyst to consider models for the missingness as a function of the unobserved data.

Sensitivity analysis has been widely advocated in such missing data scenarios where the missingness mech-

anism may depend on the unobserved data (e.g., Little and Rubin, 1987; Scharfstein, Rotnitzky, and Robins, 1999; Kenward, Goetghebeur, and Molenberghs, 2001; Rotnitzky et al, 2001). As we noted in Section **3.2**, under such informative missingness, the missing data cannot be ignored, and valid inferences are predicated on the correctness of assumed models for the missingness mechanism as a function of unobservable quantities, which cannot be checked using the observed data. In many cases, assumed such models may be "overparameterized," so that standard estimation techniques may fail due to the model being non-identifiable. This is problematic, as it may not be known a priori whether the model is identifiable. The common approach of postulating a range of models and inspecting the results to gain insight into sensitivity to model assumptions is not only subject to these issues but also is difficult to interpret. A sensitivity analysis methodology that supports principled, formal synthesis of evidence regarding the missingness mechanism would be particularly useful.

Two general strategies for sensitivity analysis have been discussed in the literature, local and global, which involve frameworks incorporating a "sensitivity parameter" representing departures from ignorability. A local sensitivity approach assesses the impacts of uncertainties on inferences over a range of models specified by the sensitivity parameter in a small "neighborhood" of a known value of this parameter (Copas and Eguchi, 2001; Verbeke et al., 2001; Troxel et al., 2002; Todem, Kim, and Lesaffre, 2006). One may assess the effects of small perturbations of the ignorable model in the directions of nonignorable models. If estimates of parameters of interest are locally insensitive to these values, then, under the assumption of small violations of ignorability, the desired inferences can be reasonably be carried out assuming ignorable missingness. Such methodology is useful but does not permit assessments of large deviations of the sensitivity parameters on inferences. Global analyses focuses on the impact of such deviations. Under qualitative assumptions regarding the missingness mechanism, it may be possible to derive bounds on the behavior without imposing further parametric assumptions (Manski, 2003; Horowitz and Manski, 2006). Alternatively, more structured parametric models for the missingness mechanism are often employed in practice on grounds of conceptual simplicity and ease of implementation, with a small number of easily interpreted sensitivity parameters (Shepherd et al., 2006; Vansteelandt et al., 2006). A limitation is that formal inference is not available. The usual strategy is to present parameter estimates at various values of the sensitivity parameters, accompanied by pointwise confidence intervals based on the assumption of a correctly specified model. Inference is ad hoc, ignoring that multiple tests are conducted and that the model may be misspecified.

We will develop an approach to sensitivity analysis that attempts to formalize the objectives of such ad hoc techniques. Instead of informally reporting results at particular values of the sensitivity parameters, corresponding to particular model assumptions that are unverifiable, we will develop a rigorous approach to combining information across such models to enable a conservative synthesis of evidence, without requiring that any of those models be correct. We will focus on the context of cancer clinical trials, but the approach will be broadly applicable to general studies and settings involving potentially nonignorable missingness.

4 PRELIMINARY STUDIES

4.1 Investigators

The research will be carried out by a highly qualified team of investigators from all three institutions. Marie Davidian of North Carolina State University (NCSU), a PD/PI for the overall program project, will serve as Project Leader. She has expertise in longitudinal data analysis and joint modeling of longitudinal and survival data (Song, Davidian, and Tsiatis, 2002; Tsiatis and Davidian, 2004), missing data and causal inference methods (Lunceford and Davidian, 2004; Davidian et al., 2005; Tsiatis and Davidian, 2007; Cao et al., 2009), and clinical trials (Tsiatis et al., 2008; Zhang et al., 2008), making her ideally suited to lead the project. Anastasios Tsiatis (NCSU) and Joseph G. Ibrahim (University of North Carolina, UNC) will serve as Project co-Leaders. Dr. Tsiatis is an internationally recognized expert on semiparametric theory and missing data (Tsiatis, 2006) and joint models (Tsiatis, DeGruttola, and Wulfsohn, 1995; Wulfsohn and Tsiatis, 1997; Song et al., 2002; Tsiatis and Davidian, 2004). Dr. Ibrahim is a world authority on missing data (e.g., Chen and Ibrahim, 2006), joint models (Brown and Ibrahim, 2003; Chi and Ibrahim, 2007), and model diagnostics (Zhu et al., 2007, 2008).

Dr. Davidian will lead Aim 1, working with Dennis Boos (NCSU), Howard Bondell (NCSU), Sin-Ho Jung (Duke), Len Stefanski (NCSU), Anastasios A. Tsiatis (NCSU), and Helen Zhang (NCSU). This aim will involve considerable work on model selection. Drs. Boos and Stefanski are both experts in model selection and are the originators of the FSR approach (Wu et al., 2007; Boos et al., 2008). Dr. Boos' expertise in bootstrap

methodology (e.g., Boos, 2003) will be exploited in the second and third sub-aims. Model selection is Dr. Bondell's research focus (Bondell and Reich, 2008a, 2009; Bondell and Li, 2009); Dr. Zhang is also a recognized authority in this area (e.g., Zhang et al., 2004; Zhang, 2006; Lin and Zhang, 2007; Zhang and Lu, 2007; Zou and Zhang, 2009) and will contribute heavily to the first, second, and third sub-aims. Semiparametric theory and missing data methods will also play a key role; Dr. Tsiatis will provide expertise and is co-developer of the adjustment methods studied in the first three sub-aims (Lu and Tsiatis, 2008; Tsiatis et al., 2008; Zhang et al., 2008). Dr. Jung has been involved in cancer clinical trials for over 15 years and has published numerous papers on trial design and analysis methods (e.g., Jung et al., 2005; Jung, 2008; Jung, Kim, and Chow, 2008); accordingly, he will bring an important perspective to the research.

Efforts on Aim 2 will be led by Dr. Tsiatis, whose expertise in semiparametrics and missing data will be critical and whose work with Dr. Davidian on adjustment methods and doubly robust estimators (Tsiatis and Davidian, 2007; Cao et al., 2009) forms the basis for both sub-aims. Dr. Davidian will also lend expertise in these areas. Dr. Ibrahim will provide additional expertise on missing data, Dr. Jung on clinical trials issues, and Drs. Bondell and Zhang will contribute expertise on model selection. Aim 3 will be led by Dr. Ibrahim, working with Drs. Davidian and Tsiatis. Drs. Davidian, Ibrahim, and Tsiatis will draw on their expertise in joint models and missing data, and Dr. Ibrahim on his background in diagnostic methods. Aim 4 will be led by Jason Fine (UNC), who is an expert on sensitivity analysis (Todem et al., 2006; Todem and Fine, 2008) as well as survival analysis. Jianwen Cai (UNC) and Dr. Tsiatis will work with Dr. Fine, lending expertise on missing data. Dr Cai will bring her expertise on missing data/measurement error and joint model methods (Greene and Cai, 2004; Schaubel and Cai, 2006; Liu, Zhou, and Cai, 2009) to this work.

Pre-existing collaborations and connections among subsets of these investigators ensure that this team will work effectively within and across institutions and aims. Trans-institutionally, Drs. Davidian and Tsiatis work with Drs. Cai, Fine, and Ibrahim, through their membership in the Biostatistics Core of the UNC Clinical and Translational Science Award, and both are members of the UNC Center for Innovative Clinical Trials, directed by Dr. Ibrahim. Drs. Fine and Tsiatis have also published together (e.g., Fine and Tsiatis, 2000). Drs. Davidian and Tsiatis are also adjunct faculty at Duke, where they spend one day a week and interact with Dr. Jung. Drs. Bondell, Boos, Cai, Fine, Ibrahim, and Zhang are or have been associate editors for *Biometrics* and worked regularly with Dr. Davidian in her capacity as Coordinating Editor and Executive Editor. Within NCSU, Drs. Bondell, Boos, Stefanski, and Zhang are members of a research group on model selection and have published together (e.g., Boos et al., 2008); more generally, all NCSU investigators publish together (e.g., Huang, Stefanski, and Davidian, 2009) and have or are co-directing doctoral dissertations (e.g., Boos-Stefanski, Davidian-Stefanski, Davidian-Tsiatis, Bondell-Zhang, Davidian-Zhang). Similarly, at Drs. Ibrahim, Cai, and Fine are members of the Lineberger Comprehensive Cancer Center, and collaborate on several substantive projects.

Neil Spector, MD, Associate Professor of Medicine (Oncology) at Duke University Medical Center, will work closely with the team, offering subject-matter guidance from his vantage point as Director of Translational Research in Oncology and a Co-Director of the Experimental Therapeutics Program at the DCCC. He is former Director of Exploratory Medical Sciences in Oncology at GlaxoSmithKline (GSK). While at GSK, he focused on elucidating the biological effects of targeted agents in preclinical models in order to identify novel molecular biomarkers with which to develop cancer therapeutics. Dr. Spector directed the Nelarabine project, which received FDA approval for the treatment of childhood acute lymphoblastic leukemia; and he directed the development of lapatinib (Tykerb), an important epidermal growth factor receptor inhibitor for treating women with advanced Her2 positive breast cancer. Dr. Spector's experience leading these therapies from bench to clinic trials to FDA approval will be invaluable in guiding the practical relevance and application of the methods.

4.2 Preliminary Studies

Aim 1: Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates. In Zhang et al. (2008), Drs. Davidian and Tsiatis outline a broad framework for covariate adjustment in clinical trials for general types of outcomes and two or more treatments based on taking a semiparametric perspective that places minimal restrictions on the nature of the data; this framework is presented formally in Section **5.1**. They derive classes of all consistent estimators and tests that make use of the data on all of outcome, treatment assignment, and auxiliary baseline covariates and identify the form of the most efficient; i.e., that theoretically leading to the most precise inferences on the effect of interest with the available data. Although the paper did not appear in *Biometrics* (published quarterly) until September 2008, it is among the most heavily downloaded

articles appearing in the journal in all of 2008, reflecting the considerable interest the approach has generated among clinical trials biostatisticians. In Lu and Tsiatis (2008), such adjustment methods are developed for comparison of two survival distributions under the usual assumption of noninformative censoring; Tsiatis et al. (2008) describe the approach in detail for practitioners in the routine special case where the parameter of interest is the difference in two treatment outcome means. In all three papers, simulation studies demonstrate impressive gains in efficiency over unadjusted and competing adjusted analyses. In the latter simple setting, on which we focus here for definiteness, the efficient estimator for the difference is obtained by incorporating correct representations of the conditional expectations of outcome given the covariates in each treatment group. Thus, roughly speaking, the most precise inferences will be attained when models for these conditional expectations are developed that come as close as possible to their true forms, which motivates our study of competing model selection techniques to identify models that achieve this.

As noted in the citations in Section **4.1**, Drs. Bondell, Boos, Stefanski, and Zhang have all developed methods for model selection. To date, only limited studies of use of model selection in the context of the semiparametric covariate adjustment methods are available. Simulations by Tsiatis et al., (2008) and Zhang et al., (2008) consider only traditional forward selection with a single fixed "α to enter" tuning constant in limited scenarios with continuous and binary outcomes. In larger sample sizes (on the order of > 600), there is no discernible effect of model selection on inferences on treatment effect, as predicted by the theory. However, in smaller sample sizes (200 or less), they found in simulations not reported that coverage of 95% confidence intervals for the true effect falls short of the nominal level, suggesting that optimistic inferences will result if the use of model selection is not taken into appropriate account. Boos et al. (2008) use their FSR model selection methods with the semiparametric methods in Tsiatis et al. (2008) to compare treatment means in a large HIV clinical trial; through various subsampling exercises with these data, they conclude that the FSR method may offer an improvement over traditional methods; see Section **5.1**. Because of simplicity and ease of use of the version of FSR in Boos et al. (2008) in practice, we will focus on it heavily.

The selection bias arising in the preference arm of partially randomized studies is analogous to the confounding encountered in the analysis of treatment effects from observational studies. Drs. Davidian and Tsiatis have considerable experience in this area (e.g., Anstrom and Tsiatis, 2001; Lunceford and Davidian, 2004; Johnson and Tsiatis, 2005). The proposed methods for these trials in the fourth sub-aim will merge these techniques with the principles of the semiparametric adjustment methods above.

Aim 2: Methods for Primary and Longitudinal Analyses in the Presence of Drop-out. Missingness due to drop-out may be cast in its most general setting as a censored data problem. Censored data methods for survival analysis have been studied extensively assuming noninformative censoring; however, in reality, the censoring mechanism may be more complex, and may be thought of as a special (albeit more complicated) example of continuous-time monotone missingness assumed to be MAR. From this vantage point, it is possible to consider realistic censoring mechanisms where censoring may depend on auxiliary information and to consider censoring in contexts other than survival analysis where the outcome of interest is not necessarily a time-to-event and occurs after some "lag time." Dr. Tsiatis has developed methodology for such lag-time problems in Zhao and Tsiatis (1997, 1999, 2000), studying censored QOL measures, and Bang and Tsiatis (2000, 2002) in the setting of censored medical costs. A general strategy for deriving estimators for parameters of interest when data are missing due to such censoring using the theory of semiparametrics is outlined in Sections 9.3 and 10.4 of Tsiatis (2006). This strategy is applicable to the specific problems involving drop-out mechanisms assumed to depend on observed auxiliary covariates to be studied in this aim, where the outcome of interest may be a time-to-event or other censored time-lagged measure.

Considering the simple problem of estimating a mean outcome when the outcome may be MAR, Kang and Schafer (2007) present examples where doubly robust estimators advocated by Bang and Robins (2005) performed poorly when both the models for the missingness mechanism and the full data are misspecified even slightly. Implementation of the doubly robust estimators requires estimation of the parameters in both postulated models. Kang and Schafer (2007) consider only the case where those in the full data model were estimated by least squares. In Cao et al. (2009), Drs. Davidian and Tsiatis demonstrate that there are many possible doubly robust estimators depending on how the full data model parameters are estimated; see also related work by Tan (2006). They show that the estimator for these parameters leading to the optimal doubly robust estimator for the mean is not least squares but is rather a weighted least squares estimator using a specific weighting

scheme. Cao et al. (2009) give an argument for why such an estimator should not only be more efficient than doubly robust estimators using least squares when the model for the missingness mechanism is correct but also should be more robust to model misspecification, and they present empirical studies based on simulation scenarios used by Kang and Schafer (2007) that show dramatic such improvements when using their doubly robust estimator with weighting. The principles used in this simple case will be exploited in the more complex setting of monotone MAR missingness due to drop-out to derive doubly robust estimators for general quantities of interest in longitudinal studies in Section **5.2** and should lead to similar improvements.

Aim 3: Diagnostic Measures for Longitudinal and Joint Models in the Presence of Missing Data. Motivated by applications in the biomedical, social, and ecological sciences, Dr. Ibrahim and colleagues have an extensive record of systematic study of important statistical models, including generalized linear, longitudinal, and survival models with and without missing outcome and/or covariate data; spatial models for ecological and neuroimaging data; and joint models for longitudinal and survival data (Gu and Zhu, 2001; Ibrahim et al., 2001; Chi and Ibrahim, 2007). Especially relevant to Aim 3 are their significant contributions to model diagnostics (Zhu et al., 2007; Zhu et al., 2008). They have developed a differential-geometric framework, the perturbation manifold, for carrying out sensitivity analysis of any parametric model (Zhu et al., 2007), showing that the metric tensor of the perturbation manifold provides important information for selecting an appropriate perturbation of a model. They also use the concept of a geodesic on the perturbation manifold to introduce new local influence measures for any objective functions at any point. The development of the perturbation manifold allows explicit measurement of the amount of perturbation, the extent to which each component of a perturbation vector contributes, and the degree of orthogonality for the components of the perturbation vector. These advances are critical for formally carrying out sensitivity analysis in complicated models, such as hierarchical models.

Dr. Ibrahim and co-workers have developed case-deletion measures for assessing the influence of several observations for a variety of statistical models for missing data (Cho et al., 2009) and goodness-of-fit statistics based on local influence measures (Zhu and Zhang, 2004). They have also proposed a general local influence approach for assessing effects of minor perturbations to statistical models for missing data. The key idea is to generalize Cook's (1977, 1986) approaches to the conditional likelihood of the complete-data likelihood function in the expectation-maximization (EM) algorithm (Zhu and Lee, 2001). This not only dramatically reduces the computational burden of deriving case-deletion and the local influence measures, but also yields better interpretation of diagnostic results for hierarchical models such as longitudinal and multilevel models, compared with standard methods (Cook, 1977; Cook, 1986) and facilitates identification of influential observations and general model checking for many complicated models (e.g., Lee and Tang, 2004).

Dr. Ibrahim and colleagues have also developed a diagnostic procedure based on the empirical likelihood for combining estimating equations without assuming any parametric distributions (Zhu et al., 2008), which consists of case-deletion measures for assessing influence of individual observations, a local influence approach for assessing small perturbations to estimating equations, and a goodness-of-fit statistic for testing potential misspecifications of the equations. Based on the theory of empirical processes (van der Vaart and Wellner, 2000; Kosorok, 2008), they develop the statistic using multiple pseudo-residual processes and devise a resampling method to approximate its null distribution to calculate the critical value, which corrects the multiple comparisons to control for the family-wise error rate while accounting for correlations among all estimating equations.

Aim 4: Inference for Sensitivity Analyses Techniques of Missing Data. In current work (Todem and Fine, 2008), Dr. Fine has proposed sensitivity analysis methods in the special case of longitudinal analysis of a binary outcome, as in E1684, where there is potentially informative drop-out. Assuming a specific longitudinal random effects binary outcome model for which the probability of drop-out depends on the random effects (e.g., Wu and Carroll, 1988; ; Albert and Follmann, 2000; Ten Have et al., 2002), inducing informative missingness, formal inferences for sensitivity analysis are developed for maximum likelihood inference under this model. Because of identifiability concerns, the analysis profiles across the random effects parameter, which quantifies the extent of informative missingness, rather than estimating it. If δ is the sensitivity parameter derived from the random effects distribution, then $\delta = 0$ corresponds to noninformative missingness, and $\delta \neq 0$ defines informative missingness, where larger magnitude δ implies stronger informativeness. A conservative test of the null hypothesis of no treatment effect is carried out simultaneously across the support of δ using an infimum test statistic, taking into account that inferences are carried out simultaneously across the entire range of δ and that the model may be misspecified under certain values of δ, as might occur if a nonignorable model is fit, when, in reality, miss-

ingness is ignorable. In addition, simultaneous confidence bands for the identifiable parameter are proposed, enabling pointwise tests of the null that control the overall type I error rate. This enables evaluation of how the treatment effect changes as δ is varied, as well as providing confidence intervals that are valid simultaneously across all values of δ. Such simultaneous statements are the most rigorous approach to inference in a sensitivity analysis and formalize current ad hoc practices of informally inspecting evidence across models. They may be especially useful in scenarios where one may not know in advance whether a model is identifiable, in which case proceeding as if the model is identifiable may yield misleading results. A careful theoretical analysis of the estimator of the treatment effect parameter as a process in δ is conducted; because of the complexity of the distribution of the process, a bootstrap procedure is proposed for practical implementation. Preliminary empirical results of application of the methods to data from a clinical trial of fluvoxamine, a serotonin re-uptake inhibitor, in patients diagnosed with depression, obsessive-compulsive disorder, or panic disorder, are presented; such patients often suffer comorbidities that lead to drop-out (Burton, 1991, Kenward, Lesaffre, and Molenberghs, 1994; Lesaffre, Molenberghs, and Dewulf, 1996). Under the framework described above for the binary outcome of side effects, the null hypothesis of no temporal effect of treatment on side effects is evaluated by testing across a continuum of values for δ, each defining a different level of informative missingness, suggesting strong evidence of a time-varying effect of fluvoxamine across a wide range of values for δ.

The research detailed in Section **5.4** will build on and significantly extend these ideas to general models.

5 RESEARCH DESIGN AND METHODS

Some symbols may represent different quantities in each section. Plans for software development and dissemination and a timetable of activities for all four aims are presented at the end of this section.

5.1 Aim 1: Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates

Semiparametric Covariate Adjustment Framework. For simplicity, we consider a univariate outcome Y; the methods are also applicable to multivariate outcomes. Let the data for subject i from a k-arm randomized trial, $k \geq 2$, be $(Y_i, \boldsymbol{X}_i, Z_i)$, $i = 1, \ldots, n$, independent and identically distributed (iid) across i, where Y_i is outcome, $\boldsymbol{X}_i$ is a vector of baseline auxiliary covariates, and $Z_i = g$ indicates assignment to treatment g with known randomization probability $P(Z = g) = \pi_g$, $\sum_{g=1}^{k} \pi_g = 1$. Randomization guarantees $Z \perp\!\!\!\perp \boldsymbol{X}$, where "$\perp\!\!\!\perp$" means "independent of." Let β be a vector of parameters involved in making treatment comparisons. When $k = 2$ and interest is in a difference in treatment means, the quantity of interest is $E(Y|Z = 2) - E(Y|Z = 1)$, which equals β_2 in $E(Y|Z) = \beta_1 + \beta_2 I(Z = 2)$, $\beta_1 = E(Y|Z = 1)$, $\beta = (\beta_1, \beta_2)'$, and $I(\cdot)$ is the indicator function. When $k = 3$ and Y is binary, interest may focus on log odds ratios β_2 and β_3 for treatments 2 and 3 relative to a control treatment 1, expressed in $\text{logit}\{E(Y|Z)\} = \beta_1 + \beta_2 I(Z = 2) + \beta_3 I(Z = 3)$, $\beta = (\beta_1, \beta_2, \beta_3)'$, $\text{logit}(p) = \log\{p/(1-p)\}$. In general, the focus is on parameters β in a model describing aspects of the conditional distribution of Y given Z, which in more complex settings could involve additional nuisance parameters γ. Standard unadjusted analyses make inference on β based only on (Y_i, Z_i), $i = 1, \ldots, n$. Standard adjusted analyses also take into account $\boldsymbol{X}$, usually through a regression model, e.g., for continuous outcome in the first example, the analysis of covariance (ANCOVA) model $E(Y|\boldsymbol{X}, Z) = \alpha_0 + \boldsymbol{\alpha}_1' \boldsymbol{X}_I + \phi I(Z = 2)$, where ϕ is the adjusted estimator for β_2 above, and $\boldsymbol{X}_I$ contains functions of elements of $\boldsymbol{X}$ pre-specified for the analysis or deemed "important" in the sense of being associated with Y. This framework links inference on treatment effect to $\boldsymbol{X}$. Although ϕ is a consistent and asymptotically normal estimator for β_2 in this simple setting, in situations like the second example, estimation of the log odds ratios β_2, β_3 by postulating an analogous (nonlinear) logistic regression model including $\boldsymbol{X}_I$ would not necessarily lead to consistent estimation; see Zhang et al. (2008).

Zhang et al. (2008) derive alternative adjustment methods by considering a semiparametric model for $(Y, \boldsymbol{X}, Z)$ placing no restrictions on the joint densities that could have generated these data except that $Z \perp\!\!\!\perp \boldsymbol{X}$ and the π_g are known; see Zhang et al. (2008) for details. Under this model, they use semiparametric theory to identify the class of all unbiased estimating functions for $\theta = (\beta', \gamma')'$ in the assumed model for the relevant aspects of the conditional distribution of $Y|Z$ based on using all the data $(Y, \boldsymbol{X}, Z)$. Unbiased estimating functions suggest estimating equations that lead to consistent and asymptotically normal estimators. They show that elements of the class of all such estimating functions have the form $m^*(Y, \boldsymbol{X}, Z; \theta) = m(Y, Z; \theta) - \sum_{g=1}^{k} \{I(Z = g) - \pi_g\} a_g(\boldsymbol{X})$, where $m(Y, Z; \theta)$ is the estimating function one would use to estimate the desired elements of β in an unadjusted analysis by solving in θ $\sum_{i=1}^{n} m(Y_i, Z_i; \theta) = 0$; and $a_g(\boldsymbol{X})$, $g = 1, \ldots, k$, are arbitrary

functions of $\boldsymbol{X}$. Solving $\sum_{i=1}^{n} m^*(Y_i, \boldsymbol{X}_i, Z_i; \theta) = 0$ with appropriate functions $a_g(\boldsymbol{X})$ yields an adjusted estimator for β; Zhang et al. (2008) show that the optimal choice leading to the most efficient such estimator is $a_g(\boldsymbol{X}) = E\{m(Y, Z; \theta)|\boldsymbol{X}, Z = g\}$. They argue via semiparametric theory that using linear models to represent these treatment-specific regressions, whether or not they are correct, and fitting these by least squares, guarantees a gain in efficiency over the unadjusted analysis. If whatever parametric models one postulates to represent the regressions contain the true functions $E\{m(Y, Z; \theta)|\boldsymbol{X}, Z = g\}$ of $\boldsymbol{X}$, then the greatest possible gain will result. Zhang et al. (2008) argue that, in general, the closer the predictions from these models are to those from the true functions of $\boldsymbol{X}$, the closer the resulting estimator for β will come to achieving the precision of the optimal estimator. They propose an adaptive strategy for solving the equations and advocate standard errors for the resulting estimator obtained by the usual sandwich technique (Stefanski and Boos, 2002). Analogous development is possible for optimal adjusted test statistics; for definiteness, we focus on estimation here.

These results are very general; in most situations, like those above, because of the simple form of $m(Y, Z; \theta)$, finding the most efficient estimator reduces to modeling $E(Y|\boldsymbol{X}, Z = g)$ for each g. In fact, for the difference of $k = 2$ treatment means above, the optimal adjusted estimator $\widehat{\beta}_2$ is asymptotically equivalent to an expression of the form

$$\overline{Y}_2 - \overline{Y}_1 - \sum_{i=1}^{n} \{I(Z_i = 2) - n_2/n\}\{n_1^{-1} E(Y|\boldsymbol{X}, Z = 1) + n_2^{-1} E(Y|\boldsymbol{X}, Z = 2)\} \tag{5.1}$$

where $n_g = \sum_{i=1}^{n} I(Z_i = g)$, and $\overline{Y}_g = n_g^{-1} \sum_{i=1}^{n} I(Z_i = g) Y_i$, $g = 1, 2$. Thus, the optimal adjusted estimator achieves efficiency gains by "augmenting" the usual unadjusted estimator $\overline{Y}_2 - \overline{Y}_1$ by a mean-zero term involving the treatment-specific regressions. For simplicity of exposition in the sequel, we focus on this setting and (5.1); during the project period, we will consider the general case.

Model Selection Methods for Covariate Adjustment. The theory implies that estimation of parameters in the postulated models for $E(Y|\boldsymbol{X}, Z = g)$, $g = 1, 2$, whether or not the models are correct, will have no effect asymptotically on consistency or precision of the resulting estimator $\widehat{\beta}_2$, and as above, efficiency gain over the unadjusted estimator is guaranteed. Thus, theoretically, one could simply use models involving all elements of $\boldsymbol{X}$, including, e.g., all linear, quadratic and first-order interactions. However, the theory is likely optimistic in the sample sizes seen in many cancer trials, suggesting use of model selection techniques to "pare down" the models to incorporate only "important" covariate effects predicting outcome in order to achieve the most dramatic efficiency gains.. A small simulation illustrates. Here, Z_i are iid Bernoulli with $\pi_1 = \pi_2 = 0.5$; outcome is generated as $Y_i = \zeta_{0g} + \zeta_{1g}' \boldsymbol{X}_i + \epsilon_i$ for $\epsilon \sim N(0, \sigma^2)$, so $\beta_2 = \zeta_{02} - \zeta_{01} - (\zeta_{12} - \zeta_{11})' E(\boldsymbol{X})$; $\boldsymbol{X}_i$ is 20-dimensional with a mix of correlated and independent normal and Bernoulli elements; and ζ_g are such that only 4 or 5 of the 20 covariates have non-zero coefficients, $g = 1, 2$, the set of such covariates overlaps only slightly in each group, and 16 or 15 covariates are "unimportant." For $n = 100$ and each of 5000 Monte Carlo data sets, we computed the unadjusted estimator (UN); (5.1) with $E(Y|\boldsymbol{X}, Z = g) = \zeta_{0g} + \zeta_{1g}' \boldsymbol{X}$ including all of $\boldsymbol{X}$, fitted by least squares (FULL); (5.1) with these models for $E(Y|\boldsymbol{X}, Z = g)$ fitted using traditional forward selection with entry $\zeta = 0.05$ (FOR) and the adaptive LASSO (ALAS, Zou, 2006); and as benchmarks (unachievable in practice) the true models fitted by least squares (TRUE) and with the true values of ζ_{0g}, ζ_{1g} substituted (IDEAL). All estimators for β_2 were unbiased; for brevity, we show only Monte Carlo efficiency based on mean square error relative to UN, and coverage probability of Wald 95% confidence intervals for β_2 using sandwich standard errors:

	UN	FULL	FOR	ALAS	TRUE	IDEAL
Rel. Eff.	1.00	0.97	1.14	1.18	1.20	1.25
Cov. Prob.	0.95	0.94	0.91	0.93	0.94	0.95

The results show that, contrary to the theory, overfitting of the models negates efficiency gains. Using model selection techniques to eliminate unimportant effects results in impressive gains approaching those achievable if the true models were known. However, coverage of the associated intervals falls short of the nominal, and an effect of estimating the parameters in the true models is evident. Standard competing methods, e.g., ANCOVA, will suffer similar performance, as they are in fact equivalent to estimators in our class but do not have the appealing feature of separating modeling and effect estimation in their implementation; see Tsiatis et al. (2008). While for continuous outcomes as here the effect of estimation and model selection disappears for sample sizes in the range of $n = 400$, in other settings, e.g., binary outcome and odds ratios, the sample sizes required for this are much larger (Zhang et al., 2008), and we conjecture similar behavior for censored time-to-event problems.

This evidence demonstrates that developing the treatment-specific models that effect the gains in efficiency is critical; a "kitchen sink" approach wherein all covariates are simply included in order to sidestep the effort involved in model-building could backfire, particularly with a large number of covariates. Indeed, in cases where the number of covariates approaches or exceeds n, as may arise with genomic information, this strategy is untenable, and model selection is a necessity. Nonnegligible efficiency gains are possible, but appealing to theory that does not take into account model building can compromise inference. These general observations underscore the need for the large-scale, comprehensive study of the use of available model selection methods in this context and the accompanying development of methods for "corrected" inference in the next sub-aim we propose. Accordingly, our first step in developing guidance and recommendations on implementing the semi-parametric covariate adjustment methods in cancer clinical trials will be an extensive investigation of model selection techniques in this context. Because of the need to consider numerous factors that may impact performance, as detailed below, such a study, which is an essential step toward promoting use of these methods, will of necessity be empirical. A study of this breadth would be difficult to carry out without access to extensive programming and computational resources and varied expertise. The Program Project is the ideal and perhaps only setting in which to conduct an investigation of this scope.

Assume that, for each $g = 1, 2$, there is a set of functions of the elements of $\boldsymbol{X}$, denoted $\boldsymbol{X}_{Ig}$, such that the true $E(Y|\boldsymbol{X}, Z = g)$ is a function of $\boldsymbol{X}$ only through $\boldsymbol{X}_{Ig}$; thus, e.g., in a linear model $E(Y|\boldsymbol{X}, Z = g) = q_g(\boldsymbol{X}, \zeta_g)$, a number of the true elements of ζ_g would be zero, corresponding to functions of $\boldsymbol{X}$ that are not "important." Most (parametric) model selection methods, make an assumption on the form of q_g, e.g., that it is linear in terms involving elements of $\boldsymbol{X}$ such as linear, quadratic, and first-order interaction effects, and then seek to include the "important" subset of these while not "over-including" terms that are unnecessary; ideally, the resulting model would include only $\boldsymbol{X}_{Ig}$. The extent to which these goals are met along with how well error of prediction of outcomes by resulting models is controlled are the usual criteria by which the methods are evaluated. It is important to recognize, however, that the treatment-specific models to be developed in our context are of little interest in their own right; their sole function is to enhance efficiency of the estimated treatment effect. Intuitively, as noted by Zhang et al. (2008) above, models that yield predictions at each $\boldsymbol{X}_i$ that get as close as possible to the truth should correspond to the greatest increases in efficiency. Whether or not they achieve the best performance in terms of identifying $\boldsymbol{X}_{Ig}$ accurately may or may not be critical.

We will consider a number of model selection methods; given our focus on providing guidance for practice, we will include but not limit to techniques for which software is available or that are otherwise convenient to implement, in order to promote the widest application. Traditional methods where a fixed tuning constant is employed, such as forward, backward, and stepwise selection, are widely available in software such as the regression procedures in SAS and the `leaps` package in R, and we will include them in their role as "off-the-shelf" methods favored by many practitioners. We conjecture that methods that are tuned adaptively may lead to better performance; accordingly, we will consider in addition several such methods. The LASSO, the adaptive LASSO, and SCAD, cited in Section **3.1**, are shrinkage methods involving penalizing the least squares or other objective function. The LASSO is similar to ridge regression but involves an L_1 rather than L_2 penalty; in implementation, the tuning constant, the penalty parameter, is chosen via a criterion such as cross-validation, generalized cross-validation, or inspection of information criteria such as the Akaike (AIC) or Bayesian (BIC) criteria. In contrast to traditional methods, where selection of variables is discrete, selection in the LASSO is continuous, and hence may be more stable; moreover, it is computationally feasible even with a large number of covariates. However, a purported drawback is that the LASSO does not possess the so-called "oracle property," meaning in linear regression that it can correctly select non-zero coefficients of predictors with probability converging to 1 and that the estimators of these coefficients are asymptotically normal with the same means and covariance matrix that they would have if their status had been known in advance. Other penalized methods, such as SCAD, do possess the oracle property (Fan and Li, 2001). However, the objective function is not convex and hence it is more difficult to compute; an R package is available. The adaptive LASSO proposed by Zou (2006) is similar to the LASSO but uses a weighted L_1 penalty with weights determined by an initial estimator of the regression parameter, and Zou (2006) has shown that the resulting estimator has the oracle property.

An alternative approach to tuning is the false selection rate (FSR) methods of Wu et al. (2007) and Boos et al. (2008). FSR is an approach to estimate the tuning constant of any model selection method, such as the "α to enter" of forward selection, so that the proportion of "unimportant" covariates for predicting outcome that enter the selected model is on average equal to a pre-specified small value, such as 0.05. When used with

forward selection to estimate "α to enter," FSR is a type of adaptive False Discovery Rate method (Benjamini and Hochberg, 2000) with intuitive appeal, leading to parsimonious models with good prediction error performance. The original version of FSR proposed by Wu et al. (2007) involves adding simulated noise variables to the real $\boldsymbol{X}$, monitoring when they enter a forward selection sequence, and using this information to tune the procedure, and hence can be time-consuming to compute. Boos et al. (2008) develop a "fast" approximation to FSR that does not require simulation and is hence easy to compute. During the project period, software for all-purpose implementation of Fast FSR in the context here will be developed.

In practice, all of the foregoing methods yield a model in which some predictors in the original set based on $\boldsymbol{X}$ are eliminated. Breiman (1996ab) has argued that some model selection procedures can result in very different models under slight perturbations of the data. To improve stability, his proposal of bagging (bootstrap aggregation) essentially averages models selected by any model selection strategy over bootstrap data sets. That is, in our context, assuming a linear model for treatment g, for the bth bootstrap data set, $b = 1, \ldots, B$, one would draw n pairs from $(Y_i, \boldsymbol{X}_i)$, $\{i : Z_i = g\}$, and run the variable selection method to obtain a model containing selected predictors $\boldsymbol{X}_{Ig,b}$, with corresponding estimated regression parameter $\widehat{\zeta}_{g,b}$ (with some elements zero corresponding to unselected variables). The final model is found by using the average $\overline{\zeta}_g^B = B^{-1} \sum_{b=1}^B \widehat{\zeta}_{g,b}$. Buhlmann and Yu (2002) show that bagging typically yields a more biased estimator for ζ_g than other methods but can yield improved prediction error, and Boos et al. (2008) present simulations showing that bagging Fast FSR works quite well in this regard. Because good predictions are so critical to efficiency gains for our covariate adjustment methods, we will consider bagging model selection techniques in our studies. Further impetus for studying bagging in the context of inference on the treatment effect is discussed below.

During the project period, we will study the foregoing model selection methods for use with the semiparametric covariate adjustment methods discussed above as well as the extension of these in the case of MAR drop-out discussed in the first sub-aim of Aim 2. We will also consider, as the current-practice competitor, standard adjustment methods such as ANCOVA with covariates and model forms specified a priori, as would be dictated by regulatory policy. In order that our study yield wide-ranging guidance, it will involve assessing performance of all of the methods over combinations of several key factors. Although we will begin by considering the difference of $k = 2$ treatment means for a continuous outcome, as above, we will study binary outcomes, censored time-to-event outcomes, and longitudinal outcomes, so that the parameter of interest, β, will have a different nature (e.g., log odds ratio, log hazard ratio) depending on the outcome. Note from the previous section that, in some more complex settings, rather than modeling treatment-specific regressions of the outcome itself on covariates, modeling will involve the "pseudo-outcome" $m(Y, Z, \theta)$, where, as in Zhang et al. (2008), θ is replaced by a preliminary estimate. A key factor will be sample size; we will consider small, moderate, and large studies reflecting those used in cancer clinical trials. The nature of the covariates is likely to play an important role. We will consider situations of only a few auxiliary baseline covariates to those involving high-dimensional information, including genomic information and the so-called "$p >> n$" case. Binary, more general categorical, and continuous covariates will be considered in combination, and the role of extent and nature of associations (e.g., correlation) among covariates and between covariates and outcomes will be systematically explored. Whether or not the true treatment-specific regressions are sparse in or involve combinations of several baseline covariates will be varied to allow evaluation of the importance of this issue in determining an appropriate selection strategy. The form of the true and assumed models is another important factor; linear and nonlinear models will be considered for both. True relationships will likely be nonlinear, as for binary outcomes, for example, and there is a potential that assumed models may be misspecified. A final important factor will be the approach for choosing the tuning constants for the model selection methods, with emphasis on the balance between model sparsity and variable selection consistency.

To develop our plan for this large-scale effort, we will first review data sets compiled by Core B as well as other cancer studies, to establish realistic settings and ranges for all of these factors. Settings representing "extreme" and "normal" ranges, as well as those for other factors that may emerge as important to study following this review, will be determined, and an initial set of "screening" simulations will be carried out using experimental design principles to gain insight into which regions of the factor space are associated with the most dramatic efficiency disparities among selection methods and possibly deleterious effects on inference on β due to both estimation and model selection. Subsequent simulations will be designed focusing on combinations and settings of factors identified as the most fruitful to study. We will compile the findings into a report providing detailed recommendations on choice and implementation of model selection techniques for covariate adjustment.

Inference Correcting for Model Selection. Identifying model selection methods that lead to the greatest gains in efficiency via covariate adjustment under situations likely to be encountered in cancer clinical trials is our main objective. However, as in our simulation above, such gains can be an accompanied by compromised inference on β based on standard, first-order asymptotic theory. Although the theory dictates that there is no effect (to first-order) of estimation of the parameters ζ_g nor of model selection in the postulated treatment-specific regression models, so that the expressions for standard errors do not involve terms explicitly taking into account uncertainty associated with these tasks, this is evidently optimistic. Undercoverage of confidence intervals is apparently due in part to estimated standard errors for $\widehat{\beta}_2$ for all of the FOR, ALAS, and TRUE cases that are too small. Based on our experience, we anticipate that our comprehensive study in sub-aim 1 will demonstrate this phenomenon more generally under configurations involving smaller n, suggesting a second-order effect of estimation of ζ_g and model selection not captured by the usual theory. We will thus study approaches to understanding and "correcting" for these effects.

A straightforward, brute-force approach to gaining insight on this effect is to carry out second-order calculations to deduce the effect of estimation of ζ_g ($p \times 1$), say, on the sampling variance of the estimator for β (e.g., Carroll, Wu, and Ruppert, 1988). To illustrate in a simple case, we focus on estimation of the single mean $\beta_1 = E(Y|Z=1)$ in the difference of two treatment means model above. Letting $\pi = \pi_1$ and $A = I(Z=1)$, it is straightforward to show in the case that the model $q_1(\boldsymbol{X}, \zeta_1)$ is correctly specified and fitted by least squares, yielding $\widehat{\zeta}_1$, that $\mathrm{var}[n^{1/2}\{\widehat{\beta}_1(\widehat{\zeta}_1) - \beta_1\}] = \mathrm{var}\left\{n^{-1/2}\sum_{i=1}^n U_i + n^{-3/2}\sum_{j=1}^p (\sum_{i=1}^n V_{ij})(\sum_{i=1}^n W_{ij})\right\} + o(n^{-1})$, where we emphasize dependence of $\widehat{\beta}_1$ on $\widehat{\zeta}_1$; $U_i = A_i Y_i/\pi - \{(A_i - \pi)/\pi\} q_1(\boldsymbol{X}_i, \zeta_1) - \beta_1$; V_{ij} is the jth element of $G^{-1} A_i q_{1\zeta}(\boldsymbol{X}, \zeta_1)\{Y_i - q_1(\boldsymbol{X}, \zeta_1)\}/\pi$, $q_{1\zeta}(\boldsymbol{X}, \zeta_1) = \partial/\partial\zeta_1\{q_1(\boldsymbol{X}, \zeta_1)\}$, and $G = E\{q_{1,\zeta}(\boldsymbol{X}, \zeta_1) q_{1\zeta}(\boldsymbol{X}, \zeta_1)'\}$; and $W_{ij} = \{(A_i - \pi)/\pi\} q_{1\zeta}(\boldsymbol{X}, \zeta_1)$. If we assume $\mathrm{var}(Y|Z=1) = \sigma^2$ and $\mathrm{var}(Y|\boldsymbol{X}, Z=1) = \sigma^2_{res}$, say, assumed constant over $\boldsymbol{X}$, then it is straightforward to show using combinatorial arguments that $\mathrm{var}[n^{1/2}\{\widehat{\beta}_1(\widehat{\zeta}_1) - \beta_1\}] = \mathrm{var}(U) + n^{-1}\mathrm{trace}\{E(VV')E(WW')\} + o(n^{-1})$. Under our assumptions, $E(VV') = \sigma^2_{res} G^{-1}/\pi$ and $E(WW') = (1-\pi)G/\pi$, so that the trace in this expression is equal to $\sigma^2_{res}(1-\pi)p/\pi^2$; also, $\mathrm{var}(U) = \sigma^2 + \sigma^2_{res}(1-\pi)/\pi$. Letting $\rho^2 = (\sigma^2 - \sigma^2_{res})/\sigma^2$, we thus obtain

$$\mathrm{var}[n^{1/2}\{\widehat{\beta}_1(\widehat{\zeta}_1) - \beta_1\}] = \sigma^2[1 + (1-\rho^2)\{(1-\pi)/\pi + (p/n)(1-\pi)/\pi^2\}] + o(n^{-1}).$$

This expression makes clear that there is a trade-off between gain of efficiency effected by increasing ρ^2 through using larger models (p) versus the penalty involved in increasing p. This result, although in an admittedly special case, may explain the simulations above. Generalization of these calculations is possible and will be investigated. These second-order results may be useful for correcting the first-order asymptotic variance for any fixed treatment-specific regression model and may be helpful in choosing among a set of competing models.

An alternative practical approach is to use bootstrap techniques to effect a correction. Although it has been argued Leeb and Pötscher (2005, 2006, 2008) that use of bootstrap methods after model selection may be problematic, we conjecture that use of a nonparametric bootstrap-t approach, which is based on a pivotal quantity and has been shown to be second-order correct in regular problems (Hall, 1986), may be fruitful: forming pivotal quantities based on the first-order asymptotic variance for $\widehat{\beta}$ may not yield exact pivotal statistics, but these may be close enough to pivotal to yield desirable results. Tsiatis et al. (2008) demonstrate how a "principled analysis" may be conducted using bootstrap methods. Yet another strategy will be to adapt the methods of Shen, Huang, and Ye (2004) for correcting inference on regression parameters in the selected models themselves to our setting. Because of the asymptotic equivalence of the semiparametric adjustment methods to ANCOVA methods in certain cases (Tsiatis et al., 2008), we expect that this will be straightforward.

During the project period, we will derive general second-order results and investigate all of these methods via extensive empirical study.

Methods for Handling Missing Covariates. All approaches for covariate adjustment, including the semiparametric methods we advocate, assume that all baseline auxiliary covariates potentially useful for enhancing efficiency are recorded on all n subjects in the trial. Missing such information is commonplace, where covariates are typically missing intermittently in different patterns across subjects with incomplete information. This reality leads many analysts and policy-makers to eschew adjusted analyses in favor of standard unadjusted ones that

do not use covariate information. Alternatively, an ad hoc and clearly problematic approach is to base the analysis only on the subjects with complete covariate data, which has the potential to lead to biased inferences if the missingness is not completely at random (MCAR). Even under MCAR, this approach is inefficient and could be grossly so, as subjects with even one missing covariate would be eliminated. Another approach is to limit the analysis to consideration of only the subset of covariates for which all subjects have complete data; which may exclude important predictors of outcome from which gains of efficiency might arise.

Missing covariates would appear at first glance to present a challenge for practical use of the semiparametric covariate adjustment methods, including the model selection efforts. We propose to develop and study several approaches to address this issue during the project period.

An advantageous feature of the semiparametric theory underlying the adjustment methods is that consideration of the assumed class of joint probability distributions for $(Y, \boldsymbol{X}, Z)$ under which it is developed it makes explicit that yet another tactic for handling missing covariates has a formal rationale. One may adopt the broader perspective that "$\boldsymbol{X}$" represents not necessarily a vector of specific covariates available at baseline but instead the more general "information" available at baseline, which includes the fact itself that some covariates are missing for some subjects according to an observed pattern. That is, we may view $\boldsymbol{X}$ as the collection of all such observed information, and redefine $\boldsymbol{X}$ accordingly. For example, we could take $\boldsymbol{X}$ to include indicators of whether or not particular variables are missing in addition to their values if they are not; that is, for the kth potentially observable baseline covariate X_k, $\boldsymbol{X}$ would include $\Delta_k = I(X_k \text{ is observed})$ and $\Delta_k X_k$. The semiparametric theory that leads to the covariate adjustment methods is unaffected by this redefinition of $\boldsymbol{X}$; randomization still ensures $Z \perp\!\!\!\perp \boldsymbol{X}$ under the redefinition. Consequently, the analyst is justified in defining $\boldsymbol{X}$ in this way and incorporating missingness patterns and indicators themselves in model-building. Indeed, that missingness of baseline information itself may be associated with prognosis supports this view. We will study the performance of this approach under a range of missingness mechanisms, patterns, and severity.

Under the foregoing rationale, making an assumption on, for example, on the nature the mechanism of the missingness, e.g., that it is MAR will have no implications for form of the adjustment methods, and hence offer no opportunity to take account of the missingness in a way that could enhance efficiency over the foregoing approach. In order to take the problem outside this class of estimators for β requires that the underlying probability model also include a specific assumption on the nature of the baseline covariates, e.g., that they are MAR and arise according to a specific model. Model selection to develop the treatment-specific regression models would then incorporate such assumptions. Approaches to model selection in a likelihood framework in this spirit have been proposed (e.g., Cavanaugh and Shumway, 1998; Hens and Molenberghs, 2006). We propose alternative approaches based on making "working" assumptions on the covariates (and possibly the outcome) and modern shrinkage methods (LASSO, adaptive LASSO) and Fast FSR that, while somewhat ad hoc, can exploit existing methods and software and have practical appeal.

For illustration, consider the admittedly restrictive working assumption that $\boldsymbol{X}$ is multivariate normal. A simple, accessible approach is to fit this parametric model via the EM algorithm and impute missing covariates via best linear unbiased predictors; the `mix` package in R is available for this purpose. Any model selection strategy may then be used with the "filled in" data set and the predicted values substituted in the semiparametric estimator for β; we will consider the LASSO, adaptive LASSO, and Fast FSR with forward selection. Taking adequate account of the variability due to imputation, along with possible effects of model selection, is then necessary for inference on β. One obvious approach is to use multiple imputation (Rubin, 1987), with standard errors for the estimator for β possibly corrected for model selection. Alternatively, model selection could be based on bagging; as discussed above, the effects of model selection may be less pronounced with this method.

In a more realistic setting with continuous and categorical covariates, the strategy of factorizing the joint distribution and modeling the components for the categorical subset and the continuous given the categorical could be used (Little and Schluchter, 1985; Schafer, 1997, Chapter 9), also implement in the R package `mix`.

A second approach involves extending the working assumption to the joint distribution of Y and $\boldsymbol{X}$. Of course, the semiparametric model underlying the adjustment methods does not require an assumption on the distribution of $Y, \boldsymbol{X}$ given Z, indeed, that is part of the appeal; thus, such a "working" assumption would be made only to facilitate the missing covariate problem, and sensitivity to it will need to be evaluated. To illustrate, again consider the restrictive case where the working joint distribution assumption is multivariate normal. Fitting this parametric model via the EM algorithm as above would yield estimated treatment-specific regression relationship for Y given all of $\boldsymbol{X}$. To carry out model selection to get closer to the "true" relationship, armed with the

estimated coefficients, one could invoke the least squares approximation (LSA) to the adaptive LASSO (Wang and Leng, 2007). This methods facilitates model selection for virtually any model by working with estimated coefficients of a "full" model rather than the model itself; in our context, the coefficients of "unimportant" covariates would be "zeroed out," yielding the final model to be substituted in the semiparametric adjusted treatment effect estimator. This idea could in principle be extended to working models using the approaches above.

The foregoing approaches, while practically appealing, have obvious limitations. E.g., the LSA adaptive LASSO approach limits the regression relationships that can be considered. An alternative approach would be to consider as a starting point a semiparametric model that incorporates missingness mechanisms and assumptions on the covariates; e.g., that they arise from a specific distribution. Semiparametric theory could then be used to characterize the class of all estimators under this model, similar to Zhang et al. (2008).

During the project period, we will study all the performance of of the all of these approaches via empirical studies in scenarios in sub-aim 1 with the additional feature of covariate missingness, where we will base the missingness patterns and covariate assumptions to be considered on a review of data from Core B. The extent to which "going to the trouble" to adopt additional assumptions on the covariates will be evaluated; our conjecture is that, unless very strong parametric assumptions are justifiably made, possible efficiency gains over the current semiparametric approach with the redefinition of $\boldsymbol{X}$ discussed above will be negligible. Robustness of the methods that rely on assumptions on $\boldsymbol{X}$ to departures from these will be systematically explored.

In all of these sub-aims, we will apply the methods to re-analyze cancer clinical trials compiled by Core B to assess the extent to which the proposed methods would have strengthened or altered primary inferences.

Adjustment Methods for Partially Randomized Trials. Consider a two-arm partially randomized trial with univariate outcome Y, where a subject recruited to the study is asked whether or not s/he is willing to be randomized. Denote those unwilling by $P = 1$; such a subject is assigned to his/her preferred treatment. Willing subjects have $P = 0$ and are randomized. The observed data are thus iid $(Y_i, \boldsymbol{X}_i, Z_i, P_i)$, $i = 1, \ldots, n$, where $\boldsymbol{X}$ is as above, and Z_i is observed treatment assignment, which, for ease of exposition we now denote by $Z = 0, 1$, made by preference or randomized as $P_i = 1$ or $P_i = 0$, with randomization probability $P(Z = 1|P = 0) = \pi$. To make explicit our assumptions, we exploit ideas from causal inference and define $Y_i^{(0)}$ and $Y_i^{(1)}$ as the potential outcomes (e.g., Lunceford and Davidian, 2004) for subject i; i.e., the outcomes s/he would exhibit were s/he to receive each treatment 0,1. The goal is to estimate $E(Y^{(1)}) - E(Y^{(0)})$, that is, the difference in population means were the entire population of patients to receive each treatment. We make the usual assumption (e.g., Rubin, 1974; Robins, Hernán, and Brumback, 2000) $Y_i = Y_i^{(1)} Z_i + Y_i^{(0)}(1 - Z_i)$ that the outcome actually observed for subject i is his/her potential outcome under the treatment that s/he actually received. This implies that outcome is the same regardless of how treatment was assigned (preference or randomization), which should be reasonable for objective treatments such as chemotherapeutic agents and vaccines, as opposed to behavioral interventions such as those studied by Long, Little, and Lin (2008). We also assume that $(\boldsymbol{X}_i, Y_i^{(0)}, Y_i^{(1)}) \perp\!\!\!\perp Z_i | P_i = 0$; this is of course true for all subjects in a fully randomized trial, for which it is well-known that $E(Y|Z = 1) - E(Y|Z = 0) = E(Y^{(1)}) - E(Y^{(0)})$, so that inference has the desired causal interpretation. Note that our focus on this overall population quantity is different from that of Long et al. (2008), who consider treatment effect with the preference group. For the preference group, assume $(Y_i^{(0)}, Y_i^{(1)}) \perp\!\!\!\perp Z_i | \boldsymbol{X}_i, P_i = 1$, the strong ignorability or no unmeasured confounders assumption, which, similar to MAR, implies that $\boldsymbol{X}$ captures all information necessary to explain treatment preference among these subjects. Such an assumption is necessary to make progress and is implicit in the competing empirical likelihood methods presented in Project 2, Aim 3. Whether or not it is plausible must be critically assessed, and we will investigate sensitivity to its violation during the project period.

Under this set-up, we assume $P(Z = 1|\boldsymbol{X}, P) = e(\boldsymbol{X}, P, \pi, \boldsymbol{\xi}) = \pi^{(1-P)}\{\varphi(\boldsymbol{X}_i, \boldsymbol{\xi})\}^P$, where $\varphi(\boldsymbol{X}, P, \boldsymbol{\xi})$ is a propensity score (e.g., Lunceford and Davidian, 2004) model for the probability of receiving treatment 1 among those with a preference, which may be fitted to the data from $\{i : P_i = 1\}$ by maximum likelihood techniques for binary regression, yielding $\widehat{\boldsymbol{\xi}}$, and π may be estimated by the sample proportion $\widehat{\pi}$ among $\{i : P_i = 0\}$ randomized to treatment 1. The proposed estimator for the treatment mean difference has the form

$$n^{-1}\sum_{i=1}^{n}\left[\frac{Z_i Y_i}{e(\boldsymbol{X}_i, P_i, \widehat{\pi}, \widehat{\boldsymbol{\xi}})} - \frac{(1 - Z_i)Y_i}{1 - e(\boldsymbol{X}_i, P_i, \widehat{\pi}, \widehat{\boldsymbol{\xi}})} - \{Z_i - e(\boldsymbol{X}_i, P_i, \widehat{\pi}, \widehat{\boldsymbol{\xi}})\}\left\{\frac{E(Y_i|\boldsymbol{X}_i, P_i, Z_i = 1)}{e(\boldsymbol{X}_i, P_i, \widehat{\pi}, \widehat{\boldsymbol{\xi}})} + \frac{E(Y_i|\boldsymbol{X}_i, P_i, Z_i = 0)}{1 - e(\boldsymbol{X}_i, P_i, \widehat{\pi}, \widehat{\boldsymbol{\xi}})}\right\}\right];$$

in practice, $E(Y|\boldsymbol{X}, P, Z = g)$, $g = 0, 1$ would be replaced by postulated regression models specific to each of the four combinations of preference and treatment. Among randomized subjects ($P = 0$), this estimator has the form of (5.1); for preference subjects ($P = 1$), it has the form of the most efficient estimator for the treatment difference under the usual causal inference assumptions (Lunceford and Davidian, 2004, equation (9)). Thus, the estimator combines adjustment both to enhance efficiency and account for the selection bias in the preference group, as required. It follows from standard theory that the estimator should be consistent and asymptotically normal and that sandwich standard errors may be derived (Stefanski and Boos, 2002).

The empirical likelihood approach in Project 1, Aim 3, has the advantage of simplicity of implementation, as it involves no modeling; however, this is at the expense of potential bias. Both estimators require a no unmeasured confounders assumption, but that proposed here takes complete account of the selection bias through models for the full probability distribution of treatment preference, while the empirical likelihood methods attempt to achieve this approximately by constraining the moments of the confounding covariates in these models to be equal in each treatment for the preference group. This may eliminate some of the selection bias, but cannot resolve it all. During the project period, we will compare these approaches in extensive simulation studies to develop guidelines outlining the sample sizes and situations where the more complex estimator here is required.

During the project period, we will extend this approach to the general setting of $k \geq 2$ treatments and more complex estimands and outcomes (e.g., censored survival), and investigate the role of model selection for the required models. We will also carry out a comprehensive evaluation of performance under sample sizes and conditions consistent with these trials in cancer research, and investigate consequences of model misspecification and departures from the no unmeasured confounders assumption.

5.2 Aim 2: Methods for Primary and Longitudinal Analyses in the Presence of Drop-out

Covariate Adjustment Methods Under MAR Drop-Out. Consider a fully randomized trial as in Aim 1. For simplicity, restrict attention to $k = 2$ treatments, and let Y denote outcome and $Z = 0, 1$ denote assignment to the two treatments, where $P(Z = 1) = \pi$. As before, we are interested in inference on a parameter β involved in treatment comparisons in a model describing aspects of the conditional distribution of Y given Z depending on θ. We now allow the possibility that some subjects drop out prior to the time T at which Y is to be recorded, which we refer to as the lag time, so that the outcome is time-lagged or a marked point process (Huang and Louis, 19998; Anstrom and Tsiatis, 2001). If the outcome of interest is survival time, then $Y = T$, but we allow the possibility that T might be time to ascertainment of some other outcome Y, which could be the same for all subjects or vary. The time lag T itself may be unobserved if a subject drops out prior to the recording of the response. We denote by C the potential drop-out time. Under these considerations, the observed data from the trial are iid $O_i = \{U_i, \Delta_i, \Delta_i Y_i, \boldsymbol{X}_i(U_i), Z_i\}$, $i = 1, \ldots, n$, where $U_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$. Thus, Y_i is observed only if $\Delta_i = 1$. $\boldsymbol{X}_i(U_i)$ contains all auxiliary covariate information observed from baseline until drop-out or ascertainment of outcome, which we represent as $\boldsymbol{X}_i(U_i) = \{\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}(U_i)\}$, where $\boldsymbol{X}_{1i}$ is a vector of baseline auxiliary covariates for which randomization guarantees $Z \perp\!\!\!\perp \boldsymbol{X}_i$, and $\boldsymbol{X}_{2i}(U_i) = \{\boldsymbol{X}_{2i}(u), 0 \leq u \leq U_i\}$ is the observed post-randomization covariate history observed on i up to time U_i.

If there were no drop-out, then the "full data" that could be observed are $F_i = \{T_i, Y_i, \boldsymbol{X}_i(T_i), Z_i\}$, $i = 1, \ldots, n$. Let $\lambda_C(u|T \geq u, F)$ be the conditional hazard of dropping out given the full data, which includes information subsequent to drop-out. We operationalize the MAR assumption as, in obvious notation, $\lambda_C(u|T \geq u, F) = \lambda_C\{u|T \geq u, \boldsymbol{X}(u), Z\}$; i.e., the probability of dropping out at time u, given that one has not achieved the lag time or yet dropped out, depends only on information that is observed up to u and not on future information.

Let $m(Y, Z; \theta)$ be an estimating function for the unadjusted analysis if one had full data, and, assuming $\lambda_C\{u|T \geq u, \boldsymbol{X}(u), Z\}$ is known for the moment, let $K_C\{t, \boldsymbol{X}(t), Z\} = \exp[-\int_0^t \lambda_C\{u|T \geq u, \boldsymbol{X}(u), Z\}\, du]$ be the conditional survival function for drop-out. Let $dM_C\{u, \boldsymbol{X}(u), Z\}$ be the increment of the martingale process $dN_C(u) - \lambda_C\{u|T \geq u, \boldsymbol{X}(u), Z\}A(u)du$, where $dN_C(u)$ is the increment of the counting process for drop-out, i.e., $N_C(u) = I(U \leq u, \Delta = 0)$, and $A(u) = I(U \geq u)$ is the "at risk" process. Then, from the theory of Robins and Rotnitzky (1992) and Tsiatis (2006, sec. 9.2), it estimating functions for β based on O should have the form

$$\frac{\Delta m(Y, Z; \boldsymbol{\theta})}{K_C\{U, \boldsymbol{X}(U), Z\}} + (Z - \pi)a(\boldsymbol{X}_1) + \int dM_C\{u, \boldsymbol{X}(u), Z\} b\{u, \boldsymbol{X}(u), Z\}, \tag{5.2}$$

where $a(\boldsymbol{X}_1)$ and $b\{u, \boldsymbol{X}(u), Z\}$ are arbitrary functions. The drop-out survival function $K_C\{U, \boldsymbol{X}(U), Z\}$ will not be known in practice and may be modeled, e.g., a stratified (by treatment) proportional hazards model,

and the fitted model substituted in (5.2). If this model is correct, the estimator for β be consistent; moreover, following the theory, the optimal $a(\boldsymbol{X}_1)$ and $b\{u, \boldsymbol{X}(u), Z\}$ leading to the most efficient estimator are $a(\boldsymbol{X}_1) = E\{(Z-\pi)m(Y,Z;\boldsymbol{\theta})|\boldsymbol{X}_1\}/\{\pi/(1-\pi)\}$ and $b\{u, \boldsymbol{X}(u), Z\} = E\{m(Y,Z;\boldsymbol{\theta})|T \geq u, \boldsymbol{X}(u), Z\}/K_C\{u, \boldsymbol{X}(u), Z\}$. In practice, as before, one would postulate (parametric) models for these conditional expectations. In the special case of the difference of two treatment means β_2 in Aim 1, one may use (5.2) to estimate the two treatment means separately and form the difference; in this case, letting $\mu_g = E(Y|Z=g)$, $g=0,1$, the optimal estimator for μ_g should take the form

$$n^{-1}\sum_{i=1}^{n}\left[\frac{\Delta_i Z_i^g(1-Z_i)^{1-g}Y_i}{K_C\{U_i, \boldsymbol{X}(Y_i), Z_i\}} + (-1)^g(Z_i-\widehat{\pi})E(Y|\boldsymbol{X}_{1i}, Z_i) + Z_i^g(1-Z_i)^{1-g}\int \frac{dM_C\{u, \boldsymbol{X}(u), Z_i\}E\{Y|T_i \geq u, \boldsymbol{X}(u), Z\}}{K_C\{u, \boldsymbol{X}(u), Z\}}\right].$$

During the project period, we will use semiparametric theory to validate rigorously the above results, and we will use the results in Tsiatis (2006, sec. 9.1) to derive standard error formulæ for estimators for β found using (5.2). We will derive the form of estimators in the particular case where the outcome is a possibly censored time-to-event. Consistency of treatment effect estimator requires the model K_C to be correct; to relax this, we will derive a doubly robust version of the methods via a continuous time version of the methods in the next sub-aim. We will carry out extensive simulations to evaluate performance of the methods under sample sizes and scenarios reflecting those in cancer clinical trials from Core B and elsewhere, and apply the methods to studies involving drop-out. Studies of sensitivity to modeling assumptions will be conducted.

Doubly Robust Methods for Longitudinal Analysis Under MAR Dropout. The development is related to that in the previous sub-aim; here, we consider specifically the case of a longitudinal study in a cancer trial where a measure such as QOL score is to be recorded at pre-specified times $t_0, t_1, \ldots, t_{M+1}$, yielding observed outcomes $Y_0, Y_1, \ldots, Y_{M+1}$; we include the possibility of measurement at baseline (time $t_0 = 0$) as in E1694, but this is not necessary. If a subject drops out between times t_j and t_{j+1}, only $(Y_0, \ldots, Y_j)$ would be observed, and $(Y_{j+1}, \ldots, Y_{M+1})$ would be missing. Let $\boldsymbol{L}_0$ be baseline covariates, and suppose that $\boldsymbol{L}_j$, $j=0,\ldots,M+1$, is the vector of information collected at time t_j, which includes Y_j but also may include auxiliary variables. Let R be the index of the last time at which information is available prior to drop-out, so $R=j$ means drop-out occurred between t_j and t_{j+1} and $R=M+1$ that drop-out did not occur. Write $\overline{\boldsymbol{L}}_j = (\boldsymbol{L}_0, \ldots, \boldsymbol{L}_j)$ to denote all information collected through t_j, and write $\overline{\boldsymbol{L}} = \overline{\boldsymbol{L}}_{M+1}$. The observed data can be summarized by iid $(R_i, \overline{\boldsymbol{L}}_{R_i})$, $i=1,\ldots,n$. The MAR assumption may be formalized as $P(R=j|\overline{\boldsymbol{L}}) = \pi_j(\overline{\boldsymbol{L}}_j)$, $j=0,\ldots,M$; i.e., the probability of dropping out between t_j and t_{j+1} conditional on all intended information depends only on the data available through t_j and not on data not observed under drop-out. $P(R=M+1|\overline{\boldsymbol{L}}) = \pi_{M+1}(\overline{\boldsymbol{L}})$, the probability of not dropping out, in fact depends only on $\overline{\boldsymbol{L}}_M$ because $\pi_{M+1}(\overline{\boldsymbol{L}}) = 1 - \sum_{j=0}^{M}\pi_j(\overline{\boldsymbol{L}}_j)$, and we write $\pi_{M+1}(\overline{\boldsymbol{L}}_M)$.

Suppose we are interested in a parameter $\boldsymbol{\beta}$ in some statistical model for the full, intended data, and suppose that $m(\overline{\boldsymbol{L}}, \boldsymbol{\beta})$ is an unbiased estimating function such that one would estimate $\boldsymbol{\beta}$ based on full data by solving $\sum_{i=1}^{n} m(\overline{\boldsymbol{L}}_i, \boldsymbol{\beta}) = \boldsymbol{0}$. For instance, if interest focused on the mean change in QOL from baseline to t_{M+1}, $\beta = E(Y_{M+1} - Y_0)$, which would be estimated by solving $\sum_{i=1}^{n}(Y_{M+1,i} - Y_{0,i} - \beta) = 0$. Robins et al. (2005) showed that, when the drop-out probabilities are known and MAR holds, all consistent and asymptotically normal estimators for $\boldsymbol{\beta}$ are solutions to estimating equations of the form

$$\sum_{i=1}^{n}\left[\frac{I(R_i = M+1)m(\overline{\boldsymbol{L}}_i, \boldsymbol{\beta})}{\pi_{M+1}(\overline{\boldsymbol{L}}_i)} + \sum_{j=0}^{M}\frac{\{I(R_i=j) - \lambda_j(\overline{\boldsymbol{L}}_{j,i})I(R_i \geq j)\}}{K_j(\overline{\boldsymbol{L}}_{j,i})}f_j(\overline{\boldsymbol{L}}_{j,i})\right] = 0, \qquad (5.3)$$

where, for $j=0,\ldots,M$, $\lambda_j(\overline{\boldsymbol{L}}_j) = P(R=j|R \geq j, \overline{\boldsymbol{L}}_j)$, $K_j(\overline{\boldsymbol{L}}_j) = P(R>j|\overline{\boldsymbol{L}}_j) = \prod_{\ell=1}^{j}\{1-\lambda_j(\overline{\boldsymbol{L}}_\ell)\}$, $\pi_{M+1}(\overline{\boldsymbol{L}}) = \pi_{M+1}(\overline{\boldsymbol{L}}_M) = K_M(\overline{\boldsymbol{L}}_M)$; and $f_j(\overline{\boldsymbol{L}}_j)$ are arbitrary functions of $\overline{\boldsymbol{L}}_j$.

The optimal choices $f_j(\overline{\boldsymbol{L}}_j)$, $j=0,\ldots,M$, those leading to the estimator of $\boldsymbol{\beta}$ solving (5.3) with smallest asymptotic variance, are $f_j(\overline{\boldsymbol{L}}_j) = E\{m(\overline{\boldsymbol{L}}, \boldsymbol{\beta}_0)|\overline{\boldsymbol{L}}_j\}$, where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. In practice, we know neither the drop-out mechanism, summarized by the discrete hazards $\lambda_j(\overline{\boldsymbol{L}}_j)$, $j=0,\ldots,M$, nor these conditional expectations, and the usual strategy is to postulate parametric models. For the discrete hazards, we write $\lambda_j(\overline{\boldsymbol{L}}_j, \boldsymbol{\psi})$, $j=0,\ldots,M$, depending on a parameter $\boldsymbol{\psi}$; typically, these would be binary (e.g., logistic) regression models, and $\boldsymbol{\psi}$ would be estimated by maximum likelihood, so maximizing $\prod_{j=0}^{M}\prod_{i:R_i \geq j}\{\lambda_j(\overline{\boldsymbol{L}}_{j,i}, \boldsymbol{\psi})\}^{I(R_i=j)}\{1-\lambda_j(\overline{\boldsymbol{L}}_{j,i}, \boldsymbol{\psi})\}^{I(R_i>j)}$ to obtain $\widehat{\boldsymbol{\psi}}$. Similarly, in a way analogous to the approach in Zhang et al. (2008), models for the conditional expectations of the form $f_j(\overline{\boldsymbol{L}}_j, \boldsymbol{\xi})$, $j=0,\ldots,M$, depending on a parameter $\boldsymbol{\xi}$, could be

developed and ξ estimated by some estimator $\widehat{\xi}$; we discuss estimation of ξ in detail shortly. Substituting in (5.3), one would thus solve

$$\sum_{i=1}^{n}\left[\frac{I(R_i=M+1)m(\overline{L}_i,\beta)}{K_M(\overline{L}_i,\widehat{\psi})}+\sum_{j=0}^{M}\frac{\{I(R_i=j)-\lambda_j(\overline{L}_{j,i},\widehat{\psi})I(R_i\geq j)\}}{K_j(\overline{L}_{j,i},\widehat{\psi})}f_j(\overline{L}_{j,i},\widehat{\xi})\right]=\mathbf{0}. \tag{5.4}$$

It may be shown that the estimator for β obtained by solving (5.4) is doubly robust. That is, it will be consistent and asymptotically normal if *either* the model for drop-out is correctly specified, $P(R=j|R\geq j,\overline{L}_j)-\lambda_j(\overline{L}_j,\psi)$ for all $j=0,\ldots,M$; *or* if the model for the conditional expectations is correctly specified, i.e., $E\{m(\overline{L},\beta_0)|\overline{L}_j\}=f_j(\overline{L}_j,\xi)$ for all $j=0,\ldots,M$ and some ξ in the parameter space. As noted in Section **3.2**, the doubly robustness property is desirable because only one set of models need be correctly specified to ensure consistency of the resulting estimator for β. In the present context, this implies that the analyst need only do a good job at modeling the drop-out mechanism or the conditional expectations to obtain reliable inferences; in a complex setting such as this, this feature is especially appealing.

In (5.4), ψ is estimated by the maximum likelihood estimator (MLE). However, how best to estimate ξ is not clear. For the double robustness property to hold, $\widehat{\xi}$ must converge to the value ξ_0 such that $E\{m(\overline{L},\beta_0)|\overline{L}_j\}=f_j(\overline{L}_j,\xi_0)$, $j=0,\ldots,M$, regardless of whether or not the drop-out probabilities are correctly specified. Such estimators may be obtained as solutions to estimating equations based on the estimating function

$$\sum_{j=0}^{M}I(R>j)q_j(\overline{L}_j)\{f_{j+1}(\overline{L}_{j+1},\xi)-f_j(\overline{L}_j,\xi)\}, \tag{5.5}$$

where $q_j(\overline{L}_j)$ is a vector of functions of $\overline{L}_j$ for each j of dimension the same as ξ, and $f_{M+1}(\overline{L},\xi)=m(\overline{L},\beta)$. Bang and Robins (2000) suggested such an estimator for ξ.

As recounted in Sections **3.2** and **4.2**, usual doubly robust estimators can perform poorly, as shown in simulations by Kang and Schafer (2007) in the case of estimation of the mean of a single outcome, where the outcome may be MAR, which can be deduced as a special case of the drop-out set-up here by taking $M=0$, so that $\overline{L}=L_0$; $m(\overline{L},\beta)=Y_0-\beta$, with β the unconditional mean of Y_0; and $f_0(L_0)=E(Y_0|L_0)-\beta$, so that ξ is the parameter in a regression model for $E(Y_0|L_0)$. As noted in Section **4.2**, Kang and Schafer considered the doubly robust estimator with ξ estimated by least squares based on the complete cases, corresponding to taking $q_0(L_0)$ in (5.5) equal to $\partial/\partial\xi'\{f_0(\overline{L}_0,\xi)\}$. Their results translate in the current context to suggesting that, if the drop-out model and models for the conditional expectations $E\{m(\overline{L},\beta_0)|\overline{L}_j\}$ are misspecified even slightly, then the resulting estimator for β may be unstable, especially so when the probabilities $K_M(\overline{L}_i,\widehat{\psi})$ are very small for some i. In Cao et al. (2009), in the case of a single mean ($M=0$ as above), we have shown that the performance of doubly robust estimators for the mean β is highly dependent on how the parameters ξ in a model for $E(Y_0|L_0)$ are estimated. We demonstrate that, if we consider the class of doubly robust estimators indexed by ξ, then there exists an optimal estimator for β even if the regression model $f_0(L_0,\xi)$ is misspecified, and we derive an estimator for ξ that leads to this optimal estimator by deducing an appropriate, alternative choice of $q_0(L_0)$ in (5.5). The resulting estimator for β is doubly robust and optimal in that it has smallest asymptotic variance among all doubly robust estimators even if the regression model $f_0(L_0,\xi)$ is misspecified. Importantly, it does not exhibit the unstable performance of standard doubly robust estimators using least squares based on the complete cases to estimate ξ as above, and analytical arguments are presented supporting why this so.

In the considerably more complicated setting here of a longitudinal study with drop-out, we believe that it should be possible to derive similar results, thus providing analysts with for analysis of longitudinal studies that both have the appealing double robustness property and offer stable performance and reliable performance. We conjecture that the optimal estimator (in the sense described above for the simple case of $M=0$) for ξ will be found by taking

$$q_j(\overline{L}_j)=-\frac{\sum_{\ell=1}^{j}\lambda_j(\overline{L}_j)\partial/\partial\xi'\{f_j(\overline{L}_j,\xi)\}/K_\ell(\overline{L}_\ell)}{K_\ell(\overline{L}_j)},\quad j=1,\ldots,M. \tag{5.6}$$

(5.6) assumes that the drop-out model defined by the discrete hazards is known. Similar to Cao et al. (2009), we will find the optimal $q_j(\overline{L}_j)$ when these hazards are modeled in terms of ψ and ψ is estimated by the MLE.

During the project period, we will verify that (5.6) leads to the optimal doubly robust estimator for β, and we will derive its large sample properties. We will develop feasible computational algorithms to combine estimation

of ψ in the drop-out model and solution of (5.5) with the optimal choice (5.6) for estimation of ξ with solution of (5.4) to obtain the optimal doubly robust estimator for β. We will carry out extensive simulations based on longitudinal studies in cancer clinical trials such as E1694, E1684, CALGB 90401, CALGB 49907, and those compiled by Core B to evaluate the performance of the methods. This will also involve study of the impact of different severity and types of misspecification of both sets of models, and especially the case of "slight" misspecification of both, in order to confirm that the choice (5.6) achieves the desired performance. Based on these results, similar to Cao et al. (2009), we will derive an analytical argument to explain this behavior.

5.3 Aim 3: Diagnostic Measures for Longitudinal and Joint Models in the Presence of Missing Data

We outline three sub-aims. First, we will develop new sensitivity methods for assessing minor perturbations to general parametric and semiparametric models. As an initial step toward methods for joint models, we will use these results to develop diagnostic tools for random effects models for longitudinal data with possibly missing outcomes and/or covariates; we will then extend these to joint models for longitudinal and survival data under these conditions. For each sub-aim, we will first carry out methodological/theoretical development and iterate this with empirical studies based on simulated data with known ground truth to assess performance and validate relevance of the theory in practice. Application of the methods to data from the studies discussed in Section **3.3** will then be undertaken to demonstrate the utility of the methods. The tools developed here will formally help clinicians to characterize the relationship between the longitudinal measure (e.g., QOL or immune response) and survival within each treatment arm as in ECOG 1694.

General Methodology for Sensitivity Analysis. Let D_{obs} and D_{mis} be the observed data and the missing data, and $D_{com} = (D_{mis}, D_{obs})$ be the complete data. We develop a new geometric framework, the perturbation manifold, to measure each perturbation ω in a perturbation set Ω to statistical models with incomplete data, denoted by $p(D_{com};\theta)$. The perturbation model $\mathcal{M} = \{p(D_{com};\theta,\omega) : \omega \in \Omega\}$ has a natural geometrical structure. Because Ω can be an infinite dimensional set, we must develop a manifold for the infinite dimensional space, which includes the finite dimensional manifold as a submanifold (Friedrich, 1991; Lang, 1999; Zhu et al., 2007). Assume that $C(t) : p_c\{\omega(t)\} = p\{D_{com};\theta,\omega(t)\}$ is a differentiable function mapping from $t \in \mathcal{I} \subset \mathbb{R}$ to the manifold $\mathcal{M}$ with $p_c\{\omega(0)\} = p(D_{com};\theta,\omega)$, where $\mathcal{I}$ is an open interval covering 0. At each ω, there is a tangent space $T_\omega\mathcal{M}$ of $\mathcal{M}$ defined by $T_\omega\mathcal{M} = [v(\omega) = d\log p_c\{\omega(t)\}/dt|_{t=0} : E\{v(\omega)\} = 0$ and $v(\omega) \in L^2(P) < \infty]$, where $L^2(P) = \{g : \int g^2 dP < \infty\}$ is a Hilbert space. The inner product of any two tangent vectors $v_1(\omega)$ and $v_2(\omega)$ in $T_\omega\mathcal{M}$ is defined as $g(v_1,v_2)(\omega) = E\{v_1(\omega)v_2(\omega)\}$. We can calculate the Levi-Civita connection, denoted by $\nabla_v u$, and the geodesic on $\mathcal{M}$, and then introduce the notion of exponential and logarithm maps to generalize the concept of moving "straight" in the direction of a tangent vector. In particular, an *appropriate perturbation* to any model corresponds to choosing the orthogonormal basis of $T_\omega\mathcal{M}$.

We develop several global influence measures for quantifying the effects of perturbing $p(D_{com};\theta)$. Without loss of generality, let $p_c(\omega^0)$ and $p_c(\omega)$ represent the unperturbed and perturbed complete-data distributions, respectively. Let $C(t) = p_c\{\omega(t)\} : [-\gamma,\gamma] \to \mathcal{M}$ be a smooth curve on $\mathcal{M}$ joining $p_c(\omega^0)$ and $p_c(\omega)$ such that $C(0) = p_c(\omega^0)$ and $C(1) = p_c(\omega)$, where $\gamma > 1$. Consider a smooth function of interest $f(\omega) = f\{p_c(\omega)\} : \mathcal{M} \to \mathbb{R}$ for sensitivity analysis, which is often chosen to be a functional of the unperturbed and perturbed distributions (Cook, 1986; Zhu et al., 2007). We introduce a global influence measure along the smooth curve $C(t)$ as $\mathrm{GI}_{f,C(t)}(\omega^0,\omega) = \{f(\omega) - f(\omega^0)\}^2/\{S_C(\omega^0,\omega)\}^2$, where $S_C(\omega^0,\omega)$ is the length of the curve $p_c\{\omega(s)\}$ as $s \in [0,1]$. Considering the geodesic joining $p_c(\omega^0)$ and $p_c(\omega)$, we can define an intrinsic global influence measure, denoted by $\mathrm{IGI}_f(\omega^0,\omega)$. Moreover, we can extend these global influence measures to quantifying a set of perturbations $\omega \in \Omega_1 \subset \Omega$ to $p(D_{com};\theta)$ by using $\sup_{\omega\in\Omega_1} \mathrm{IGI}_f(\omega^0,\omega)$.

We develop first- and second-order local influence measures for quantifying effects of perturbing $p(D_{com};\theta)$ by considering the local behavior of $f\{\omega(t)\}$ as $t \to 0$ along all possible smooth curves. Let $\dot{f}\{\omega(0)\}$ and $\ddot{f}\{\omega(0)\}$ denote the first- and second-order derivatives of $f\{\omega(t)\}$ with respect to t evaluated at $t = 0$. We distinguish two cases: $\dot{f}\{\omega(0)\} \neq 0$ for some smooth curves $\omega(t)$ and $\dot{f}\{\omega(0)\} = 0$ for all smooth curves $\omega(t)$. If $\dot{f}\{\omega(0)\} \neq 0$ holds for some $\omega(t)$, introduce the first-order local influence measure as $\mathrm{FI}_f[v]\{\omega(0)\} = \lim_{t\to 0} \mathrm{GI}_{f,C(t)}\{\omega(0),\omega(t)\} = [df[v]\{\omega(0)\}]^2/g(v,v)\{\omega(0)\}$, where $df[v]$ denotes the directional derivative of $f(\omega)$ in the direction of $v \in T_\omega\mathcal{M}$. $\mathrm{FI}_f[v]\{\omega(0)\}$ is independent of specification of $C(t)$. We use $\ddot{f}\{\omega(0)\}$ to assess second-order local influence of ω. However, for a general smooth curve $\omega(t)$ on $\mathcal{M}$, $\ddot{f}\{\omega(0)\}$ is not geo-

metrically well-behaved (Lang, 1999; Zhu et al., 2007), thus we only consider the geodesic and define a second-order influence measure in the direction $v \in T_{\omega(0)}\mathcal{M}$ as $\mathrm{SI}_f[v]\{\omega(0)\} = \mathrm{Hess}(f)(v,v)\{\omega(0)\}/g(v,v)\{\omega(0)\}$, where $\mathrm{Hess}(f)(v,v)\{\omega(0)\}$ is a covariant (or Riemmanian) Hessian (Lang, 1999).

We will apply these measures to sensitivity analysis in longitudinal and joint models with/without missing data as discussed below, examining various perturbation schemes to choose the appropriate ones.

Methodology for Diagnostic Methods for Longitudinal Data. Consider a general random effects model for longitudinal outcome vectors $\boldsymbol{y}_i$ involving completely and partially observed baseline covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, random effects $\boldsymbol{b}_i$, and vector of missingness indicators $\boldsymbol{r}_i$ and define the associated n independent clusters $D_{com} = \{\boldsymbol{d}_i = (\boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{r}_i, \boldsymbol{y}_i),\ i = 1,\ldots,n\}$. Moreover, let $\boldsymbol{z}_{m,i}$ and $\boldsymbol{z}_{o,i}$ denote the missing and observed components of $\boldsymbol{z}_i$, respectively. Assume that the density for the ith cluster can be written $p(\boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{r}_i, \boldsymbol{y}_i|\eta) = p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{z}_i, \eta)p(\boldsymbol{b}_i|\eta)p(\boldsymbol{x}_i, \boldsymbol{z}_i|\eta)\ p(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \eta)$, where η denotes the vector of unknown parameters. As an illustration, consider the simpler case where $D_{com} = \{\boldsymbol{d}_i = (\boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{y}_i),\ i = 1,\ldots,n\}$, and assume that $p(\boldsymbol{b}_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \eta) = p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i, \eta)p(\boldsymbol{b}_i|\eta)$. Consider two classes of perturbation schemes ω to the longitudinal models: The single-case perturbation scheme is defined by $\prod_{i=1}^n p(\boldsymbol{b}_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \eta, \omega) = \prod_{i=1}^n p(\boldsymbol{b}_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \eta, \omega_i)$, where ω_i denotes the perturbation to the ith observation; in the global perturbation scheme, mainly for assessing the robustness of model assumptions to small perturbations (Copas and Eguchi, 2005; Troxel et al., 2004; Huang et al., 2009), the perturbed complete-data density is defined by $\prod_{i=1}^n p(\boldsymbol{b}_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \eta, \omega) = \prod_{i=1}^n p(\boldsymbol{b}_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \eta, \omega, \epsilon_i)$, where ϵ_i may be independently simulated observations (Huang et al., 2009).

To apply the general sensitivity analysis, we will first calculate the metric tensor and other geometric quantities such as curvatures for a given perturbation scheme (Zhu et al., 2007). Based on the metric tensor $g(v_1, v_2)$, we can choose an appropriate perturbation scheme at least locally. Second, we choose an objective function, such as the MLE $\widehat{\eta}(\omega)$ under the perturbed model and the likelihood displacement function (Cook, 1986). For most finite-dimensional perturbation schemes, it is straightforward to calculate the local influence measures $\mathrm{FI}_f[\mathrm{v}]\{\omega(0)\}$ and $\mathrm{SI}_f[v]\{\omega(0)\}$, which are simply vectors and matrices. For the global measure $\mathrm{GI}_{f,C(t)}\{\omega(0), \omega\}$, we need to calculate $f(\omega)$, $f\{\omega(0)\}$, and $S_C\{\omega(0), \omega\}$ using numerical methods. E.g., if we choose the MLE as $f(\omega)$, we need to compute the MLEs for both $\omega(0)$ and ω.

Case-deletion Measures. To detect outliers and influential points, we define case-deletion measures, including Cook's distance and the Q-displacement, to examine the effects of deleting individual observations/clusters on the MLE of η, $\widehat{\eta}$. To quantify the effects of deleting the ith cluster on $\widehat{\eta}$, define an estimator of η for a subsample $D_{com[i]}$, in which $\boldsymbol{d}_i$ is deleted from D_{com}. Then define $\widehat{\eta}_{[i]}$ as the maximizer of $Q_{[i]}(\eta|\widehat{\eta}) = E\{\log p(D_{com[i]}, \eta)|D_{obs}, \widehat{\eta}\}$, where the expectation is taken with respect to $p(D_{mis}|D_{obs}, \widehat{\eta})$. Recall that $\widehat{\eta} = \mathrm{argmax}_\eta Q(\eta|\widehat{\eta})$. Thus, Cook's distance is $\mathrm{CD}_i = (\widehat{\eta}_{[i]} - \widehat{\eta})'\{-\partial^2_\eta Q(\eta|\widehat{\eta})\}|_{\eta=\widehat{\eta}}(\widehat{\eta}_{[i]} - \widehat{\eta})$ (Cook and Weisberg, 1982; Zhu et al. 2001). The Q-displacement is defined by $\mathrm{QD}_i = 2\{Q(\widehat{\eta}|\widehat{\eta}) - Q(\widehat{\eta}_{[i]}|\widehat{\eta})\}$. We can similarly define case-deletion measures for deleting each observation and a subset of observations. It can be shown that QD_i and CD_i have nice decompositions, whereas standard Cook's distance and likelihood-displacement do not.

An intriguing issue is that the cluster sizes m_i may vary dramatically across i so that CD_i and QD_i are not directly comparable. We will develop a size correction A_i for the ith cluster $\boldsymbol{d}_i$, which accounts for the cluster size and the specific model for $\boldsymbol{d}_i$, and calculate an adjustment factor $A_i(\widehat{\eta})$ and the corrected case-deletion measures $\mathrm{QD}(i)^c = \mathrm{QD}(i)/A_i(\widehat{\eta})$, $i = 1,\ldots,n$. We propose an empirical Bayes method based on a "two-group" model to identify rigorously influential observations. Specifically, for a particular observation, compare the corrected case-deletion measures with a benchmark to determine whether it is influential or not. Thus, for the ith cluster (or observation), we might regard the corrected case-deletion measures as pseudo-test statistics and set up the null and alternative pseudo-hypotheses as H_i^0 : the ith cluster is not influential vs. H_i^1 : the ith cluster is influential. If the ith cluster is influential, then H_i^1 is true. In this way, we formalize the detection of influential observations as a multiple hypothesis testing problem. Following Efron (2004, 2007), we assume that the the corrected case-deletion measures follow a mixture distribution given by $f(z) = \pi_0 f_0(z) + (1 - \pi_0)f_1(z)$, where $\pi_0 = P\{null\}$, and $f_0(z)$ and $f_1(z)$ are the distributions of non-influential and influential clusters, respectively. We can then define the local pseudo false discovery rate and use the q-value, denoted $\widehat{q}(z)$, as a reference quantity for deciding which observations are influential (Storey, 2002; Efron, 2007).

Conditional Residuals. For simplicity, we focus primarily on residuals for $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{z}_i, \eta)$. Define the residual for the jth observation in the ith cluster as $R_{i,j}(\widehat{\eta}) = y_{i,j} - E_b\{g(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b}_i, \eta)\}$, where E_b is taken with

respect to $p(\boldsymbol{b}_i)$, and $g(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b}_i, \eta) = E(y_{i,j}|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b}_i)$. However, because $\boldsymbol{z}_{m,i}$ and $y_{i,j}$ may be missing, $R_{i,j}(\widehat{\eta})$ cannot be directly calculated for cases with missing data. There are many ways of eliminating $y_{i,j}$ and $\boldsymbol{z}_{m,i}$ from $R_i(\widehat{\eta})$. We will focus on $\mathrm{CR}_{i,j}^{(1)}(\widehat{\eta}) = r_{i,j}^{y}[y_{i,j} - E\{g(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b}_i, \eta)|\boldsymbol{x}_i, \boldsymbol{z}_{o,i}\}]/\Delta_{i,j}$ and $\mathrm{CR}_{i,j}^{(2)}(\widehat{\eta}) = r_{i,j}^{y}[y_{i,j} - E\{g(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b}_i, \eta)|\boldsymbol{x}_i, \boldsymbol{z}_{o,i}, \boldsymbol{r}_i, y_i\}]/\Delta_{i,j}$, where $\Delta_{i,j}$ can be either 1 or some probability weights (Robins et al., 1995). We will choose different $\Delta_{i,j}$ to increase power for detecting model misspecification.

Goodness of Fit Test with/without Incorporating the Missing Data. We use the proposed conditional residuals to construct various goodness-of-fit statistics to evaluate $H_0 : E_b\{g(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b}_i, \eta)\} = E_b\{E(y_{i,j}|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b}_i)\}$ for some η. Due to the presence of missing outcomes and covariates, it is not trivial to construct test statistics for H_0. First, we only consider $\boldsymbol{x}_{i,j}$ in constructing stochastic processes as follows: $J_1^{(k)}\{(\boldsymbol{\varphi}, t\}; \widehat{\eta}) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(\boldsymbol{x}_{i,j}'\boldsymbol{\varphi} \le t)\mathrm{CR}_i^{(k)}(\widehat{\eta})$ for $k = 1, 2$, where $(\boldsymbol{\varphi}, t) \in \Pi = \{\boldsymbol{\varphi} \in \mathbb{R}^{p_1} : \boldsymbol{\varphi}'\boldsymbol{\varphi} = 1\} \times [-\infty, \infty]$. Second, we use $E\{z_{m,i}|\boldsymbol{x}_i, \widehat{\alpha}\}$ to fill in $z_{m,i}$, which results in a new value, $z_{m,i}^{*}$, and then construct several stochastic processes, such as $J_2\{(\widetilde{\boldsymbol{\varphi}}, t\}; \widehat{\eta}) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(\boldsymbol{c}_{i,j}'\widetilde{\boldsymbol{\varphi}} \le t) \times \mathrm{CR}_i^{(2)}(\widehat{\eta})$, in which $\boldsymbol{c}_{i,j} = (\boldsymbol{x}_{i,j}', \boldsymbol{z}_{i,j*}')'$ and $\widetilde{\boldsymbol{\varphi}} \in \{\widetilde{\boldsymbol{\varphi}} \in R^p : \widetilde{\boldsymbol{\varphi}}'\widetilde{\boldsymbol{\varphi}} = 1\}$. Because $I(\boldsymbol{x}'\boldsymbol{\varphi} \le t)$ in $J_1^{(k)}\{(\boldsymbol{\varphi}, t\}; \widehat{\eta})$ does not involve the missing covariates $\boldsymbol{z}$, this can lead to loss of power in detecting the misspecification involving $\boldsymbol{z}$. In particular, if the fraction of missing covariates is moderate to large, then it is inefficient to drop all the information in $\boldsymbol{z}$. We will develop test statistics based on these stochastic processes, such as the integrated conditional moment (ICM) test of Bierens (1990) and establish their asymptotic null distributions and power (Bierens and Ploberger, 1997; Inglot and Ledwina, 2001, 2004, 2006). Finally, we will use a resampling method to approximate p-values of the ICM statistics (Lin, Wei, and Ying, 2001; Zhu et al., 2008; Kosorok, 2008).

Methodology for Diagnostic Methods for Joint Models for Longitudinal and Survival Data. Because the development for joint models is close to that for longitudinal models above, we just highlight their differences. Consider n independent clusters $\boldsymbol{d}_{o,i} = \{\boldsymbol{X}_i = (\boldsymbol{x}_{i,j} : j = 1, \ldots, m_i), \boldsymbol{z}_i, \delta_i, V_i\}$, $i = 1, \ldots, n$, where $\boldsymbol{z}_i$ is a vector of baseline covariates, $V_i = T_i \wedge C_i$ and $\delta_i = I(T_i \le C_i)$, where T_i and C_i are the potential failure and censoring times. Let $\boldsymbol{b}_i$ be a vector of random effects for the ith cluster, and assume that $p(\boldsymbol{b}_i, \boldsymbol{X}_i, \boldsymbol{z}_i, V_i, \delta_i; \eta) = p(V_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{z}_i, \eta)p(\boldsymbol{X}_i|\boldsymbol{b}_i, \boldsymbol{z}_i, \eta)p(\boldsymbol{b}_i; \eta)$. Assume that the hazard and survival functions of T_i are $\lambda_t(V_i|\mathrm{c}_i) = h_0(V_i)\exp\{k(\boldsymbol{b}_i, \boldsymbol{z}_i, \eta)\}$ and $S_t(V_i|\mathrm{c}_i) = \exp[-\exp\{k(\boldsymbol{b}_i, \boldsymbol{z}_i, \eta)\}H_0(V_i)]$, where $H_0(y) = \int_0^y h_0(v)dv$, and assume that the cumulative baseline hazard function $H_0(V)$ is a step function with jumps only at the V_i so that $H_0(V_i) = \sum_{V_j \le V_i} h_0(V_j)$, $i = 1, \ldots, n$.

Case-deletion Measures. Because of the presence of the infinite dimensional parameter $h_0(\cdot)$, we develop a profile likelihood method to quantify the effects of deleting a particular observation/cluster on the finite dimensional parameter η. Let $L_c\{\eta, h_0(\cdot)|D_{com}\}$ be the logarithm of the complete-data likelihood. Define $Q_0(\eta|\widehat{\eta}) = E[L_c\{\eta, \widehat{h}_0(\cdot|\eta)|D_{com}\}|D_{obs}, \widehat{\eta}]$ and $Q_{0[i]}(\eta|\widehat{\eta}) = E[L_c\{\eta, \widehat{h}_{0[i]}(\cdot|\eta)|D_{com[i]}\}|D_{obs}, \widehat{\eta}]$ for each i, where $\widehat{h}_0(\cdot|\eta) = \mathrm{argmax}_{h_0}E[L_c\{\eta, h_0(\cdot)|D_{com}\}|D_{obs}, \widehat{\eta}]$ and $\widehat{h}_{0[i]}(\cdot|\eta) = \mathrm{argmax}_{h_0}E[L_c\{\eta, h_0(\cdot)|D_{com[i]}\}|D_{obs}, \widehat{\eta}]$ for fixed η. Then define Cook's distance and the Q-displacement as $\mathrm{CD}_i = (\widehat{\eta}_{[i]} - \widehat{\eta})'\{-\partial_{\eta}^2 Q_0(\eta|\widehat{\eta})\}|_{\eta=\widehat{\eta}}(\widehat{\eta}_{[i]} - \widehat{\eta})$ and $\mathrm{QD}_i = 2\{Q_0(\widehat{\eta}|\widehat{\eta}) - Q_0(\widehat{\eta}_{[i]}|\widehat{\eta})\}$, respectively, where $\widehat{\eta}_{[i]} = \mathrm{argmax}_{\eta}Q_{0[i]}(\eta|\widehat{\eta})$. It can be shown that both CD_i and QD_i have a nice decomposition, which facilitates model diagnostics. For $h_0(\cdot)$, we use $\mathrm{dh}_{0,i} = \max_j |\widehat{h}_0(V_j) - \widehat{h}_{0[i]}(V_j)|$ to measure the difference between $\widehat{h}_0(\cdot)$ and $\widehat{h}_{0[i]}(\cdot)$ and $\mathrm{dH}_i = \max_j |\widehat{H}_0(V_j) - \widehat{H}_{0[i]}(V_j)|$ to measure the distance between $\widehat{H}_0(\cdot)$ and $\widehat{H}_{0[i]}(\cdot)$, where $\widehat{H}_0(u) = \sum_{V_j \le u}\widehat{h}_0(V_j)$ and $\widehat{H}_{0[i]}(u) = \sum_{V_j \le u}\widehat{h}_{0[i]}(V_j)$.

Conditional Martingale Residuals. Define a conditional martingale residual process as follows: $\mathrm{CM}_i(t, \eta) = N_i(t) - \int_0^t I(V_i \ge u)E[\exp\{k(\boldsymbol{b}_i, \boldsymbol{z}_i, \eta)\}|\boldsymbol{d}_{o,i}, \eta]h_0(u)du$, $i = 1, \cdots, n$, where $N_i(t) = \delta_i I(y_i \le t)$ and the expectation is with respect to the conditional distribution of $\boldsymbol{b}_i$ given $\boldsymbol{d}_{o,i}$. The conditional martingale residual process $\mathrm{CM}_i(t)$ evaluated at $\widehat{\eta}$ is given by $\widehat{\mathrm{CM}}_i(t, \widehat{\eta}) = N_i(t) - \int_0^t I(V_i \ge u)E[\exp\{k(\boldsymbol{b}_i, \boldsymbol{z}_i, \eta)\}|\boldsymbol{d}_{o,i}, \widehat{\eta}]\widehat{h}_0(u)du$. We will define and examine score vector residuals, Cox-Snell residuals, censoring consistent residuals, and Schoenfeld residuals (Lawless, 2003; Leon and Tsai, 2004) in the presence of missing data and study their theoretical properties as with the conditional martingale residuals. It should be noted that we can also define residuals for the model $p(\boldsymbol{X}_i|\boldsymbol{b}_i, \boldsymbol{z}_i, \eta)$ of longitudinal data as discussed above.

Goodness of Fit Tests. We will use the proposed conditional martingale residuals to construct various goodness-

of-fit statistics to check formally correct specification of the joint models. We will develop a resampling method to approximate the p-value of the corresponding test statistics and theoretically examine the asymptotic properties of the test statistics under the null and alternative hypotheses. We will also develop specific strategies for incorporating the missing data into goodness-of-fit statistics, and consider goodness-of-fit statistics based on other types of residuals, such as the censoring consistent residuals (Leon and Tsai, 2004).

5.4 Aim 4: Inferences for Sensitivity Analyses of Missing Data

Profile Estimation. We consider a much more general framework than that described in Section **4.2** for conducting formal inferences for sensitivity analyses. Our set-up does not require specifying the entire model for both the full data and missingness mechanism; as in the case where a likelihood analysis would be undertaken. A partially specified model is sufficient as long as the parameter of interest, β, can be estimated for each value of the sensitivity parameter, δ, which make the proposed methods more broadly applicable; in particular, there is an extensive literature on estimating equations having this structure.

Assume, then, for a general problem, an iid sample of size n is available, and, for inference on β, there exists an estimating equation $\boldsymbol{U}_n(\beta, \delta)$ such that $\widehat{\beta}(\delta)$ satisfies $\boldsymbol{U}_n\{\widehat{\beta}(\delta), \delta)\} = \mathbf{0}$ for fixed δ. With maximum likelihood, $\boldsymbol{U}_n$ is the score equation. Although one might try to estimate δ by solving simultaneously for β and δ, the model used to define $\boldsymbol{U}_n$ is generally unverifiable, so that results based on this approach may be very misleading, as there may be multiple competing models which are equally compatible with the observed data. Even if one has a strong prior belief about the model form, there may be insufficient information in the data to identify the model parameters, leading to instability in estimation. To make indirect inference on β, one needs to consider how the estimate of β behaves as δ is varied, which demands careful analysis of $\widehat{\beta}(\delta)$.

Consider $\widehat{\beta}(\delta)$ for δ in a compact set, i.e., $\delta \leq \Delta$, $0 < \Delta < \infty$. The upper bound Δ on δ may be chosen to reflect a level of informativeness which is not expected in practice, but represents an extreme scenario. It is straightforward to permit δ to have dimension > 1 with little additional complexity beyond defining the relevant region for δ. It is well known that, under a correctly specified model fit with a given $\delta = \delta_0$, where δ_0 is the true value of δ, and under mild regularity conditions, the estimator for β is consistent for the true value β_0 and asymptotically normal, assuming $\boldsymbol{U}_n$ is a valid estimating equation.

Under a misspecified dropout process, i.e., $\delta \neq \delta_0$, and for large n, assume that $\widehat{\beta}(\delta) \xrightarrow{p} \beta_*(\delta)$, where $\beta_*(\delta)$ satisfies $\lim_{n\to\infty} \boldsymbol{U}_n\{\beta*(\delta), \delta\} = \mathbf{0}$ with $\beta_*(\delta)$ not necessarily equal to β_0. Moreover, even with misspecified δ, as $n \to \infty$, the score function $\boldsymbol{U}_n$ is generally roughly quadratic in the neighborhood of $\beta_*(\delta)$ for fixed δ, and the limiting distribution of $n^{1/2}\{\widehat{\beta}(\delta) - \beta_*(\delta)\}$ is mean zero normal with variance-covariance matrix $\Gamma(\delta)$, which has a sandwich form, similar to that for the maximum likelihood estimator under a misspecified model (Stefanski and Boos, 2002). In certain applications, it may be possible to develop robust variance estimators $\widehat{\Gamma}(\delta)$, say, along the lines of Todem and Fine (2008). However, with a general model and estimating equation $\boldsymbol{U}_n$, a simple plug-in variance estimator is unclear. An alternative variance estimator is described below.

These results are important. If one is fitting several models with different missingness mechanisms, then it is not generally possible for all models to be correct. At best, one model might be valid. Presentation of from multiple analyses would then in theory need to account for the fact that model misspecification is present in at least one of those analyses. Moreover, it may not be clear which of the analyses is correctly specified, if any.

Following the arguments in Todem and Fine (2008), these pointwise results may be made uniform under certain smoothness conditions, using empirical process arguments (Kosorok, 2008). That is, one may show that $\widehat{\beta}(\delta)$ is uniformly consistent for $\beta_*(\delta)$ and $J(\delta) = n^{1/2}\{\widehat{\beta}(\delta) - \beta_*(\delta)\}$ converges weakly to a Gaussian process with mean 0 and positive definite covariance function. In general, estimators $\widehat{\beta}(\delta)$ at different values of δ will be highly correlated, and treating the estimates as independent could lead to misleading inferences. Uniformity may seem highly technical and of limited practical relevance; however, it provides a key technical justification for the inferential methods described below.

One may use the bootstrap method to estimate the standard errors of the estimators of $\beta_*(\delta)$ (Efron and Tibshirani, 1993); it may also be used to approximate the distribution of $\widehat{\beta}(\delta)$ as a process, which is quite complex and does not lead to simple analytic testing procedures. The validity of the bootstrap follows automatically from empirical process theory under the regularity conditions given in van der Vaart and Wellner (2000), and holds even under model misspecification, as the estimators are functions of the empirical data distribution.

Sensitivity Testing. In general, we wish to use $\widehat{\beta}(\delta)$ to evaluate the null hypothesis $H_{01} : C\beta = c$, where β $p \times 1$ is a parameter of the outcome process, C $r \times p$ is a contrast matrix and c is a $r \times 1$ vector of constants. This framework allows for composite hypotheses. In the special case of testing a particular covariate effect, such as treatment or some risk factor, on the outcome process, C is the $1 \times p$ vector with a one in the position of the regression parameter for the covariate and zeros elsewhere, and $c = 0$. If a non-identifiable, non-ignorable model is assumed to generate both the outcome and dropout processes under investigation, H_{01} cannot be tested without additional restrictions. For $\delta \in \mathbb{R}^+$, if the true value δ_0 of the sensitivity parameter is known, the true hypothesis is $H_{02} : C\beta_*(\delta_0) = c$, where $\beta_*(\delta_0) = \beta$. As discussed above, in many cases, δ_0 is unknown and may not be estimable from observed data. We propose a global sensitivity test that uses the trivial inequalities $\inf_\delta \|C\beta_*(\delta) - c\| \leq \|C\beta_*(\delta_0) - c\| \leq \sup_\delta \|C\beta_*(\delta) - c\|$ to make rigorous inferential statements about H_{02}, where $\|.\|$ is the Euclidean norm. We can evaluate H_{02} by conservatively testing the hypothesis $\inf_\delta \|C\beta_*(\delta) - c\| = 0$. Clearly, when $\inf_\delta \|C\beta_*(\delta) - c\|$ is strictly greater than 0, $\|C\beta_*(\delta_0) - c\|$ will be greater than 0 as well. The infimum hypothesis is formally defined as $H_{03} : \inf_\delta \|C\beta_*(\delta) - c\| = 0$. In the case of a one-dimensional covariate, an ad hoc test may be conducted by constructing simultaneous confidence bands for $\beta_*(\delta)$ given δ. If the band includes 0 for any δ, then H_{02} cannot be rejected. In general, if a simultaneous confidence band for $C\beta_*(\delta)$ excludes c for all δ, then one rejects the associated null. Such bands also identify those δ at which the null of is rejected, which may be useful in understanding how covariate effects change as a function of the missingness model.

We now propose a formal infimum statistic to evaluate H_{03},

$$T = \inf_\delta \left\{ (C\widehat{\beta}(\delta) - c)'(C\widehat{\Gamma}(\delta)C')^{-1}(C\widehat{\beta}(\delta) - c) \right\}$$

where $\widehat{\beta}(\delta)$ is the corresponding estimator for $\beta_*(\delta)$ and $\widehat{\Gamma}(\delta)$ is an estimator of the variance-covariance matrix of $\widehat{\beta}(\delta)$, as described above. The test statistic T rejects the null for unusual large values. While for each fixed δ, the test process follows a simple chi square distribution under the null, the asymptotic distribution of the estimator $\widehat{\beta}(\delta)$ as a process in δ is quite complicated, and the distribution of T is analytically intractable.

One may use the nonparametric bootstrap to generate the distribution of the infimum test statistic and to construct the confidence bands, using the approach described in Section 2. Let $\widehat{\beta}^s(\delta)$ and T^s, $s = 1, \ldots, S$ denote the estimators and infimum tests computed in S bootstrap samples. One rejects the null at level α if the observed test statistic is larger than the $(1 - \alpha)$ percentile of empirical distribution of $T^s, s = 1, \ldots, S$. To define the confidence bands for the kth component of $\beta_*(\delta)$, denoted by $\beta_{*,k}(k = 1, \ldots, p)$, we define ϑ_α as the $(1 - \alpha)$th empirical percentile of $\{\sup_{\delta \in [0,\Delta]} \|\widehat{\beta}^s_k(\delta) - \widehat{\beta}_k(\delta)\|\}_{s=1}^S$, where the subscript k refers to the kth component of the corresponding vector. A $(1 - \alpha)$ simultaneous confidence band for $\{\beta_{*,k}(\delta) : \delta \in [0, \Delta]$ is $\{b_{*,k}(\delta) : [0, \Delta] \to \mathbb{R}^p; \ \|b_{*,k}(\delta) - \widehat{\beta}_k(\delta)\| < \vartheta_\alpha\}$. In general, simultaneous confidence bands for $C\beta_*(\delta) - c$ take the form $\{x(\delta) : [0, \Delta] \to \mathbb{R}^r; \ \|x(\delta) - C\widehat{\beta}(\delta) + c\| < \widetilde{\vartheta}_\alpha\}$, where $\widetilde{\vartheta}_\alpha$ is the $(1 - \alpha)$th empirical percentile of $\{\sup_{\delta \in [0,\Delta]} \|C\widetilde{\beta}^s(\delta) - C\widehat{\beta}(\delta)\|\}_{s=1}^S$.

A pointwise approach might also be used to carry out sensitivity testing for $\|C\beta_*(\delta) - c\| = 0$ given a finite number of values for δ. The method consists of letting the sensitivity parameter take values in a set $A = \{\delta_1, \ldots, \delta_Q\} \subset \mathbb{R}^+$, $Q < \infty$, and evaluating the hypothesis $\|C\beta_*(\delta) - c\| = 0$. Specifically, for $\delta \in A$, simultaneous pointwise confidence intervals may be constructed for $\{C\beta_*(\delta) - c\}$ using standard error estimates for $\widehat{\beta}(\delta)$, with a multiplicity adjustment. If all intervals exclude 0, then $C\beta_*(\delta) \neq c$ at those δ. This approach should be undertaken carefully, as the choice of points in A may be arbitrary and may miss δ where H_{03} holds. Thus, one cannot formally test H_{01} using finite δ. Moreover, there may be reduced power with large Q, where the multiplicity adjustment for controlling the overall type I error may be quite conservative. The global approach described above provides a systematic method for dealing with these issues.

Numerical Studies. Extensive simulation studies will be carried out to evaluate performance of the procedures in small, moderate, and large sample sizes typical those in cancer clinical trials. Simulation scenarios will be based on data sets compiled by Core B, the International Breast Cancer Study Group (IBCSG) Trial VI, and other CALGB studies. Size and power of the infimum tests and coverage of the simultaneous confidence bands will be assessed; the former will be compared to standard tests that assume ignorability and to alternative tests procedures that attempt to estimate the sensitivity parameter instead of profiling in estimation. Analyses of E1684, E1694, and other cancer trials compiled by Core B will be conducted to develop experience in applying

the sensitivity testing procedures. Comparison with naive approaches, in which multiple analyses are presented without formal adjustment will be undertaken.

5.5 Software Development and Dissemination

All public-use software developed in this project will be made available on dedicated pages on the Program Project website, including downloadable code and instructions, documentations, and examples.

Aim 1: We will work with Core C to translate our code into a software package for covariate adjustment that has modules to carry out treatment-specific model selection according to a user-specified choice (among those recommended for this purpose) and to construct the final treatment effect estimate; naive and adjusted (for model selection and estimation of the regressions) standard errors and confidence intervals will be provided. Depending on our results, we may include a module to carry out imputation or model selection using the LSA adaptive LASSO approach in the case of missing covariates. Because our target users are trialists who may be involved in regulatory submissions, it will be essential to provide a SAS implementation. We will also develop a separate software package that can be used to obtain the proposed estimator for partially randomized trials.

Aim 2: We will work with Core C to adapt our code for both methods for handling drop-out into user-friendly software. For the covariate adjustment methods, we will develop a separate module handling this case.

Aim 3: Each of the new methods will first undergo developmental implementation until the procedure has been validated via simulation studies and data analyses. Core C will then assist with taking the implementation and developing it into a usable and robust software package appropriate for dissemination in both R and SAS formats. In addition to the materials on the web page, the new software will be communicated through presentations and short courses at appropriate professional meetings.

Aim 4: Software will be developed for R and SAS, which will be compartmentalized, so that the end user provides a function that carries out the analysis at a fixed value of δ, enabling broad applicability in oncology. The user will also need to specify a range of interest for the sensitivity parameter. The software will provide as output the infimum test and its corresponding p-value, as well as point estimates of the parameters β as a function of δ, and simultaneous confidence bands. Bootstrap variance estimates will be provided.

5.6 Timetable

For all Aims, manuscripts will be prepared and submitted as results worthy of publication become available.

Aim 1: Years 1 and 2 will be devoted to the comprehensive study of model selection methods. Once recommended model selection strategies have been established, in Year 3 we will begin work on the approaches to inference after model selection and missing covariates. In Year 4, we will study the proposed method for partially randomized trials. Throughout, as methods mature, we will begin working with Core C on translating them into software. In Year 5, we will complete software development and manuscripts, including a paper for a clinical audience explaining and promoting the covariate adjustment methods.

Aim 2: Years 1 and 2 will on focus study of the extension of the covariate adjustment methods to the case of drop-out and their implementation in software. Years 3 and 4 will be devoted to study of the methods for longitudinal analysis and software implementation. In Year 5, we will wrap up work on both tasks.

Aim 3: Each of the three phases of research will require about one year of development and theoretical work, about 6 months of simulation and data analysis evaluation, and 6 months to 1 year of implementation and development. We will stagger initiation of each phase so that they are a year apart and the theoretical work can be done sequentially. Results of the first phase will be completed by the middle of Year 2, and we should be able to complete the first three phases and start the fourth phase before the end of Year 5.

Aim 4: Years 1 and 2 will be spent on work related to the foundations of the methodology. A proof of concept paper will be written for an applied statistical audience to disseminate the main ideas. Years 3 and 4 will involve rigorous theoretical study of the methods, extensive simulation work, and data analysis. Years 4 and 5 will be focused on software development and preparing additional papers for publication. These will include a paper on the theoretical aspects for a biostatistical methods journal and a paper targeted to practitioners and clinicians in the oncology community explicating in layman's terms why such sensitivity analyses should be undertaken, the methods at a conceptual level, and the pitfalls of reporting results without such adjustments.

6 INCLUSION ENROLLMENT REPORT

N/A

7 BIBLIOGRAPHY AND REFERENCES CITED

Albert, P. S. and Follmann, D. A. (2000). Modeling repeated count data subject to informative dropout. *Biometrics* **56**, 667–677.

Anstrom, K. J. and Tsiatis, A. A. (2001). Using propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics* **57**, 1207–1218.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–972.

Bang, H. and Tsiatis, A. A. (2000). Estimating medical costs with censored data.. *Biometrika* **87**, 329–343.

Bang, H. and Tsiatis, A. A. (2002). Median regression with censored cost data. *Biometrics* **58**, 643–649.

Benjamani, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**, 60–83.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica* **58**, 1443–1458.

Bierens, H. J. and Ploberger, W. (1997). Asymptotic theory of integrated conditional moment test. *Econometrica* **65**, 1129–1151.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and clustering of predictors with OSCAR. *Biometrics* **64**, 115–123.

Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, in press.

Bondell, H. D. and Li, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *Journal of the Royal Statistical Society, Series B*, in press.

Boos, D. D. (2003). Introduction to the bootstrap world. *Statistical Science* **18**, 168–174.

Boos, D. D., Stefanski, L. A., and Wu, Y. (2008). Fast FSR variable selection with applications to clinical trials. *Biometrics*, doi:10.1111/j.1541-0420.2008.01127.x. NIHMSID:NIHMS76165.

Breiman, L. (1996s). Bagging predictors. *Machine Learning* **24**, 123–140.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350–2383.

Brown, E. R. and Ibrahim, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221–228.

Buhlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics* **30**, 927–961.

Burton, S. W. (1991). A review of fluvoxamine and its uses in depression. *International Clinical Psychopharmacology* **6**, 1–17.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, in press.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Boca Raton, Florida: Chapman and Hall/CRC Press.

Carroll, R. J., Wu, C. F. J, and Ruppert, D. (1988). The effect of estimating weights in generalized least squares. *Journal of the American Statistical Association* **83**, 1045–1054.

Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference* **67**, 45–65.

Chen, Q. and Ibrahim, J. G. (2006). Semiparametric models for missing covariate and response data in regression models. *Biometrics* **62**, 177–184.

Chi, Y. Y. and Ibrahim, J. G. (2007). Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions. *Statistica Sinica* **17**, 445–462.

Cho, H., Ibrahim, J. G., Sinha, D., and Zhu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, in press.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics* **64**, 1062–1069.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19**, 15–18.

Cook, R. D. (1986). Assessment of local influence (with discussion), *Journal of the Royal Statistical Society, Series B* **48**, 133–169.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

Copas, J. B. and Eguchi, S. (2001). Local sensitivity approximation for selectivity bias. *Journal of the Royal Statistical Society, Series B* **63**, 871–895.

Copas, J. B. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society, Series B* **67**, 459–513.

Crews, H. B., Boos, D. D., and Stefanski, L. A. (2008). Fast false selection rate methods for analyzing response surface experimental designs. Institute of Statistics Mimeo Series # 2619, Raleigh.

Critchley, F., Atkinson, R. A., Lu, G., and Biazi, E. (2001). Influence analysis based on the case sensitivity function. *Journal of the Royal Statistical Society, Series B* **63**, 307–323.

Davidian, M., Tsiatis, A.A., and Leon, S. (2005) Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with discussion and rejoinder). *Statistical Science* **20**, 261–301. PMCID:PMC2600547.

Diggle, P. J, Farewell, D,. and Henderson, R. (2007). Longitudinal data without dropout: objectives, assumptions and a proposal (with discussion). *Applied Statistics* 56, 499–550.

Diggle, P. J., Sousa, I., and Chetwynd, A. G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture. *Statistics in Medicine* **27**, 2981–2998.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of American Statistical Association* **99**, 96–104.

Efron, B. (2007). Size, power and false discovery rates. *Annals of Statistics* **35**, 1351–1377.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle property. *Journal of the American Statistical Association* **96**, 1348–1360.

Fine, J. P. and Tsiatis, A. A. (2000). Testing for differences in survival with delayed ascertainment. *Biometrics* **56**, 145–153.

Friedrich, T. (1991). Die Fisher-Information und symplektische Strukturen. *Mathematische Nachrichten* **153**, 273–296.

Greene, W. F. and Cai, J. (2004). Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics* **60**, 987–996.

Grouin, J. M., Day, S., and Lewis, J. (2004). Adjustment for baseline covariates: an introductory note. *Statistics in Medicine* **23**, 697–699.

Gu, M. G. and Zhu, H. T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society, Series B* **63**, 339–355.

Hall, P. (1986). On the bootstrap and confidence intervals. *Annals of Statistics* **14**, 1431–1452.

Henderson, R, Diggle, P, and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Hens , N., Aerts , M., and Molenberghs , G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine* **25**, 2502–2520.

Hogan, J. W. and Laird, N. M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine* **16**, 259–72.

Horowitz, J. L. and Manski, C. F. (2006). Identification and estimation of statistical functionals from incomplete data. *Journal of Econometrics* **132**, 445–469.

Huang, X., Stefanski, L. A., and Davidian, M. (2009). Latent-model robustness in joint models for a primary endpoint and a longitudinal process. *Biometrics*, in press. NIHMSID:NIHMS89685.

Huang, Y. and Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika* **85**, 785–798.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer.

Inglot, T. and Ledwina, T. (2001). Intermediate approach to comparison of some goodness-of-fit tests. *Annals of the . Institute of Statistical Mathematics* **53**, 810–834.

Inglot, T. and Ledwina, T. (2004). On consistent minimax distinguishability and intermediate efficiency of Cramer-von Mises test. *Journal of Statistical Planning and Inference* **124**, 453–474.

Inglot, T. and Ledwina, T. (2006). Intermediate efficiency of some max-type statistics. *Journal of Statistical Planning and Inference* **136**, 2918–2935.

Johnson, B.A. and Tsiatis, A. A. (2005). Semiparametric inference in observational duration-response studies, with duration possibly right-censored. *Biometrika* **92**, 605–618.

Jung S. H., Hui, S. (2002). Sample size calculations to compare K different survival distributions. *Lifetime Data Analysis* **8**, 361–373.

Jung S. H., Kang, S. J., McCall, L., Blumenstein, B. (2005). Sample size computation for noninferiority log-rank test. *Journal of Biopharmaceutical Statistics* **15**, 957–967.

Jung, S. H. (2008). Sample size calculation for the weighted rank statistics with paired survival data. *Statistics in Medicine* **27**, 3350–3365.

Jung S. H., Kim, C., Chow, S. C. (2008). Sample size calculation for the log-rank tests for multi-arm trials with a common control. *Journal of Korean Statistical Society* **37**, 11–22.

Kang, D. Y. J. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion and rejoinder). *Statistical Science* **20**, 661–301.

Kenward, M. G., Goetghebeur, E. J. T., and Molenberghs, G. (2001). Sensitivity for incomplete categorical data. *Statistical Modelling* **1**, 31–48.

Kenward, M. G., Lesaffre, E., and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**, 945–953.

Kirkwood, J. M., Ibrahim, J. G., Sosman, J. A., Sondak, V. K., Agarwala, S. S., Ernstoff, M. S., and Rao, U. (2001). High-dose interferon Alfa-2b significantly prolongs relapse-free and overall survival compared with the GM2-KLH/QS-21 vaccine in patients with resected stage IIB-III melanoma: results of intergroup trial E1694/S9512/C509801. *Journal of Clinical Oncology* **19**, 2370–2380.

Kirkwood, J. M., Strawderman, M. H., Ernstoff, M. S., Smith, T. J., Borden, E. C., and Blum, R. H. (1996). Interferon alfa–2b adjuvant therapy of high-risk resected cutaneous melanoma: The Eastern Cooperative Oncology Group Trial EST 1684. *Journal of Clinical Oncology* **14**, 7–17.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Lang, S. (1999). *Fundamentals of Differential Geometry*. Graduate Texts in Mathematics, 191. New York: Springer.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd edition. Hoboken: Wiley.

Lee, S. Y. and Tang, N. S. (2004). Local influence analysis of nonlinear structural equation models. *Psychometrika* 69, 573–592.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: facts and fiction", *Econometric Theory* **21**, 29–59.

Leeb, H. and Pötscher, B. M. (2005). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* **34**, 2554–2591.

Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges's estimator. *Journal of Econometrics* **142**, 201–211.

Leon, L. F. and Tsai, C. L. (2004). Functional form diagnostics for Cox's proportional hazards model. *Biometrics* **60**, 75–84.

Lesaffre, E., Molenberghs, G., and Dewulf, L. (1996). Effects of dropouts in longitudinal study: An application of a repeated ordinal model. *Statistics in Medicine* **15**, 1123–1141.

Lesaffre, E. and Senn, S. (2003). A note on nonparametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine* **22**, 3586–3596.

Lin, D. Y., Wei, L. J., and Ying, Z. L. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* 58, 1–12.

Lin, Y. and Zhang, H. H. (2007). Component Selection and smoothing in smoothing spline analysis of variance models - COSSO. *Annals of Statistics* **34**, 2272–2297.

Liu, Y., Zhou, H., and Cai, J. (2009). Estimated pseudo-partial-likelihood method for correlated failure time data with auxiliary covariates. *Biometrics*, in press.

Little, R. J. A. and Rubin, D.B. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models: Comments. *Journal of the American Statistical Association* **94**, 1130–1132.

Little, R. J. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497–512.

Livingston, P. O., Wong, G. Y. C., Adluri, S., Tao, Y., Padavan, M., Parente, R., Hanlon, C., Calves, M. J., Helling, F., and Ritter, G. (1994). Improved survival in stage III melanoma patients with GM2 antibodies: a randomized trial of adjuvant vaccination with GM2 ganglioside. *Journal of Clinical Oncology* **12**, 1036–1044.

Long, Q., Little, R. J., and Lin, X. (2008). Causal inference in hybrid intervention trials involving treatment choice. *Journal of the American Statistical Association* **103**, 474–484.

Lu, X. and Tsiatis, A. A. (2008). Improving the efficiency of the logrank test using auxiliary covariates. *Biometrika* **95**, 679–694. NIHMSID:NIHMS:79934.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects. *Statistics in Medicine* **23**, 2937–2960.

Manski, C. (2003). *Partial Identification of Probability Distributions*. New York: Springer.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* **89**, 633–644.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment, and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine* **21**, 2917–2930.

Raab, G. M., Day, S., and Sales, J. (2000). How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials* **21**, 330–342.

Robins, J. M., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology-Methodological Issues*, eds. N. Jewell, K. Dietz, V. Farewell. New York: Birhauser, pp. 297–331.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated measures outcomes with non-ignorable non-response. *Journal of American Statistical Association* **93**, 1321–1339.

Rotnitzky, A., Scharfstein, D. O., Su, T.-L., and Robins, J. (2001). Methods for conducting sensitivity analysis for trials with potentially nonignorable competing causes of censoring. *Biometrics* **57**, 103–113.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1987). *Multiple Imputations for Nonresponse in Surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

Schaubel, D. E. and Cai, J. (2006). Rate/mean regression for multiple-sequence recurrent events data with missing event category. *Scandinavian Journal of Statistics* **33**, 191–207.

Senn, S. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* **8**, 467–475.

Shen, X., Huang, H.-C., and Ye, J. (2004). Inference after model selection. *Journal of the American Statistical Association* **99**, 751–762.

Shepherd, B. E., Gilbert, P. B., Jemimimai, Y., and Rotnitzky, A. (2006). Sensitivity analysis comparing outcomes in a subset selected post randomization, conditional on covariates, with application to HIV trials. *Biometrics* **62**, 332–342.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58**, 742–753.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56**, 29–38.

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101**, 1619–1637.

Tan, Z. (2007). Understanding OR, PS,, and DR. *Statistical Science* **22**, 560–568.

Tangen, C. M. and Koch, G. G. (1999). Nonparametric analysis of covariance for hypothesis testing with logrank and Wilcoxon scores and survival-rate estimation in a randomized clinical trial. *Journal of Biopharmaceutical Statistics* **9**, 307–338.

Ten Have, T. R., Reboussin, B. A., Miller, M. E., and Kunselman, A. (2002). Mixed effects logistic regression models for multiple longitudinal binary functional limitation responses with informative dropout confounding by baseline outcomes. *Biometrics* **58**, 137–144.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Todem, D., Kim, K., and Lesaffre, E. (2006). A sensitivity approach to modeling longitudinal bivariate ordered data subject to informative dropouts. *Health Service and Outcomes Research Methodology* **6**, 37–57.

Todem, D. and Fine, J.P. (2008). Global sensitivity testing for longitudinal data subject to potentially informative dropout. *Biometrics*, invited revision.

Troxel, A. B., Ma, G., and Heitjan, D. F. (2004). An index of local sensitivity to nonignorability. *Statistica Sinica* **14**, 1221–1237.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809–834.

Tsiatis, A.A. and Davidian, M. (2007) Discussion of "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," by J.D.Y. Kang and J.L. Schafer. *Statistical Science* **22**, 569–573. PMCID: PMC2397555.

Tsiatis, A.A. and Davidian, M. (2007) Discussion of "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," by J.D.Y. Kang and J.L. Schafer. *Statistical Science* **22**, 569–573. PMCID: PMC2397555.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27**, 4658–4677. NIHMSID:NIHMS46485.

van der Vaart, A. W. and Wellner, J. A. (2000). *Weak Convergence and Empirical Processes, With Applications to Statistics, Second Edition*. New York: Springer.

Vansteelandt, S., Goetghebeur, E., Kenward, M.G., and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in sensitivity analysis. *Statistica Sinica* **16**, 957–979.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. (2001). Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics* **57**, 7–14.

Wang, H. and Leng, C. (2007). Unified LASSO estimation via least squares approximation. *Journal of the American Statistical Association* **102**, 1039–1048.

Wu, M.C. and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association* **102**, 235–243.

Wulfsohn, M. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Zeng, D. and Cai, J. (2005).Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Annals of Statistics* **33**, 2132–2163.

Zhang, H. H. and Lu, W. (2007). Adaptive LASSO for Cox's proportional hazards model. *Biometrika* **93**, 1–13.

Zhang, H. H., Wahba, G., Lin, Y., Volker, M., Ferris, M., Klein, R., and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association* **99**, 659–672.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inference in randomized clinical trials using auxiliary covariates. *Biometrics* **64**, 707–715. PMCID: PMC2574960.

Zhao, H. and Tsiatis, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* **84**, 339–348.

Zhao, H. and Tsiatis, A. A. (1999). Efficient estimation of the distribution of quality-adjusted survival time. *Biometrics* **55**, 1101–1107.

Zhao, H. and Tsiatis, A. A. (2000). Estimating mean quality adjusted life-time with censored data. *Sankhya, Series B* **62**, 175–188.

Zhu, H. T., Ibrahim, J. G., Lee, S. Y., and Zhang, H. P. (2007). Perturbation selection and influence measures in local influence analysis. *Annals of Statistics* **35**, 2565–2588.

Zhu, H. T., Ibrahim, J. G., Tang, N. S., and Zhang, H. P. (2008). Diagnostic measures for empirical likelihood of general estimating equations. *Biometrika* **95**, 489–507.

Zhu, H. T. and Lee, S. Y. (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B* **63**, 111–126.

Zhu, H. T., Lee, S. Y., Wei, B. C., and Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika* **88**, 727–37.

Zhu H. T. and Zhang, H. P. (2004). A diagnostic procedure based on local influence. *Biometrika* **91**, 579–589.

Zhu, L. X. (2005). *Nonparametric Monte Carlo Tests and Their Applications*. New York: Springer.

Zou, H. (2006) The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429

Zou, H., and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, in press.

8 PROTECTION OF HUMAN SUBJECTS

Although the proposed research indirectly involves human subjects through the preparation, in Core B, of de-identified data sets from identifiable patient data sources, the investigators on Project 2 will have access only to the de-identified data. Thus, the investigators on Project 2 will have no access to any identifiable patient information.

9 INCLUSION OF WOMEN AND MINORITIES

The methods we develop will be applicable to studies with both women and minorities and also to studies which examine treatment differences adjusted for gender, ethnicity and race. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, many of the existing data sets to be studied and provided by Core B include women and minorities, although we will not be generating any new data involving human subjects.

10 TARGETED/PLANNED ENROLLMENT TABLE

N/A

11 INCLUSION OF CHILDREN

The methods we develop will be applicable to studies with children and also to studies which examine treatment differences adjusted for age. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, some of the existing data sets to be studied and provided by Core B may include children, although we will not be generating any new data involving human subjects.

12 VERTEBRATE ANIMALS

N/A

13 SELECT AGENT RESEARCH

N/A

14 MULTIPLE PD/PI LEADERSHIP PLAN

N/A

15 CONSORTIUM/CONTRACTUAL ARRANGEMENTS

If the present application is funded, the University of North Carolina at Chapel Hill will execute subcontracts with the consortium institutions (Duke University and North Carolina State University). These inter-institutional agreements will be written consistent with the NIH consortium agreement policy.

16 LETTERS OF SUPPORT - None

17 RESOURCE SHARING PLAN(S)

(a) Data sharing plan: The data-related resources generated by the proposed research consists of new statistical methodology, software packages for implementation of the methodology, and tutorials for the software. The statistical methodology will be shared through peer reviewed publications and national meetings and through other standard means. All accepted publications will be deposited in PubMed Central in accordance with the NIH Public Access Policy. Summaries of the methodology, the software and tutorials will be shared through a public web site managed by Core A, while Core C will assist in preparation of the software and tutorials for dissemination. This project will use de-identified data prepared by Core B to test the methods and to create demonstrations of use of the methods to be included in tutorials. This project will not be involved in sharing of these data; this function will be addressed by Core B.

(b) Sharing model organisms: N/A

(c) GWAS: N/A

PROJECT 3

METHODS FOR POST MARKETING SURVEILLANCE AND COMPARATIVE EFFECTIVENES RESEARCH

Project Leader: Joseph G. Ibrahim, PhD

PROJECT SUMMARY (See instructions):

The Sentinel Initiative mandated by the Food and Drug Administration will lead to an enormous number of studies being planned post-market that will require analyzing and combining data from several different studies. The proposed project will address this challenge through developing new and flexible methods for meta-analysis using a variety of models, including models for binary and discrete data, models for longitudinal data, and models for time-to-event data. A related issue that will also be addressed is design, sample size, and power considerations using these types of meta-analytic models. Such models and data collected post-market can be quite useful in designing future clinical studies such as non-inferiority, equivalence, and superiority cancer clinical trials. The proposed project will also develop methods for meta-analytic studies of diagnostic tests to facilitate evidence-based medicine. We will also create flexible and robust methodology for accurately comparing rare adverse event rates in cancer for different drugs and for determining how those rates are affected by important prognostic factors. The proposed project will also explore statistical methods for the analysis of large cancer data sets for calibrating treatment dose in the presence of potentially conflicting factors, such as length and quality of life and economic costs. We will explore these tradeoffs rigorously, using a utility based approach traditionally employed in the analysis of health policy at the population level. The proposed statistical methodology will be broadly applicable to complex, large scale, data sets arising in phase III clinical trials and post-marketing studies.

RELEVANCE (See instructions):

The proposed statistical methodology will be broadly applicable to the statistical analysis and interpretation of complex, large scale, data sets arising in phase III clinical trials and post-marketing studies. The research will improve public health be facilitating discovery of important benefits and risks of cancer treatment.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location			
Organizational Name: The University of North Carolina at Chapel Hill			
DUNS: 608195277			
Street 1: Office of Sponsored Research, CB #1350	Street 2: 104 Airport Dr., Suite 2200		
City: Chapel Hill	County: Orange	State: NC	
Province:	Country: USA	Zip/Postal Code: 27599-1350	
Project/Performance Site Congressional Districts: NC-004			

Additional Project/Performance Site Location			
Organizational Name: North Carolina State University			
DUNS: 042092122			
Street 1: Research Admin/ SPARCS	Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514		
City: Raleigh	County: Wake	State: NC	
Province:	Country: USA	Zip/Postal Code: 27695-7514	
Project/Performance Site Congressional Districts: NC-02			

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Ibrahim, Joseph G.	JOE_IBRAHIM	UNC-CH	Project 3 Leader
Bondell, Howard D.		NC State University	Co-Investigator
Carpenter, William R.	Wcarpenter	UNC-CH	Co-Investigator
Chu, Haitao	hchu11	UNC-CH	Project Co-Leader
Fine, Jason P.	Jasonp3p	UNC-CH	Co-Investigator
Kosorok, Michael R.	Michael_Kosorok	UNC-CH	Project Co-Leader
Sandler, Robert S.	ROBERT_SANDLER	UNC-CH	Co-Investigator
Zhang, H. Helen		NC State University	Project Co-Leader

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project

Human Embryonic Stem Cells ☒ No ☐ Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

RESEARCH PLAN

1 INTRODUCTION TO RESUBMISSION/REVISION APPLICATION - N/A

2 SPECIFIC AIMS

The primary goal of this project is to develop, test, and evaluate new statistical methodology for Bayesian meta-analysis; design, sample size, and power considerations for future studies using meta-analytic models; meta-analysis of diagnostic tests; meta-analysis for regression analysis of rare adverse events; and for identifying optimal individualized therapies. To achieve this goal, we have developed five specific aims as follows:

Aim 1. Develop methodology for Bayesian meta analysis. We will develop novel Bayesian parametric and semiparametric models for meta-analysis for aggregated data, time to event data, discrete data, and longitudinal data. Specifically, we will consider: 1) Normal random effects models and develop novel Bayesian derivation of Q function for assessing heterogeneity across different studies for aggregated data; 2) Random effects generalized linear models for continuous or discrete data; 3) Mixed effects models for longitudinal data; and 4) Random effects Cox models with gamma process priors for time-to-event data. We will incorporate missing covariates and/or responses in all these models for various data types.

Aim 2. Develop methodology for Bayesian trial design using meta-analytic models. We will develop a new Bayesian approach of sample size determination (SSD) for the design of non-inferiority clinical trials using the novel meta-analytic models developed in Aim 1. First, we will extend the fitting and sampling priors of Wang and Gelfand (2002) to Bayesian SSD using meta-analytic models with a focus on controlling type I error, type II error, and power. Second, we will develop novel simulation-based Bayesian SSD using meta-analytic random effects generalized linear models, generalized linear mixed models, and random effects Cox models with gamma process priors.

Aim 3. Develop meta-analytic methodology of diagnostic tests without a gold standard. First, we will develop statistical methods for estimating accuracies of two and multiple (i.e., ≥ 3) diagnostic tests in a meta-analysis in the absence of a gold standard using maximum likelihood and full Bayesian methods. Second, we will reanalyze the meta-analysis data of 17 studies to evaluate the accuracy of microsatellite instability testing (MSI) and mutation analysis (Chen, Watson, and Parmigiani 2005), and a multi-center data set from NCI Colorectal Cancer Family Registry Study to evaluate the accuracy of 10 biomarkers in predicting Lynch syndrome and other data sets.

Aim 4. Develop methodology for regression analysis of rare adverse events for post-marketing safety evaluation. First, we will develop semi-parametric methods of inference for evaluating drug and risk factor effects for rare time-to-event outcomes in clinical trials and epidemiological studies. Second, we will develop semi-parametric methods of inference for extremely rare time-to-event outcomes. Third, we will extend both of the results to the adjudicated endpoint setting. Fourth, we will extend these results to the meta-analytic setting involving collections of clinical studies, registry data and health insurance claims data.

Aim 5. Develop methodology for identifying optimal individualized therapies from existing clinical trial data using meta-analysis, utility functions, classification and regression. We will develop a general inferential tool for determining optimal individualized therapies. First, we will propose an multi-attribute utility function for accommodating complex survival information, as well as cost and quality of life considerations. Second, we will develop rigorous inferential procedures for optimal dosing which will be broadly applicable to individualized therapies based on subject specific characteristics, including genomic as well as demographic and disease severity predictors. Third, we will utilize machine learning and other high dimensional statistical learning and regression techniques in addition to more traditional approaches to statistical modeling.

For each of the new methods developed for each aim, there will be four phases to the research: a methodological phase in which we construct new methods, create beta software for development, and establish universal properties based on statistical theory; a simulation phase were we verify validity of the theoretical predictions using highly controlled simulation studies; a data analysis stage where we evaluate performance using real data; and a software implementation and dissemination phase where we refine our software, test it in practical settings, and disseminate the software in a manner useful and accessible for practitioners.

3 BACKGROUND AND SIGNIFICANCE

3.1 Bayesian Meta-analysis

The recent FDA mandate through its Sentinel Initiative that all approved drugs must undergo post-marketing assessment of safety will lead to the collection and analysis of large datasets from multiple studies. Dr. Ibrahim has taken a great interest in this issue due to his direct involvement with the Sentinel Initiative for establishing guidelines for conducting such post-marketing studies. Assessment of safety of a post-market drug is a much more different and difficult problem to address than the assessment of safety in the course of a cancer clinical trial. In a clinical trial, a regimen leading to severe adverse events in 1% of the patients enrolled in the trial may be deemed acceptable whereas such a figure is generally not acceptable in the general population once a drug gets marketed. These types of post-market studies, many of which are now ongoing, will involve analysis of large datasets from multiple studies in order to accurately assess safety as well as efficacy of a given drug. There has been an enormous literature on frequentist methods using random effects models for meta-analysis for the past 20 years. Now, random effects modeling for meta-analysis has become a well accepted standard from both a frequentist and Bayesian perspective. See, for example, Whitehead and Whitehead (1991), Hardy and Thompson (1996), Higgins and Whitehead (1996), Biggerstaff and Tweedie (1997), Aitkin (1999), Normand (1999), Brockwell and Gordon (2001), Lopes, Müller, and Rosner (2003), Burr and Doss (2005), McLeod et al. (2007), and Sutton and Higgins (2008). However, frequentist methods based on random effects models, in general, cannot accommodate complex modeling schemes and the associated inference is often very difficult and unwieldy. For example, since it is not straightforward to obtain model-based estimate of heterogeneity between studies, exact confidence regions, and the null distributions of test statistics for tests of hypotheses in the context of generalized linear models, models for longitudinal data, and survival models. One must rely on asymptotic methods in this context. On the other hand, due to the nature of the random effects modeling in meta-analysis, Bayesian methods appear much better suited and more powerful inferential tools for handing meta-analysis from both a parametric and semiparametric perspective.

Although clinical trials are the gold standard for evaluating the safety and efficacy of cancer therapies, only less than 2% of patients with incident cancers enroll on NCI sponsored clinical trials. Furthermore, the fraction of trial enrollees is lower in racial/ethnic minority groups as well as older patients (Murthy, Krumholz and Gross 2004). Those limitations have created huge gaps for the generalization of clinical trial findings to general populations. To overcome those limitations, it is very important to develop comprehensive analysis methodology to pool different data resources together to maximize our understanding of hard-to-measure outcomes. Motivated by our collaborative work with Drs Carpenter, Goldberg and Sandler at UNC, using the integrative analysis of the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) Network, Accelerated Community Oncology Research Network Data Warehouse (ACORN), Cancer Care Outcomes and Research Consortium (CanCORS), the Surveillance, Epidemiology and End Results (SEER) and the Adjuvant Colon Cancer End Points (ACCENT) and other data sets, we will develop novel Bayesian parametric and semiparametric models for meta-analysis of continuous or discrete data, longitudinal data, and time to event data. We will also use the above-mentioned datasets in designing future clinical trials as discussed in Aim 2. There are several challenges in the meta-analysis using Bayesian methods compared to standard Bayesian methods in mixed models. These include (i) the development of general and flexible meta-analytic models that can account for (a) subject-to-subject heterogeneity and study-to-study heterogeneity, (b) aggregate and/or individual response and/or covariate data, and (c) missing data; (ii) the methods for assessing subject-to-subject heterogeneity and study-to-study heterogeneity; (iii) the properties such as the propriety of posteriors; and (iv) the computation. In the proposal, we will precisely address these important issues. In particular, we develop novel and general Bayesian meta-analytic models by introducing subject random-effects and treatment random-effects to account for subject-to-subject heterogeneity and study-to-study heterogeneity. We also develop new methods for assessing subject-to-subject heterogeneity and study-to-study heterogeneity and new meta-analytic models to account for aggregate and/or individual response and/or covariate data as well as missing data, and new semiparametric Bayesian meta-analytic models for time to event data. In addition, we will examine the properties such as the propriety of posteriors of the proposed meta-analytic models and develop efficient Markov chain Monte Carlo sampling algorithms to carry out challenging posterior computations under the complex parametric and semiparametric Bayesian meta-analytic models.

3.2 Bayesian Design of Trials Using Meta-analytic Models

Recently, FDA released "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials" (May 23, 2006, www.fda.gov/cdrh/meetings/072706-bayesian.html). This document provides guidance on statistical aspects of the design and analysis of Bayesian clinical trials for medical devices. It lays out detailed guidance on the determination of sample size in a Bayesian clinical trial. This document also provides guidance on the evaluation of the operating characteristics of a Bayesian design. Specifically, the evaluation of a Bayesian design should include type I error (probability of erroneously approving an ineffective or unsafe device), type II error (probability of erroneously disapproving a safe and effective device), and power (the converse of type II error: the probability of appropriately approving a safe and effective device).

In this aim, we are particularly interested in sample size determination (SSD) in non-inferiority therapeutic trials. Non-inferiority cancer clinical trials have become quite frequent and popular in recent years since designing superiority or equivalence trials for many regimens, especially in cancer, may not be realistic or practical. Thus, methods for design of non-inferiority trials are greatly lacking and there is a great need for new methodology in this area. There has been a vast literature on the frequentist methods of SSD in various non-inferiority trials, which includes, for example, D'Agostino Sr., Massaro, and Sullivan (2003), Hung et al (2003), Rothmann et al. (2003), Wang and Hung (2003a, b), Hung, Wang, and O'Neill (2005, 2007), Kieser and Friede (2007), and Fleming (2008). As discussed in Wang and Gelfand (2002), the limitations of the frequentist approach are evident. For instance, the frequentist methods require an estimate of the variability in the data or of the standard error of the parameter estimate in certain cases or models. In some other cases, one needs a value of the parameter vector in order to calculate the non-centrality parameter as a function of sample size.

The Bayesian approach in this context may be more attractive as it does not require an estimate of the variability or the specification of any nuisance parameters. The literature on Bayesian SSD has been growing due to recent advances in Bayesian computation and Markov chain Monte Carlo sampling in particular. The articles by Joseph, Wolfson, and Du Berger (1995a,b), Lindley (1997), Rubin and Stern (1998), Katsis and Toman (1999), and Inoue, Berry, and Parmigiani (2005) are the ones cited in the 2006 FDA Guidance Document. Some recent work includes Rahme and Joseph (1998), Simon (1999), Adcock (1997), Wang and Gelfand (2002), and M'Lan, Joseph, and Wolfson (2006, 2008). The existing literature on Bayesian SSD, however, primarily focuses on simple normal and binomial one or two sample problems and standard normal linear regression, and generalized linear models. Although the literature on Bayesian SSD discusses a variety of performance criteria, the widely used ones include Bayes factors (Weiss, 1997), the average posterior variance criterion (APVC) (see, for example, Wang and Gelfand, 2002), the average coverage criterion (ACC), the average length criterion (ALC), and the worst outcome criterion (WOC) (e.g., Joseph, Wolfson, and Du Berger, 1995a,b, and Joseph and Bélisle, 1997). Lindley (1997), Pham-Gia (1997), and Lam and Lam (1997) provide SSD through a maximization of expected utility or a minimization of the Bayes risk. M'Lan, Joseph, and Wolfson (2006, 2008) extend ACC and ALC to ACC_k and ALC_k. However, most of these criteria do not directly link to or do not control the type I error, type II error, and power, which are the most important operating characteristics of a Bayesian design specified in the 2006 FDA Guidance Document. In addition, none of the aforementioned Bayesian articles directly address the important yet practically most useful design and analysis of non-inferiority trials.

From both the frequentist and Bayesian perspectives, the literature on trial design using meta-analytic models is essentially non-existent. In addition, there is virtually no literature on addressing sample size and other trial design issues for random effects generalized linear models, generalized linear mixed models, and random effects survival models, such as the random effects Cox model with gamma process priors on the cumulative baseline hazard in the meta-analytic framework. In this aim, we wish to develop a general methodology for Bayesian trial design using meta-analytic models, with an emphasis on non-inferiority trials, since it is precisely in these types of trials that meta-analytic models can be most useful, efficient, and powerful. We then extend the fitting and sampling priors of Wang and Gelfand (2002) to Bayesian SSD in meta-analytic models with a focus on controlling the type I error, type II error, and power. We will develop novel simulation-based Bayesian SSD in meta-analytic models for random effects generalized linear models, generalized linear mixed models, and the random effects Cox model with gamma process priors. The use of meta-analytic models for trial design is quite new and in particular the use of these models for trial design in non-inferiority trials is quite innovative.

3.3 Meta-analytic Methodology of Diagnostic Tests without a Gold Standard

Accurate diagnosis of a disease status such as cancer is often the first step toward its control and prevention. The rapid growth of evidence-based medicine has led to a dramatic increase in attention to evidence-based diagnosis by meta-analysis of diagnostic test accuracy studies (Egger, Smith, and Altman 2001). The performance of a binary diagnostic test is usually represented by sensitivity and specificity (Zhou, Obuchowski, and McClish 2002; Pepe 2003). When a "gold standard" is available, random effects models including the hierarchical summary receiver operating characteristic model (Rutter and Gatsonis 2001) and the bivariate random effects model on sensitivities and specificities (van Houwelingen, Arends, and Stijnen 2002; Reitsma et al. 2005; Chu and Cole 2006), which are very closely related and sometimes identical (Harbord et al. 2007; Chu and Guo 2009), have been recommended to take into account the potential heterogeneity between studies (Zwinderman and Bossuyt 2008; Riley, Thompson, and Abrams 2008). In the absence of a gold standard reference test, there is a considerable literature discussing the challenges and approaches to assess the performance of diagnostic tests from a single or two population (Gart and Buck 1966; Hui and Walter 1980; Joseph, Gyorkos, and Coupal 1995; Andersen 1997; Johnson, Gastwirth, and Pearson 2001). Generally, the literature on meta-analytic studies of diagnostic test accuracies using random effects models to account for heterogeneity in the absence of a gold standard is very sparse. In a recent meta-analysis of seventeen studies to evaluate the accuracy of microsatellite instability testing (MSI) and mutation analysis (MUT) in predicting Lynch syndrome, the most common familial colorectal cancer syndrome, a Bayesian approach was proposed to handle missing data resulting from partial testing and lack of a gold standard (Chen, Watson, and Parmigiani 2005). However, the existing meta-analysis assumed that the sensitivities and specificities of both tests did not differ from study to study. Furthermore, after categorizing the studies into a registry-based recruitment group and a family-based recruitment group, the prevalence was assumed to be homogeneous within each group. Due to the differences in study design, study population, and laboratory techniques, between-study heterogeneity is intrinsic in almost all meta-analysis. Not adequately accounting for this heterogeneity when it is present may result in biased estimation and/or underestimation of uncertainty.

It is crucial for us to develop cutting edge statistical methods that can be directly applied to meta-analysis of diagnostic accuracy studies when a gold standard is not available. In this project, we will tackle some of the most important and most challenging issues. The methods to be developed will be used to reanalyze the meta-analysis data of 17 studies to evaluate the accuracy of microsatellite instability testing (MSI) and mutation analysis, and a multi-center data set from NCI Colorectal Cancer Family Registry Study to evaluate the accuracy of 10 biomarkers in predicting Lynch syndrome, and other data sets. The statistical methodology for estimating accuracies of diagnostic tests without a gold standard in a meta analysis setting is quite new and innovative.

3.4 Regression Analysis of Rare Adverse Events for Post-marketing Safety Evaluation

Proper rare adverse event analysis is crucial in both cancer treatment and cancer screening studies. This issue can potentially have enormous public health consequences if not addressed in an effective and timely manner. The most famous, recent example of this is the drug Vioxx which was used to treat inflammation and was pulled from the market in September, 2004, even though early indications of an increased risk of heart attack had emerged as early as 2001 (Mukherjee et al., 2001). While the risk of heart attack was low, when cumulated over a large number of patients, it had a dramatic public health impact. Cancer registries (see, e.g., Lucas, 1993), meta-analysis of clinical trials (Berlin and Colditz, 1999; Temple, 1999), and the FDA's active surveillance system MedWatch (Kessler, 1993) are important and key sources of data for monitoring for rare adverse events, and there are special added levels of oversight for pediatric drug surveillance (Smith et al., 2008). Rare events are also important in cancer screening studies, such as the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial conducted by the National Cancer Institute (Gohagan et al., 1994) which is still being actively evaluated (Ahn et al., 2008). We are also motivated by the need for property tools of analysis for data such as the ACORN and SEER data sets mentioned above in Section 2.1. Rare event analysis is crucial for many disease in addition to cancer, including cardiovascular disease, as highlighted by the Vioxx example, vaccine safety, and in many other health settings.

The key issue, however, is correct analysis and interpretation of rare event data. Without this, the best data possible will not improve human health. Indeed, some intellectuals believe that our current financial crisis is due to improper use of statistical methods for the analysis of rare event data in the financial markets: see the very provocative article by Taleb (2008). The issue is that rare events behave differently in a probabilistic sense than

other statistical quantities for which society has an intuitive feel for and that many non-statistical professionals were interpreting probabilities as certainties and not appropriately adjusting for the true uncertainty. Clearly, correct statistical methodology is crucially important for studying rare health events. What is different about rare events? The main issue is that the central limit theorem no longer holds for rare event statistics. Consider, for example, trying to estimate the probability of a rare event (or collection of rare events) p from the observed proportion $\hat{p}_n$ out of a large sample of size n. As long as the true probability is not too small, $\sqrt{n}(\hat{p}_n - p)$ will have an approximately normal distribution with mean zero and variance $p(1-p)$. For rare events, when p is close to zero as n gets large, this result is not longer true. We will classify the setting where p goes to zero but np goes to infinity, as n goes to infinity, as the "rare" event setting; we will classify the setting where np goes to a constant $0 < k_0 < \infty$ as the "extremely rare" setting; and we will classify the setting where np goes to zero as the "almost impossible" setting. We will not study the almost impossible event setting further , since it is unrealistic to expect to learn about such events since they may never be observed even in extremely large studies with extensive followup. For rare events, $\sqrt{n/\hat{p}_n}(\hat{p}_n - p_0)$ is approximately standard normal (even though the usual version of the central limit theorem does not hold); but for extremely rare events, $n\hat{p}_n$ goes to a Poisson random variable with rate k_0. The point of this is that there are fundamental differences in the statistical properties of rare event analysis from more standard analysis settings. Another complication for many rare event settings is the presence of events that need adjudication before being confirmed as actual primary events. Many late phase clinical trials and post-submission safety studies require the adjudication of primary event outcomes in order to ensure that events are classified correctly, with minimal bias, and consistently across clinical centers involved in the study and across the entire study period. Consider, for example, risk of endometrial and breast cancer in patients taking raloxifene to reduce postmenopausal osteoporosis as evaluated in a follow-up to the Multiple Outcomes of Raloxifene Evaluation (MORE) trial (Cummings et al., 1999; Cauley et al., 2001). Screening for endometrial cancer requires confirmation (adjudication) by an independent board of gynecologists based on complete clinical data. This issue can be even more complicated for cardiovascular endpoints for which the adjudication process can be very complex and involve several primary endpoints (see, e.g., Barter at al., 2007). At the time of data analysis, some endpoints have been adjudicated and some have not, and thus the factuality of some of the purported endpoints is uncertain.

The goal of the proposed research is to develop effective and valid methods of analysis for regression of risk factor and treatment effects on rare and extremely rare events for both adjudicated and un-adjudicated data. This includes developing flexible semi-parametric approaches for survival (time-to-event) data with rare and extremely rare rates of events. We also develop methods for both single post-marketing epidemiological studies as well as meta-analytic approaches. We note that logistic regression, while sharing some similarities with survival models in the rare event setting, is importantly different from survival analysis in several key ways, including the fundamental difference that patient follow-up times must be factored into survival analysis. For this reason, we will not discuss it further except for a few brief instances where it is useful for illustration.

Although some work has been done for rare analysis, there are many unresolved issues, especially for semi-parametric methods and adjudication. Exact conditional permutation approaches have been developed for both logistic regression (Hirji et al., 1987) and the log-rank test under equal (Mehta and Patel, 2000) and unequal (Heinze et al., 2003) follow-up. Unfortunately, these exact approaches can be very computationally intense, have a limited range of possible models, and are not suitable for sample size formula development. These concerns can largely be addressed through the development of large sample approximations we propose. Another method for addressing rare events in the design of studies is to oversample the rare event, as is done, for example, in case-control studies and other biased-sampled studies. However, this approach is not useful in monitoring for events yet to occur, as is the focus of this research, and so we will not pursue sample-based approaches further. Important large sample approximation work for rare events include the logistic regression methodology of King and Zeng (2000) which has been implemented in commercially available software. Other work on logistic regression for rare events includes Buzas and Stefanski (1996) and an evaluation by Cepeda et al. (2003). Methodology and theory for the semi-parametric log-rank test for rare survival events—as in the context of the PLCO cancer trial—was developed by Buyske et al. (2000), although their results do not apply to the extremely rare event setting. The use of the Cox model for rare events is studied in Li at al. (2007), but they did not derive any theoretical properties and asymptotic (large sample) approximations, and thus their results do not provide insight into solving the necessary inferential problems. Some results for survival analysis for rare events in a meta-analytic setting are developed by Shuster et al. (2007) that may be at least partly applicable to

our setting. Shuster et al. show that standard use of random effects to adjust for intra-study heterogeneity may be problematic for rare events. All of these results are applicable to the rare setting but not the extremely rare setting for which are we are not aware of any such work.

Cook and Kosorok (2004) developed rigorous semi-parametric survival analysis methodology for adjudicated endpoints, but their approach has not yet been extended to the rare event setting. Thus theoretically and scientifically valid methodology for semi-parametric methodology is very much underdeveloped, with the exception of the log-rank test for rare events, and the area is very much open to development. Moreover, no work has been done for extremely rare events or for adjudicated rare events. Thus the proposed research will fill a very important vacuum in the study of rare events, and the new statistical methods can lead to significant public health improvements in cancer and other diseases.

3.5 Identifying Optimal Individualized Therapies from Existing Clinical Trial Data Using Meta-Analysis, Utility Functions, Classification and Regression

In clinical trials, like those in oncology, it is common to report separate analyses of endpoints of interest. A typical example of such a study comes from a clinical trial conducted by the International Breast Cancer Study Group (IBCSG) Trial VI (IBCSG, 1996). This study was conducted in premenopausal women with node-positive breast cancer to explore both the duration of adjuvant chemotherapy and the reintroduction of delayed chemotherapy. The design was a 2 by 2 factorial, comparing three versus six cycles of oral cyclophosphamide, methotrexate, and flouracil with or without the introduction of three single courses of delayed chemotherapy. Each participant was randomly assigned to receive either the six initial courses of CMF at months 1 through 6 with or without reintroduction of CMF at months 9, 12, or 15, or to receive three initial courses of CMF at months 1 through 3 with or without reintroduction of CMF at months 6, 9, and 12. In addition to the survival endpoints, disease free survival (DFS) and overall survival (OS), there was considerable interest in the impact of the treatment regimens on toxicity outcomes and their influences on quality of life, which was assessed using a self-assessment questionnaire. In understanding these benefits in absolute terms, it is important to adjust for other patient specific factors, like age, number of positive nodes, and estrogen receptor (ER) status, which are tightly linked to the natural history of breast cancer. IBCSG VI is a large, international collaborative, with data from multiple continents. Randomization was stratified by participating institution, type of surgery, and ER status. To control for culture, country of residence may be used as a proxy, see Chen and Ibrahim (2006).

The study is prototypical, in that there are several endpoints of clinical interest. Each of these endpoints will impact the overall evaluation of treatment benefit and influence patient management. Typically, one would conduct separate regression analyses of each of the endpoints to understand the impact of patient specific predictors and treatment. In IBCSG VI, a proportional hazards model (Cox, 1972) would ordinarily be fit to DFS and OS. For repeated QOL measurements, longitudinal data models would be utilized, accounting for the potential missing QOL measurements due to dropout and or death (Zeger and Xu, 2001; Tsiatis and Davidian, 2004). In other studies where cost data is available, similar longitudinal regression models might be employed. In addition to the longitudinal analyses, one might consider regression analysis for summary measures of QOL and cost, quality adjusted lifetime (Glasziou, Simes, and Gelber, 1990; Fine and Gelber, 2001) and lifetime medical cost (Lin, Feuer, Etzioni, Wax, 1997; Lin, 2003), respectively. While such analyses are very useful in understanding the effects of treatment, disease severity factors, and other predictors, it may be difficult to synthesize these results in order to determine optimal treatment strategies for individual patients. The challenge is that it is not clear how to formalize the potential tradeoffs amongst the endpoints, e.g., a decrease in quality of life due to more aggressive treatment may be counterbalanced by better survival outcomes. Typically, such tradeoffs are evaluated informally, without a systematic quantitative evaluation.

The goal of this aim is to address the lack of inferential methodology to address these tradeoffs. To our knowledge, issues related to inference for optimal dose determination have not been considered in the setting considered here. Considerable work has been done on dose calibration using for example, linear regression, where the dose to achieve a minimally acceptable mean response is of interest, or logistic regression, where the dose necessary to achieve a particular response level is of interest; see Davidian (2002) for a review of this area. The work involves modeling a single response and multi-attribute utility functions does not need to accommodate tradeoffs between multiple outcomes.

The use of utility functions has been examined in the context of reinforcement learning, where optimal regimens may be determined either adaptively in clinical trials or using existing data. Recent work on dynamic

methods for adaptively and efficiently identifying such regimens have been studied, as surveyed in the seminal discussion paper of Murphy (2003). A limitation is that while such methods may be helpful in incorporating utility weights, formal inference may be difficult, with almost no results available. An assessment of uncertainty in estimates of the optimal regimes and an evaluation of the sensitivity of regime choice to changes in the utility function has not been considered. The current research will provide new methodology to fill these gaps.

4 PRELIMINARY STUDIES

4.1 Investigators

Dr. Ibrahim is Project Leader. Drs. Chu, Kosorok and Zhang are project co-Leaders; and Drs. Bondell, Carpenter, Fine and Sandler are project co-Investigators. Joseph G. Ibrahim, PhD, is Alumni Distinguished Professor of Biostatistics and Director of Biostatistics in the Lineberger Comprehensive Cancer Center at the University of North Carolina at Chapel Hill (UNC-CH). His research expertise is Bayesian inference, missing data problems, and cancer genomics. He has written two texts on Bayesian inference (Ibrahim et al., 2001; Chen et al., 2002). Haitao Chu, PhD, MD, is Research Associate Professor of Biostatistics at UNC-CH. His expertise is in clinical trials, statistical methods in epidemiology, and meta-analysis. Michael R. Kosorok, PhD, is Professor and Chair of Biostatistics at UNC-CH. His expertise is in clinical trials, survival analysis, semiparametric inference and empirical processes. He has written a text on the last two topics (Kosorok, 2008). Hao Helen Zhang, PhD, is Associate Professor of Statistics at North Carolina State University (NCSU). Her expertise is in model selection, machine learning and nonparametric smoothing methods. Howard D. Bondell, PhD, is Assistant Professor of Statistics at NCSU. His expertise is in variable and model selection and nonparametric and semiparametric statistical methods. William R. Carpenter, PhD, MHA, is Research Assistant Professor of Health Policy and Management and a Research Fellow at the Cecil G. Sheps Center for Health Services Research at UNC-CH. His expertise is in the analysis of large data sets for understanding public health aspects of cancer treatment. Jason P. Fine, PhD, is Professor of Biostatistics at UNC-CH. His expertise is in survival analysis, and nonparametric and semiparametric inference. Robert S. Sandler, MD, is chief of the Division of Gastroenterology and Hepatology and Nina C. and John T. Sessions Distinguished Professor of Medicine in the UNC School of Medicine. Dr. Sandler is an oncologist and has a wealth of expertise in cancer, especially in gastrointestinal (GI) cancer. Dr. Sandler has a joint publication with Drs. Chu and Ibrahim (Qu et al., 2008). Dr. Ibrahim also has a highly influential joint publication with Drs Goldberg and McLeod (Hoskins et al., 2007) on meta-analysis for the UGT1A1*28 allele (see Project 4).

4.2 Preliminary Studies

We have made some preliminary progress on developing statistical methods for diagnostic accuracy studies in meta-analysis using random effects models in the presence of a gold standard (Chu and Cole 2006; Chu and Guo 2009). To simplify the discussion, we focus on a bivariate random effects model without any covariates here. Let $n_{11}^i, n_{00}^i, n_{01}^i$ and n_{10}^i be the number of true positives, true negatives, false positives and false negatives, and $n_{1\cdot}^i$ and $n_{0\cdot}^i$ be the number of diseased and non-diseased patients in the i^{th} diagnostic accuracy studies from a meta-analysis, respectively. Conditional on the number of diseased and non-diseased patients in each study, the bivariate random-effects meta-analysis model assumes, $n_{00}^i \sim$ Bin $(n_{0\cdot}^i, Sp_i)$, $n_{11}^i \sim$ Bin $(n_{1\cdot}^i, Se_i)$, logit$(Se_i) = \mu_0 + \mu_i$, and logit$(Sp_i) = \nu_0 + \nu_i$ where logit$(w) = \log(w) - \log(1-w)$ and the random effects $(\mu_i, \nu_i)^T$ are bivariate normally distributed with mean zero. The expected sensitivity for a chosen specificity is given by logit$(Se) = a + b$logit(Sp) where $a = \mu_0 - \rho\nu_0\sigma_\mu/\sigma_\nu$ and $b = \rho\sigma_\mu/\sigma_\nu$. The median sensitivity for a given specificity on the back-transformed scale is given by expit$\{a + b\text{logit}(1-Sp)\}$ where expit$(w) = 1/(1+\exp(-w))$. The area under the summary operating characteristic (sROC) curve (AUC) based on the median sensitivity for a given specificity on the back-transformed scale can be estimated as $\int_0^1 \text{expit}\{a + b\text{logit}(1-Sp)\}\, dSp$.

We have made some preliminary progress on developing two new technical tools in empirical processes that pave the way for developing methods of inference for semi-parametric techniques for both rare and extremely rare time-to-event data. Let $N(t)$ be a counting process for the rare endpoint events in a sample under study and let $Y(t)$ be the corresponding at-risk process. This is the standard counting process notation for time-to-event data (Fleming and Harrington, 1991). We will let time t be on the scale of time since enrollment for each individual, i.e., $N(t)$ is the number of endpoints observed for all patients who have been enrolled for no more than time t. Also let $\lambda(t)$ denote the hazard for this endpoint (allowing it to become smaller and smaller

as the total sample size $n \to \infty$), and let $\Lambda(t) = \int_0^t \lambda(s)ds$. A good choice for an estimator of Λ is the well-established Nelson-Aalan estimator $\hat{\Lambda}(t) = \int_0^t [Y(s)]^{-1} dN(s)$. Provided we have no covariates, one can use the usual continuous-time martingale central limit to establish that $\left[\int_0^t \frac{dN(s)}{Y^2(s)}\right]^{-1/2} \left\{\int_0^t \frac{dN(s)}{Y(s)} - \Lambda(t)\right\} \rightsquigarrow N(0,1)$, where $\rightsquigarrow$ denotes convergence in distribution. This result can be extended to uniform in t weak convergence to a Gaussian process. This permits a large variety of possible inference approaches; see also Buyske et al. (2000) for developing log-rank tests for rare events.

The challenge with the martingale approach is that it does not directly apply to certain more complicated statistics, such Cox regression, especially under model misspecification, nor does it apply to other more complicated semi-parametric modeling settings. Fortunately, there is a very general central limit theorem for i.i.d. data that allows for the model to change with the sample size (a requirement, almost by definition, for rare events). There are several versions of this theorem, including Theorems 2.11.1 and 2.11.9 in van der Vaart and Wellner (1996) and also Theorems 11.16 and 11.18 in Kosorok (2008). The last theorem establishes validity of a modified bootstrap procedure that enables construction of confidence intervals and critical regions for hypothesis testing in this complex setting. The main idea is to view the statistics as sums of independent stochastic processes whose distributions are allowed to change with the sample size but whose index set does not change. Provided the complexity of this index set is reasonable, as measured by various kinds of entropy, such as uniform entropy, bracketing entropy, and "manageability," the standardized statistics can be shown to converge to a Gaussian limiting distribution. Our early studies show that this applies to the rare event setting.

A challenge with both the martingale and empirical process central limit theorems is that they are not applicable for extremely rare events. It turns out that some newer empirical process techniques have considerable promise here. We have discovered an unexpected connection between the behavior of $N(t)$ for extremely rare events and the behavior of an estimator of the change-point in change-point regression problems. Consider the very simple example where we observe a sample of random pairs (Y, Z), where $Y = \mu_1 + \epsilon$ if $Z \leq \xi$ and $Y = \mu_2 + \epsilon$ otherwise, and where $\mu_1 \neq \mu_2$, ξ is a real number, and ϵ is Gaussian with unknown variance σ^2. The parameters of this model are μ_1, μ_2, σ^2, and the change-point ξ. If least-squares estimate is used to estimate these parameters, the estimates of μ_1, μ_2 and σ^2 are all $\sqrt{n}$-consistent and asymptotically normal (as would be expected), but the estimate of ξ is n-consistent and converges to a compound Poisson process which we will describe in more detail later. It turns out that $N(t)$, after suitable standardization, also converges to a compound Poisson process. The reasons for this similarity are not obvious, but the similarity does emerge under

careful analysis using delicate empirical process arguments. We have also developed novel empirical process-based approaches for more complex change-point estimation and inference settings (see, for example, Kosorok and Song, 2007; and Song, Kosorok and Fine, 2008) that will be generally useful in developing semi-parametric methods of inference for extremely rare event data. These two powerful empirical process tools, the flexible central limit theorem and the compound Poisson process methodology, in combination with more classical statistical tools such as martingale methods, will enable us to make dramatic progress on this challenging research aim. Part of the reason that more progress has not been previously made on rare events is that the needed technical tools, and the connection with the change-point problem, were not available until recently.

5 RESEARCH DESIGN AND METHODS

5.1 Bayesian Meta-analysis

Bayesian justification of the heterogeneity measure Q. Suppose we consider K studies or random trials and let y_k denote the aggregate response with a known precision parameter w_k. Then, the standard normal random effects model for the meta-analysis assumes

$$y_k = \mu_k + \epsilon_k, \ \epsilon_k \sim N(0, w_k^{-1}) \quad (5.1)$$

$$\mu_k = \mu + \xi_k, \ \xi_k \sim N(0, \tau^2) \quad (5.2)$$

for $k = 1, 2, \ldots, K$. In (5.1), y_k may be viewed as the estimated effect size and μ_k is the true effect size. For example, in a collection of n two-arm randomized clinical trials, y_k is simply the observed log-odds ratio and w_k is the corresponding estimated precision of the observed log-odds ratio for the k^{th} study/trial (Higgins and Whitehead, 1996; Brockwell and Gordon, 2001). In the frequentist literature, a formal test of statistical homogeneity is performed using the test statistic proposed by Cochran (1954) given by

$$Q = \sum_{k=1}^{K} w_k (y_k - \hat{\mu})^2, \tag{5.3}$$

where $\hat{\mu} = \sum_{k=1}^{K} w_k y_k / (\sum_{k=1}^{K} w_k)$. Under the hypothesis of homogeneity ($\tau^2 = 0$), Q follows a χ^2_{K-1} distribution (Cochran, 1954). Although Q is the most popular test statistic for assessing heterogeneity across K studies, there is virtually no literature on a Bayesian justification of Q.

We will provide two Bayesian justifications of Q. Our first justification is based on the Deviance Information Criterion (DIC). Under the model specified by (5.1), the likelihood function is given by

$$f(\boldsymbol{y}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \frac{w_k^{1/2}}{\sqrt{2\pi}} \exp\{-\frac{w_k}{2}(y_k - \mu_k)^2\} = \Big[\prod_{k=1}^{K} \frac{w_k^{1/2}}{\sqrt{2\pi}}\Big] \exp\Big\{-\frac{1}{2}\sum_{k=1}^{K} w_k (y_k - \mu_k)^2\Big\}, \tag{5.4}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)'$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)'$. According to Spiegelhalter et al. (2002), we define the deviance function as $D(\boldsymbol{\mu}) = -2\log f(\boldsymbol{y}|\boldsymbol{\mu}) + 2\log f(\boldsymbol{y}|\boldsymbol{\mu}^*(\boldsymbol{y}))$, where $\boldsymbol{\mu}^*(\boldsymbol{y})$ is an estimate of the pseudotrue parameter $\boldsymbol{\mu}^t$. Let H_0 denote the hypothesis of homogeneity and D_{obs} be the observed data. Then, we define the *Bayesian Q function* as

$$Q_B = D(\tilde{\boldsymbol{\mu}}, H_0), \tag{5.5}$$

where $\tilde{\boldsymbol{\mu}}$ is either the posterior mode or the posterior mean of $\boldsymbol{\mu}$ under H_0, denoted by $E[\boldsymbol{\mu}|D_{obs}, H_0]$. For the normal model (5.1), $\boldsymbol{\mu}^*(\boldsymbol{y}) = \boldsymbol{y}$ and $f(\boldsymbol{y}|\boldsymbol{\mu}^*(\boldsymbol{y})) = \prod_{k=1}^{K} \sqrt{w_k/(2\pi)}$. Suppose we take an improper uniform prior for μ, i.e., $\pi(\mu) \propto 1$. Then, we have $E[\mu|D_{obs}, H_0] = \hat{\mu}$ and

$$Q_B = -2\log\Big[\prod_{k=1}^{K} \frac{w_k^{1/2}}{\sqrt{2\pi}}\Big] + \sum_{k=1}^{K} w_k (y_k - \hat{\mu})^2 + 2\log\Big[\prod_{k=1}^{K} \frac{w_k^{1/2}}{\sqrt{2\pi}}\Big] = Q.$$

Our second justification is based on the Bayes factor. Instead of (5.2), we assume

$$\mu_k = \mu + \xi_k, \;\; \xi_k \sim N(0, \tau^2/w_k). \tag{5.6}$$

Let H_1 denote the hypothesis of heterogeneity ($\tau^2 > 0$). We specify an inverse gamma prior $G(a, b)$ for τ^2 with density $\pi(\tau^2) \propto (\tau^2)^{-(a+1)} \exp(-b/\tau^2)$, where $a > 0$ and $b > 0$. Because both hypotheses share a common parameter μ, without loss of generality, we take $\pi(\mu) \propto 1$. After some algebra, it can be shown that the marginal likelihood under the hypothesis H_0 is given by

$$m_0 = \Big[\prod_{k=1}^{K} \frac{w_k^{1/2}}{\sqrt{2\pi}}\Big] \int \exp\Big\{-\frac{1}{2}\sum_{k=1}^{K} w_k (y_k - \mu)^2\Big\} d\mu = \Big[\prod_{k=1}^{K} \frac{w_k^{1/2}}{\sqrt{2\pi}}\Big] \exp\Big\{-\frac{1}{2}Q\Big\} \sqrt{2\pi} \Big(\sum_{k=1}^{K} w_k\Big)^{-1/2}.$$

Similarly, under the hypothesis H_1, we obtain the marginal likelihood as follows:

$$m_1 = \Big[\prod_{k=1}^{K} \frac{w_k^{1/2}}{\sqrt{2\pi}}\Big] \sqrt{2\pi} \Big(\sum_{k=1}^{K} w_k\Big)^{-1/2} \int_0^{\infty} \Big(\frac{1}{\sqrt{1+\tau^2}}\Big)^{K-1} \exp\Big\{-\frac{Q}{2(1+\tau^2)}\Big\} \frac{b^a}{\Gamma(a)(\tau^2)^{a+1}} \exp(-\frac{b}{\tau^2}) d\tau^2.$$

Thus, the Bayes factor for comparing H_1 to H_0 is given as follows:

$$B_{10} = \frac{m_1}{m_0} = \int_0^{\infty} \Big(\frac{1}{\sqrt{1+\tau^2}}\Big)^{K-1} \exp\Big\{\frac{\tau^2}{2(1+\tau^2)} Q\Big\} \frac{b^a}{\Gamma(a)(\tau^2)^{a+1}} \exp(-\frac{b}{\tau^2}) d\tau^2. \tag{5.7}$$

It is easy to see that B_{10} is an increasing function of Q. This is intuitively appealing as the larger Q, the more evidence in the data supporting the alternative hypothesis H_1. When (5.2) is assumed for the random effects instead of (5.6), an exact connection between the Bayes factor and the Q function is not available. However, in this case, we can show that under the same prior specification for μ and τ^2, the marginal likelihood under hypothesis

of heterogeneity (H_1) is given by $m_1^* = \left[\prod_{k=1}^K \frac{w_k^{1/2}}{\sqrt{2\pi}}\right]\sqrt{2\pi}\int_0^\infty \left[\left(\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}\right)^{-1/2}\left(\prod_{k=1}^K \frac{1}{\sqrt{1+w_k\tau^2}}\right) \times \exp\left\{-\frac{1}{2}\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}(y_k-\hat{\mu}(\tau^2))^2\right\}\frac{b^a}{\Gamma(a)(\tau^2)^{a+1}}\exp(-\frac{b}{\tau^2})\right]d\tau^2$, where $\hat{\mu}(\tau^2) = \left(\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}\right)^{-1}\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}y_k$. After some algebra, it can be shown that the corresponding Bayes factor is given by

$$B_{10}^* = \frac{m_1^*}{m_0} = \int_0^\infty \Big[\Big(\frac{\sum_{k=1}^K w_k}{\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}}\Big)^{1/2}\exp\Big\{\frac{1}{2}\sum_{k=1}^K w_k\frac{w_k\tau^2}{1+w_k\tau^2}(y_k-\hat{\mu})^2 + \frac{1}{2}\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}(\hat{\mu}-\hat{\mu}(\tau^2))^2\Big\} \times \frac{b^a}{\Gamma(a)(\tau^2)^{a+1}}\exp(-\frac{b}{\tau^2})\Big]d\tau^2. \quad (5.8)$$

Using (5.8), we obtain that the upper bound of B_{10}^* is

$$\int_0^\infty \Big(\frac{\sum_{k=1}^K w_k}{\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}}\Big)^{1/2}\exp\Big\{\frac{1}{2}Q + \frac{1}{2}\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}(\hat{\mu}-\hat{\mu}(\tau^2))^2\Big\}\frac{b^a}{\Gamma(a)(\tau^2)^{a+1}}\exp(-\frac{b}{\tau^2})d\tau^2 \quad (5.9)$$

and the lower bound of B_{10}^* is

$$\int_0^\infty \Big(\frac{\sum_{k=1}^K w_k}{\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}}\Big)^{1/2}\exp\Big\{\frac{w_{min}}{2(1+w_{min}\tau^2)}Q + \frac{1}{2}\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}(\hat{\mu}-\hat{\mu}(\tau^2))^2\Big\}\frac{b^a}{\Gamma(a)(\tau^2)^{a+1}}\exp(-\frac{b}{\tau^2})d\tau^2, \quad (5.10)$$

where $w_{min} = \min\{w_k,\ k = 1,2,\ldots,K\}$. Due to the two bounds of the Bayes factor B_{10}^* given by (5.9) and (5.10), we expect that B_{10}^* increases as Q increases. In the proposed grant, we will empirically examine this property via several simulation studies and examine further theoretical connections between B_{10}^* and Q for the normal model and a variety of other models, including meta-analytic models based on generalized linear models, models for longitudinal data and survival models. The Bayesian justification of Q is important and significant as it sheds light on the use of Bayesian model criteria such as DIC or Bayes factor for assessing heterogeneity across K studies without resorting the asymptotic distribution of Q and the extension of the Q function to more complex models such as meta normal mixed-effects models below.

Bayesian meta-analysis for multiple aggregate responses data per study. We consider meta-analysis models that accommodate q aggregate responses, p aggregate covariates of interest that we model as fixed-effects, and r aggregate covariates that we model as random-effects across K studies. The q aggregate responses typically would correspond to q treatments for example. Let y_{jk} denote the aggregate response with a known precision parameter w_{jk}, $\boldsymbol{x}_{jk}$ denote a p-dimensional vector of aggregate fixed-effect covariates, and $\boldsymbol{z}_{jk}$ denote a r-dimensional vector of aggregate random-effect covariates. The meta-analysis model assumes that for $k = 1,\ldots,K$ and $j = 1,\ldots,q$,

$$y_{jk} = \boldsymbol{x}_{jk}'\boldsymbol{\beta} + \boldsymbol{z}_{jk}'\boldsymbol{\gamma}_k + \mu_{jk} + \epsilon_{jk},\ \epsilon_{jk} \sim N(0, w_{jk}^{-1}), \quad (5.11)$$

independently, where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$ is the vector of regression coefficients corresponding to p aggregate fixed-effects covariates, $\boldsymbol{\gamma}_k = (\gamma_{k1},\ldots,\gamma_{kr})'$ is the vector of random regression coefficients corresponding to r aggregate random-effects covariates, and

$$\boldsymbol{\mu}_k = (\mu_{1k},\ldots,\mu_{qk})' = \boldsymbol{\mu} + \boldsymbol{\xi}_k,\ \boldsymbol{\xi}_k = (\xi_{1k},\ldots,\xi_{qk})' \sim N_q(0,V),\ \text{and}\ \boldsymbol{\gamma}_k \sim N_r(0,\Gamma), \quad (5.12)$$

where $\boldsymbol{\mu} = (\mu_1,\mu_2,\ldots,\mu_q)'$, V is a $q \times q$ covariance matrix, Γ is a $r \times r$ covariance matrix, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\gamma}_k$ are independent. In (5.11), the random-effects covariates $\boldsymbol{z}_{jk}$ may be useful in correcting potential study bias. In (5.2), when $V = 0$ and $\Gamma = 0$, then $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_K = \boldsymbol{\mu}$ and $\boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_K = 0$, indicating that there is no heterogeneity across K studies. As $\boldsymbol{\mu}_k$ in (5.12) is a q-dimensional vector of random effects, the classical definition of Q given in (refQstat) is not well defined. However, the Bayesian Q_B defined in (5.5) can be easily extended to this case. Specifically, assuming $\pi(\boldsymbol{\mu}) \propto 1$, after some algebra, we can show that $Q_B = \sum_{k=1}^K (\boldsymbol{y}_k - \hat{\boldsymbol{\theta}})'W_k(\boldsymbol{y}_k - \hat{\boldsymbol{\theta}})$, where $\boldsymbol{y}_k = (y_{1k},\ldots,y_{qk})'$, $W_k = \text{diag}(w_{1k},\ldots,w_{qk})$, $\hat{\boldsymbol{\theta}} = \left(\sum_{k=1}^K X_k^{*\prime}W_kX_k^*\right)^{-1}\sum_{k=1}^K X_k^{*\prime}W_k\boldsymbol{y}_k$,

$X_k^* = (X_k, I_q)$ is a $q \times (p+q)$ matrix, $X_k = (\boldsymbol{x}_{1k}, \ldots, \boldsymbol{x}_{qk})'$ is a $q \times p$ matrix, and I_q is the $q \times q$ identity matrix. Similar to Cochran (1954), we can show that under the hypothesis of homogeneity, i.e., $V = 0$ and $\Gamma = 0$, $Q_B \sim \chi^2_{qK-q-p}$ given that $p \ll q(K-1)$. To carry out Bayesian analysis of the meta-analysis model specified by (5.11) and (5.12), we may further take independent inverse Wishart priors for V and Γ. In the proposed research here, an efficient Gibbs sampling algorithm will be developed for carrying out posterior computations, and the theoretical connections between Q_B and the Bayes factor will also be examined in detail.

Bayesian meta-analysis via random effects generalized linear models for continuous or discrete data. To accommodate subject level continuous or discrete responses and subject level or aggregate covariates, we propose random effects generalized linear models (REGLM) for meta-analysis. Suppose that there are p covariates of interest that we model as fixed effects, r covariates of interest that we model as random effects, and q random treatment effects across K studies. Let y_{ijk} denote the subject level response, let the fixed-effects covariates for subject i be denoted by $\boldsymbol{x}_{ijk} = (x_{ijk1}, \ldots, x_{ijkp})'$ with corresponding regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, and let the random-effects covariates for subject i be denoted by $\boldsymbol{z}_{ijk} = (z_{ijk1}, \ldots, z_{ijkr})'$ with corresponding regression coefficients $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kr})'$ for $i = 1, 2, \ldots, n_{jk}$, $j = 1, 2, \ldots, q$, and $k = 1, 2, \ldots, K$. As discussed in Aitkin (1999), we let $\boldsymbol{x}_{ijk} = \boldsymbol{x}_{jk}$ for all i for aggregate covariates. We propose a REGLM for meta-analysis as follows:

$$f(y_{ijk}|\boldsymbol{x}_{ijk}, \boldsymbol{\beta}, \boldsymbol{\gamma}_k, \mu_{jk}, \sigma) = \exp\left\{a_{ijk}^{-1}(\sigma)(y_{ijk}\theta_{ijk} - b(\theta_{ijk})) + c(y_{ijk}, \sigma)\right\}, \tag{5.13}$$

where $\theta_{ijk} = \theta(\eta_{ijk})$ is the canonical parameter, $\eta_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{z}'_{ijk}\boldsymbol{\gamma}_k + \mu_{jk}$, and σ is a dispersion parameter. The functions a, b and c determine a particular family in the class. The functions $a_{ijk}(\sigma)$ are commonly of the form $a_{ijk}(\sigma) = \sigma^{-1}\omega_{ijk}^{-1}$, where the ω_{ijk}'s are known weights. When $\theta_{ijk} = \eta_{ijk}$, the link is said to be canonical. Furthermore, (5.12) is assumed for μ_{jk} and $\boldsymbol{\gamma}_k$. The REGLMs are quite general, which include the normal linear regression, logistic and Poisson regression with random effects.

Under the REGLM, we allow for missing data in either responses or covariates. We assume that missing response or covariates are nonignorably missing. Let $f(\boldsymbol{x}_{ijk}, \boldsymbol{z}_{ijk} \mid \boldsymbol{\alpha})$ denote the joint distribution of the covariates. The missing data mechanism is defined as the distribution of the $(p+1) \times 1$ random vector $\boldsymbol{d}_{ijk} = (d_{ijk0}, d_{ijk1}, d_{ijk2}, \ldots, d_{ijkp}, d_{ijk,p+1}, \ldots, d_{ijk,p+r})'$, where $d_{ijk0} = 0$ if y_{ijk} is missing and $d_{ijk0} = 1$ if y_{ijk} is observed, $d_{ijkl} = 0$ when x_{ijkl} is missing and $d_{ijkl} = 1$ when x_{ijkl} is observed for $l = 1, 2, \cdots, p$, and $d_{ijkl} = 0$ when z_{ijkl} is missing and $d_{ijkl} = 1$ when z_{ijkl} is observed for $l = p+1, 2, \cdots, p+r$,. The joint distribution of $\boldsymbol{d}_{ijk}$ is written as $f(\boldsymbol{d}_{ijk}|y_{ijk}, \boldsymbol{x}_{ijk}, \boldsymbol{z}_{ijk}, \boldsymbol{\phi})$. We assume that $\boldsymbol{x}_{ijk}$, $\boldsymbol{z}_{ijk}$ and r_{ijk} are independent of the random effects μ_k and $\boldsymbol{\gamma}_k$. Note that when some components of $\boldsymbol{z}_{ijk}$ are identical to those in $\boldsymbol{x}_{ijk}$, the dimension of $\boldsymbol{d}_{ijk}$ will be reduced. We model both distributions by a sequence of one-dimensional conditional distributions proposed by Lipsitz and Ibrahim (1996), Ibrahim, Lipsitz and Chen (1999), Ibrahim, Chen, and Lipsitz (2001), Huang, Chen, and Ibrahim (2005), and Chen et al. (2008). Since $\boldsymbol{d}_{ijk}$ is a vector of binary missing indicators, we use a binary regression model with probit for each one-dimensional conditional distribution in $f(\boldsymbol{d}_{ijk}|y_{ijk}, \boldsymbol{x}_{ijk}, \boldsymbol{z}_{ijk}, \boldsymbol{\phi})$. Due to the weak identifiability of the $\boldsymbol{d}_{ijk}$ model, the collapsing technique proposed by Huang, Chen, and Ibrahim (2005) will be used in the development of an efficient MCMC sampling algorithm.

Bayesian meta-analysis via generalized linear mixed models for longitudinal data. We extend the REGLM given by (5.13) to the generalized linear mixed model (REGLMM) with random effects to account for longitudinal data in meta-analysis. To this end, at time t, we let y_{tijk} denote the subject level response at time t, let the fixed effects covariates for subject i be denoted by $\boldsymbol{x}_{tijk} = (x_{tijk1}, \ldots, x_{tijkp})'$ with corresponding regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ for $i = 1, 2, \ldots, n_{jk}$, and let the random effects covariates for subject i be denoted by $\boldsymbol{z}_{tijk} = (z_{tijk1}, \ldots, z_{tijkr})'$ with corresponding random regression coefficients $\boldsymbol{\gamma}_{ik} = (\gamma_{ik1}, \ldots, \gamma_{ikr})'$, for $t = 1, \ldots, T_{ijk}$, $i = 1, 2, \ldots, n_{jk}$, $j = 1, 2, \ldots, q$, and $k = 1, 2, \ldots, K$. Following Ibrahim, Chen, and Lipsitz (2001), we propose the following GLMREM for longitudinal data:

$$f(y_{tijk}|\boldsymbol{x}_{tijk}, \boldsymbol{z}_{tijk}, \boldsymbol{\gamma}_{ik}, \boldsymbol{\beta}, \mu_{jk1}, \mu_{jk2}, \sigma) = \exp\left\{a_{tijk}^{-1}(\sigma)(y_{tijk}\theta_{tijk} - b(\theta_{tijk})) + c(y_{tijk}, \sigma)\right\}, \tag{5.14}$$

where $\theta_{tijk} = \theta(\eta_{tijk})$ is the canonical parameter,

$$\eta_{tijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{z}'_{tijk}\boldsymbol{\gamma}_{ik} + \mu_{jk1} + \boldsymbol{\mu}'_{jk2}\boldsymbol{g}(t), \tag{5.15}$$

$g(t)$ is a q^*-dimensional vector of known functions of t, and σ is a dispersion parameter. Let $\boldsymbol{\mu}_k = (\mu_{1k1}, \ldots, \mu_{qk1}, \boldsymbol{\mu}'_{1k2}, \ldots, \boldsymbol{\mu}'_{qk2})'$. We assume that $\boldsymbol{\gamma}_{ik}$ and $\boldsymbol{\mu}_k$ are independent, $\boldsymbol{\gamma}_{ik} \sim N_r(0, \Gamma)$ and $\boldsymbol{\mu}_k \sim N_{q(1+q^*)}(\boldsymbol{\mu}, V)$, where Γ and V are unknown covariance matrices. In (5.15), μ_{jk1} and $\boldsymbol{\mu}_{jk2}$ capture the main treatment effects and the treatment and time interaction effects. The proposed GLMREM is more general than the one proposed by Lopes, Müller, and Rosner (2003). In the proposed research, we will carefully examine model identifiability and develop an efficient computational algorithm. In addition, we will also allow for nonignorably missing responses and/or covariates.

Bayesian meta-analysis for semiparametric models for survival data with gamma process priors. Let y_{ijk} denote the subject level time-to-event (failure time) and let ν_{ijk} denote the censoring indicator such that $\nu_{ijk} = 0$ if y_{ijk} is a failure time and $\nu_{ijk} = 1$ if y_{ijk} is right censored. Also, let $\mathcal{R}(t) = \{(i,j,k) : y_{ijk} \geq t\}$ denote the set of subjects at risk at time t. We develop semiparametric models for the y_{ijk} with gamma process priors for meta-analysis. Assume the Cox proportional hazard regression model (Cox, 1972, 1975) for y_{ijk} with $h_0(y)$. Let $\eta_{ijk} = \exp(\boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{z}'_{ijk}\boldsymbol{\gamma}_k + \mu_{jk})$, then the full likelihood function is given by

$$\Big[\prod_{k=1}^{K}\prod_{j=1}^{q}\prod_{i=1}^{n_{jk}} \big(h_0(y_{ijk})\eta_{ijk}\big)^{\nu_{ijk}}\Big] \exp\Big\{-\sum_{k=1}^{K}\sum_{j=1}^{q}\sum_{i=1}^{n_{jk}} H_0(y_{ijk})\eta_{ijk}\Big\}, \tag{5.16}$$

where $D_{obs} = \{(y_{ijk}, \nu_{ijk}, \boldsymbol{x}_{ijk}, \boldsymbol{z}_{ijk}),\ i = 1, 2, \ldots, n_{jk}, j = 1, \ldots, q, k = 1, \ldots, K\}$, and $H_0(y) = \int_0^y h_0(u)du$ is the cumulative hazard function. Then, (5.12) is assumed for the random effects μ_{jk} and $\boldsymbol{\gamma}_k$.

Let $L_p(\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K | D_{obs})$ denote the partial likelihood. Sargent et al. (2000) treat the partial likelihood $L_p(\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K | D_{obs})$ as the "likelihood" and then carry out Bayesian inference for the random-effect survival analysis. In this proposal, we carry out Bayesian inference based on the full likelihood given by (5.16) via a gamma process prior for $H_0(y)$. Kalbfleisch (1978) and Sinha, Ibrahim, and Chen (2003) show that the partial likelihood can be obtained as a limiting case of the marginal posterior of $(\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$ with continuous time survival data under a gamma process prior for $H_0(y)$ using the likelihood function (5.16). Thus, the full likelihood based approach is more general than the one based on the partial likelihood of Sargent et al. (2000). Assume the baseline cumulative hazard function $H_0(y) \sim \mathcal{GP}(H^*, c_0)$, where $\mathcal{GP}(\cdot, \cdot)$ denotes a gamma process, $H^*(y)$ is a known increasing differentiable function and $c_0 > 0$. Let $y_{(1)} < y_{(2)} < \cdots < y_{(n^*)}$ be the n^* distinct failure and censoring times of the y_{ijk}'s. Write $y_{(l)} = y_{ijk_l}$ and ν_{ijk_l} denotes the corresponding censoring indicator for $l = 1, 2, \ldots, n^*$. Following Chen, Ibrahim, and Shao (2006), we can show that after integrating out $H_0(y)$, (5.16) reduces to

$$\prod_{l=1}^{n^*}\left\{\exp\Big[c_0 H^*(y_{(l)})\log\Big(\frac{c_0 + A_{l+1}}{c_0 + A_l}\Big)\Big] \prod_{(i,j,k)\in\mathcal{D}(y_{(l)})}\left[-c_0 h^*(y_{(j)})\log\left(1 - \frac{\exp(\boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{z}'_{ijk}\boldsymbol{\gamma}_k + \mu_{jk})}{c_0 + A_l}\right)\right]^{\nu_{ijk_l}}\right\}, \tag{5.17}$$

where $\mathcal{D}(y_{(l)}) = \{l : \ y_{ijk} = y_{(l)}, \nu_{ijk} = 1\}$ (i.e., the failure set at $y_{(l)}$), $A_l = \sum_{(i,j,k)\in\mathcal{R}(y_{(l)})} \exp(\boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{z}'_{ijk}\boldsymbol{\gamma}_k + \mu_{jk})$, and $h^*(y) = \frac{dH^*(y)}{dy}$. It can be shown that $\lim_{c_0\to 0} K(c_0)$, $L(\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K | D_{obs}, c_0) = L_p(\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K | D_{obs})$, where $K(c_0)$ is a parameter-free function of c_0, which is the partial likelihood used in Sargent et al. (2000) for the meta-analysis of survival data.

We specify an improper uniform prior for $\boldsymbol{\beta}$ (the coefficients for the fixed-effects covariates) and $\boldsymbol{\mu}$ (the overall mean of the random-effects $\boldsymbol{\mu}_k$) and inverse Wishart priors for V and Γ (the covariance matrices of random-effects $\boldsymbol{\mu}_k$ and $\boldsymbol{\gamma}_k$). In the proposed research, the conditions for the propriety of the resulting posterior will be established, and an efficient Markov chain Monte Carlo (MCMC) algorithm via the introduction of several sets of latent variables will be developed by extending the novel MCMC algorithm proposed by Chen, Ibrahim, and Shao (2006) to the semiparametric survival model with random-effects. In (5.17), we also allow missing covariates. Furthermore, we will also develop a more general semiparametric survival model to allow the baseline hazard function to vary across different studies. Specifically, instead of (5.16), we assume $L(\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, h_{01}, \ldots, h_{0K} | D_{obs}) = \Big[\prod_{k=1}^{K}\prod_{j=1}^{q}\prod_{i=1}^{n_{jk}} (h_{0k}(y_{ijk})\exp(\boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{z}'_{ijk}\boldsymbol{\gamma}_k + \mu_{jk}))^{\nu_{ijk}}\Big]$ $\exp\Big\{-\sum_{k=1}^{K}\sum_{j=1}^{q}\sum_{i=1}^{n_{jk}} H_{0k}(y_{ijk})\exp(\boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{z}'_{ijk}\boldsymbol{\gamma}_k + \mu_{jk})\Big\}$, where $H_{0k}(y) = \int_0^y h_{0k}(u)du$ is the cumulative

hazard function. Then, we assume $H_{0k}(y) \sim \mathcal{GP}(H^*, c_0)$ independently. In this case, H^* is allowed to depend on certain unknown parameters.

Nonparametric models for random-effects in meta-analysis. In (5.2) and (5.12), we assume parametric normal models for random-effects μ_k and γ_k. The parametric assumption can be removed by assuming a mixture of Dirichlet process (MDP) for the random effects. Burr and Doss (2005) develop a version of MDP for random-effects in the context of meta-analysis. Following Kleinman and Ibrahim (1998a, 1998b), we assume $\mu_k - \mu \sim G$ and $\gamma_k \sim F$, where G and F are general distribution functions and then assume $G \sim DP(M_G \cdot N_q(0, V))$ and $F \sim DP(M_F \cdot N_r(0, \Gamma))$, where V and Γ are unknown covariance matrices, and M_G and M_F are two positive scalars, which reflect a prior belief about how similar the nonparametric distributions G and F are to the base measures $N_q(0, V)$ and $N_r(0, \Gamma)$. In the proposed research, computational algorithms will be developed to carry out posterior computations and the properties of the posterior with a MDP prior will be examined in detail. All of the proposed models above will be applied using datasets from ACORN, CanCORS, SEER, and ACCENT, as mentioned in Section 2.1.

5.2 Bayesian Trial Design Using Meta-analytic Models

Bayesian design of non-inferiority trials. We first develop a new but general method to determine Bayesian sample size for a non-inferiority trial. Denote the data associated with a sample size of n by $\boldsymbol{y}^{(n)}$ and let $\boldsymbol{\theta}$ be the vector of all model parameter. Then, the joint distribution of $\boldsymbol{y}^{(n)}$ and $\boldsymbol{\theta}$ is written as $f(\boldsymbol{y}^{(n)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ denotes the prior distribution. Let $h(\boldsymbol{\theta})$ is a scalar function that measures the "true" size of the treatment effect. Let δ denote the non-inferiority margin. Similarly to Hung et al. (2003), we assume that the hypotheses for non-inferiority testing can be formulated as follows: H_0: $g(\boldsymbol{\theta}) \geq \delta$ versus H_1: $g(\boldsymbol{\theta}) < \delta$, where $g(\boldsymbol{\theta})$ is a known scalar function of $\boldsymbol{\theta}$. Consequently, we let Θ_0 and Θ_1 denote the parameter spaces corresponding to H_0 and H_1. In terms of non-inferiority, H_1 defines a successful trial.

Following Wang and Gelfand (2002), let $\pi^{(f)}(\boldsymbol{\theta})$ denote the fitting prior and $\pi^{(s)}(\boldsymbol{\theta})$ the sampling prior. Further we let $f^{(s)}(\boldsymbol{y}^{(n)})$ denote the marginal distribution resulted from the sampling prior. Now, we introduce the key quantity

$$\beta_s^{(n)} = E_s\Big[1\{P(h(\boldsymbol{\theta}) < \delta|\boldsymbol{y}^{(n)}, \pi^{(f)}) \geq \gamma\}\Big], \tag{5.18}$$

where the indicator function $1\{A\}$ is 1 if A is true and 0 otherwise, $\gamma > 0$ is a prespecified quantity, the probability is computed with respect to the posterior distribution given $\boldsymbol{y}^{(n)}$ and $\pi^{(f)}(\boldsymbol{\theta})$, and the expectation is taken with respect to the marginal distribution of $\boldsymbol{y}^{(n)}$ under the sampling prior $\pi^{(s)}(\boldsymbol{\theta})$.

Now, we propose a new Bayesian SSD algorithm as follows. Let $\overline{\Theta}_0$ and $\overline{\Theta}_1$ denote the closures of Θ_0 and Θ_1. Let $\pi_0^{(s)}(\boldsymbol{\theta})$ denote a "sampling prior" with support $\Theta_B = \overline{\Theta}_0 \cap \overline{\Theta}_1$. Also let $\pi_1^{(s)}(\boldsymbol{\theta})$ denote a "sampling prior" with the support $\Theta_1^* \subset \Theta_1$. For given $\alpha_0 > 0$ and $\alpha_1 > 0$, we compute $n_0 = \min\{n : \beta_{s0}^{(n)} \leq \alpha_0\}$ and $n_1 = \min\{n : \beta_{s1}^{(n)} \geq 1 - \alpha_1\}$, where $\beta_{s0}^{(n)}$ and $\beta_{s1}^{(n)}$ are given in (5.18) corresponding to $\pi^{(s)} = \pi_0^{(s)}$ and $\pi^{(s)} = \pi_1^{(s)}$. Then, the Bayesian sample size is given by

$$n_B = \max\{n_0, n_1\}. \tag{5.19}$$

According to the 2006 FDA Guidance Document, we choose $\gamma \geq 0.95$. Common choices of α_0 and α_1 include $\alpha_0 = 0.05$ and $\alpha_1 = 0.20$ so that the Bayesian sample size n_B given in (5.19) guarantees that the type I error rate is less than or equal to 0.05 and the power is at least 0.80. In addition, for a given sample size n_B, the operating characteristic curve can be constructed by varying Θ_1^* inside of Θ_1. If $h(\boldsymbol{\theta})$ is a monotonic function of the distance between Θ_1^* and Θ_B, then the far Θ_1^* is away from Θ_B, the higher the power will be. We will examine the properties of the proposed Bayesian SSD via several examples and apply this algorithm to Bayesian clinical trial design using the meta-analytic models developed in Aim 1.

A simple illustration: i.i.d. normal case. Suppose $y_1, y_2, \ldots, y_n$ are i.i.d. $N(\theta, w^{-1})$, where w is a known precision parameter. Suppose the hypotheses for non-inferiority testing are formulated as follows: H_0: $\theta \geq \delta$ versus H_1: $\theta < \delta$. We specify an improper uniform fitting prior for θ, i.e., $\pi^{(f)}(\theta) \propto 1$. In addition, we specify two point mass sampling priors for θ such that $\pi_0^{(s)}(\theta) = 1$ if $\theta = \delta$ and $\pi_1^{(s)}(\theta) = 1$ if $\theta = 0$. After some algebra, we can show that (i) a necessary condition for achieving a type I error rate of α_0 is $1 - \gamma \leq \alpha_0$ and (ii) if $1 - \gamma \leq \alpha_0$,

the Bayesian sample size in (5.19) is the smallest integer n_B satisfying

$$n_B \geq \frac{1}{w\delta^2}\Big[\Phi^{-1}(1-\alpha_1)+\Phi^{-1}(\gamma)\Big]^2, \tag{5.20}$$

where Φ denotes the $N(0,1)$ cumulative distribution function. It is interesting to note that for this simple case, $\beta_0^{(n)} \leq \alpha_0$ always holds for all n when $1-\gamma \leq \alpha_0$. We also note that the Bayesian sample size n_B in (5.20) is identical to the classical sample size formulation for a one-sided alternative hypothesis. However, if we apply ALC to a one-sided alternative hypothesis, the resulting Bayesian sample size is given by $n_{ALC} \geq \frac{1}{w\delta^2}\Big[\Phi^{-1}(\gamma)\Big]^2$ by considering a one-sided $100\gamma\%$ credible interval, which implies that ALC leads to a 50% power at $\theta = 0$. Thus, if $\alpha_1 < 0.50$, the sample size obtained from ALC cannot achieve the desirable power of $1-\alpha_1$.

Bayesian trial design assuming a single aggregate response per study. Suppose we consider a Bayesian trial design of K random non-inferiority trials and let y_k denote the aggregate response with a known precision parameter w_k. We assume the meta-analysis model given by (5.1) and (5.2) for the y_k. Suppose the hypotheses for non-inferiority testing are formulated as follows: H_0: $\mu \geq \delta$ versus H_1: $\mu < \delta$. We are interested in determining K to achieve a prespecified type I error and a prespecified power. We specify an improper uniform fitting prior for μ, i.e., $\pi^{(f)}(\mu) \propto 1$. Assume $1-\gamma \leq \alpha_0$. In this case, $\Theta_B = \{\mu = \delta\}$. We specify point mass sampling priors for μ such that $\pi_0^{(s)}(\mu) = 1$ if $\mu = \delta$ and $\pi_1^{(s)}(\mu) = 1$ if $\mu = 0$. The fitting conditional posterior distribution of μ given τ is of the form $\mu|\tau^2, \boldsymbol{y}^{(K)} \sim N(\hat{\mu}(\tau^2), \sigma^2(\tau^2))$, where $\boldsymbol{y}^{(K)} = (y_1, \ldots, y_K)'$, $\hat{\mu}(\tau^2) = \sigma^2(\tau^2)\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}y_k$, and $\sigma^2(\tau^2) = \Big(\sum_{k=1}^K \frac{w_k}{1+w_k\tau^2}\Big)^{-1}$. We assume the same proper inverse gamma for τ^2 for both the fitting prior and the sampling prior. Under these assumptions, (5.18) reduces to

$$\beta_s^{(K)} = E_s\Big(1\Big\{E_f\Big[\Phi\Big(\frac{\delta-\hat{\mu}(\tau^2)}{\sigma(\tau^2)}\Big)\Big] \geq \gamma\Big\}\Big), \tag{5.21}$$

where the expectation E_f is taken with respect to the marginal fitting posterior distribution of τ^2. When τ^2 is fixed, the Bayesian size of meta-analysis is available in closed-form. Specifically, when $1-\gamma \leq \alpha_0$, we can show that K_B is the smallest integer K satisfying the following inequality

$$\sum_{k=1}^{K}\frac{w_k}{1+w_k\tau^2} \geq \frac{1}{\delta^2}\Big[\Phi^{-1}(1-\alpha_1)+\Phi^{-1}(\gamma)\Big]^2. \tag{5.22}$$

When τ^2 is unknown, we propose a simulation-based procedure for computing $\beta_s^{(K)}$ in (5.21). We specify fitting and sampling priors for τ^2 and also assume μ and τ^2 are independent *apriori*. Let $f^{(s)}(\boldsymbol{y}^{(K)})$ denote the marginal distribution of $\boldsymbol{y}^{(K)}$ under the sampling prior $\pi^{(s)}(\mu,\tau^2)$. Also let $\{\boldsymbol{y}_i^{(K)},\ i = 1,\ldots,N\}$ denote a Monte Carlo sample from $f^{(s)}(\boldsymbol{y}^{(K)})$. For each $\boldsymbol{y}_i^{(K)}$, we generate a Monte Carlo sample $\{\tau_{ij}^2,\ j = 1,\ldots,M\}$ from the fitting posterior $\pi^{(f)}(\tau^2|\boldsymbol{y}_i^{(K)})$. Then, a Monte Carlo estimate of $\beta_s^{(K)}$ is given by

$$\hat{\beta}_s^{(K)} = \frac{1}{N}\sum_{i=1}^{N}1\Big\{\frac{1}{M}\Big[\Phi\Big(\frac{\delta-\hat{\mu}(\tau_{ij}^2)}{\sigma(\tau_{ij}^2)}\Big)\Big] \geq \gamma\Big\}. \tag{5.23}$$

Applying (5.23) to the sampling priors $\pi_0^{(s)}(\mu,\tau^2)$ and $\pi_1^{(s)}(\mu,\tau^2)$ yields the Monte Carlo estimates of $\hat{\beta}_{s0}^{(K)}$ and $\hat{\beta}_{s1}^{(K)}$. Then, we choose K_B to be the smallest integer satisfying $K_B \geq \min\{K:\ \hat{\beta}_{s0}^{(K)} \leq \alpha_0 \text{ and } \hat{\beta}_{s1}^{(K)} \geq 1-\alpha_1\}$ to achieve the prespecified type I error rate of α_0 and the power of $1-\alpha_1$. Thus, the implementation of the proposed simulation-based procedure is straightforward.

Bayesian trial design with random effects generalized linear models. We propose a general simulation-based approach for Bayesian trial design in the REGLM. Suppose we have the subject level continuous or discrete responses y_{ijk}'s and we are interested in the REGLM model in (5.13) for y_{ijk} for $i = 1,2,\ldots,n_{jk}$, $j = 1,2,\ldots,q$, and $k = 1,2,\ldots,K$. Since there are q treatment arms, we assume that the hypotheses for

non-inferiority testing can be formulated as follows: H_0: $g(\mu) \geq \delta$ versus H_1: $g(\mu) < \delta$, where $g(\mu)$ denotes a vector of G linearly or functionally independent contrasts of treatment effects μ and δ is a G-dimensional vector of prespecified margins. When $q = 2$, we may simply consider a scalar contrast $g(\mu) = \mu_2 - \mu_1$, which is sufficient for most applications in clinical trials.

For the subject level meta-analysis model, our design problem becomes much more complex than the case with only aggregate responses. To make the design via meta-analysis models more feasible, we assume

$$n_{jk} = \phi_{jk} n_k \text{ for } j = 1, 2, \ldots, q, \text{ and } n_k = \kappa_k n \text{ for } k = 1, 2, \ldots, K, \tag{5.24}$$

where both ϕ_{jk} and κ_k are prespecified nonnegative constants such that $\sum_{j=1}^{q} \phi_{jk} = 1$ and $\sum_{k=1}^{K} \kappa_k = 1$. Under this setting, the total sample size based on the entire meta-analytic model is n. We note that (5.24) is quite general and flexible, which allows $\phi_{jk} = 0$ for certain treatment arms. This implies that our setting allows an unbalanced design for certain studies. We further assume that the covariates (x_{ijk}, z_{ijk}) are generated from a prespecified joint distribution, $f(x_{ijk}, z_{ijk} \mid \alpha)$, where α is determined from the prior studies. From the design point of view, β, σ, Γ, and V are considered as nuisance parameters. Let $\theta = (\mu, \beta, \sigma, \Gamma, V)$. We specify the fitting prior as follows: $\pi^{(f)}(\theta) \propto \pi^{(f)}(\beta, \sigma, \Gamma, V)$, which assumes an improper uniform fitting prior for μ. The sampling priors are specified as follows: $\pi_0^{(s)}(\theta) = \pi_0^{(s)}(\mu)\pi^{(s)}(\beta, \sigma, \Gamma, V)$, where $\pi_0^{(s)}(\mu)$ is a proper prior with the support $\{\mu : \ g(\mu) = \delta\}$ and $\pi^{(s)}(\beta, \sigma, \Gamma, V)$ denotes the sampling prior for β, σ, Γ, and V, and $\pi_1^{(s)}(\theta) = \pi_1^{(s)}(\mu)\pi^{(s)}(\beta, \sigma, \Gamma, V)$, where $\pi_1^{(s)}(\mu)$ is a proper prior with the support $\{\mu : \ g(\mu) = 0\}$. This essentially assumes that the same fitting prior for nuisance parameters β, σ, Γ, and V in computing the type I error and the power.

Under the above setting, (5.18) reduces to

$$\beta_s^{(n)} = E_s\Big(1\Big\{E_f\Big[1\{g(\mu) < \delta\}\Big] \geq \gamma\Big\}\Big), \tag{5.25}$$

where the expectation E_f is taken with respect to the marginal fitting posterior distribution of μ and the expectation E_s is taken with respect to the sampling distribution of the y_{ijk} under the sampling prior. We will develop a simulation-based approach for computing $\beta_s^{(n)}$. Specifically, we generate (x_{ijk}, z_{ijk}) from $f(x_{ijk}, z_{ijk} \mid \alpha)$, then generate θ from the sampling prior $\pi^{(s)}(\theta)$, which is either $\pi_0^{(s)}(\theta)$ or $\pi_1^{(s)}(\theta)$, and finally we generate y_{ijk} according to (5.13). For each set of the data $(y_{ijk}, x_{ijk}, z_{ijk})$, we generate a Monte Carlo sample of μ from the fitting posterior. Using these Monte Carlo samples, we can easily obtain a Monte Carlo estimate of $\beta_s^{(n)}$ and then calculate the Bayesian sample size n_B for the entire meta analysis. Moreover, for the random effects normal linear models or the random effects normal linear mixed models, we will develop a more efficient Monte Carlo estimates of $\beta_s^{(n)}$. Finally, we note that the design of the meta-analysis given in (5.24) can be used in three different scenarios/applications: (i) no individual trials are completed; (ii) individual trials are partially completed; and (iii) all individual trials are completed. In (ii), given that K_0 individual trials are completed and the desired power is not met yet, the proposed Bayesian SD allows us to add $K - K_0$ additional new individual trials to achieve a prespecified power. In (iii), the proposed Bayesian SD can be used to determine how many studies are needed to achieve the desired power. In addition, the proposed Bayesian SD can easily account for covariates in the design of a meta-analysis.

A simulation study for random effects binomial models. We conduct a small simulation study to examine the performance of the proposed Bayesian SSD using meta-analytic models in non-inferiority clinical trials. We consider random effects binomial models with the logit link for the aggregate responses y_{jk}'s as follows $f(y_{jk}|\mu_{jk}) = \exp(y_{jk}\mu_{jk})/[1 + \exp(\mu_{jk})]^{n_{jk}}$ and $\mu_{jk} \sim N(\mu_j, \sigma_j^2)$ independently for $j = 1, 2$ and $k = 1, 2, \ldots, 8$, where "$j = 1$" denotes the test group and "$j = 2$" denotes the control group. The hypotheses for non-inferiority testing are given by $H_0 : \ \mu_1 - \mu_2 \geq \delta$ versus $H_1 : \ \mu_1 - \mu_2 < \delta$. We consider a balanced design for four trials and an unbalanced design for other four trials. The sample sizes are given as follows:

Group	Sample Size							
Test	154	279	284	694	716	269	174	221
Control	254	282	-	520	382	-	-	-

Based on the prior information from actual previous clinical trials, the non-inferiority margin, $\delta = 0.42$, was determined. The fitting prior is specified as follows $\pi^{(f)}(\mu) \propto 1$ and independently $\sigma_j^2 \sim \mathcal{G}(0.001, 0.001)$ for $j = 1, 2$. Similarly to Wang and Gelfand (2002), let $\mu_2 \sim U(-2.4, -2.0)$, reflecting the non-inferiority rates between 0.08 and 0.12. Then we set $\mu_1 = \mu_2 + \delta$ for $\pi_0^{(s)}$ and $\mu_1 = \mu_2$ for $\pi_1^{(s)}$. $\gamma = 0.95$ was used throughout the simulation study. When the fitting prior for σ_j^2 is $U(0.001, 0.005)$, reflecting a small heterogeneity across 8 trials, using the Monte Carlo sizes $N = 1,000$ and $M = 5,000$, the Monte Carlo estimates of the type I error and power are 0.004 and 0.794. When the fitting prior for σ_j^2 is $U(0.01, 0.05)$ reflecting a moderate heterogeneity across 8 trials, the estimated type I error and power are 0.01 and 0.573. When the fitting prior for σ_j^2 is $U(0.02, 0.08)$ reflecting a large heterogeneity, the estimated type I error and power are 0.012 and 0.464. These results are expected because the power using these meta-analytic models decreases when the heterogeneity increases. Surprisingly, under various sampling priors for σ_j^2's, the Bayesian type I errors are well controlled and all are less than 0.05. These empirical results demonstrate that the proposed Bayesian SSD is quite promising.

Design of Bayesian meta-analysis for semiparametric models for survival data with gamma process priors. For the semiparametric models for survival data with gamma process priors, the full likelihood function is given by (5.16). For the subject level meta-analysis model for survival data, we consider the same design setting, the same hypotheses for non-inferiority testing, and the same fitting prior and sampling priors as the ones for the random effects generalized linear models. However, for survival data, it becomes more difficult for Bayesian SSD, which may be partially due to the censored failure times and partially due to the use of gamma process priors for the cumulative baseline hazard. In particular, it is quite challenging to generate the survival times y_{ijk} from their joint marginal distribution under the sampling priors as well as to sample from the posterior distribution under the fitting prior. Our major effort here will be on developing efficient Monte Carlo algorithms for generating Monte Carlo samples from these distributions.

We first develop a simulation algorithm for sampling the y_{ijk} from the marginal distribution under the sampling prior. To generate $H_0(y) \sim \mathcal{GP}(H^*, c_0)$, let $t_0 = 0 < t_1 < t_2 < \ldots t_J$ to be J knots so that $t_j - t_{j-1}$ is small and t_J is sufficiently large so that the failure time will not exceed t_J practically. Let $h_{0j} = H_0(t_j) - H_0(t_{j-1})$ and $h_{0j}^* = H_0^*(t_j) - H_0^*(t_{j-1})$ for $j = 1, 2, \ldots, J$. Due to the property of the gamma process prior, we generate $h_{0j} \sim \mathcal{G}(c_0 h_{0j}^*, c_0)$ independently for $j = 1, 2, \ldots, J$. Then, we compute $H_0(t_j) = \sum_{l=1}^{j} h_{0j}$ for $j = 1, 2, \ldots, J$. Similarly to the random effects generalized linear models, we generate $(\boldsymbol{x}_{ijk}, \boldsymbol{z}_{ijk})$ from $f(\boldsymbol{x}_{ijk}, \boldsymbol{z}_{ijk} \mid \boldsymbol{\alpha})$, and θ from the sampling prior $\pi^{(s)}(\theta)$, and compute $\exp(\boldsymbol{x}_{ijk}'\beta + \boldsymbol{z}_{ijk}'\gamma_k + \mu_{jk})$. Then the failure time T_{ijk} is generated as follows. Let $u_{ijk} \sim U(0, 1)$. Let j^* be an integer such that

$$\exp\{-H_0(t_{j^*+1})\exp(\boldsymbol{x}_{ijk}'\beta + \boldsymbol{z}_{ijk}'\gamma_k + \mu_{jk})\} < 1 - u_{ijk} \leq \exp\{-H_0(t_{j^*})\exp(\boldsymbol{x}_{ijk}'\beta + \boldsymbol{z}_{ijk}'\gamma_k + \mu_{jk})\}.$$

Using the linear interpolation technique, we set

$$T_{ijk} = t_{j^*} + \frac{[(1 - u_{ijk}) - \exp\{-H_0(t_{j^*+1})\exp(\boldsymbol{x}_{ijk}'\beta + \boldsymbol{z}_{ijk}'\gamma_k + \mu_{jk})\}](t_{j^*+1} - t_{j^*})}{\exp\{-H_0(t_{j^*})\exp(\boldsymbol{x}_{ijk}'\beta + \boldsymbol{z}_{ijk}'\gamma_k + \mu_{jk})\} - \exp\{-H_0(t_{j^*+1})\exp(\boldsymbol{x}_{ijk}'\beta + \boldsymbol{z}_{ijk}'\gamma_k + \mu_{jk})\}}.$$

Independently we generate a censoring time C_{ijk} from a uniform distribution $U(C_1, C_2)$. Then $y_{ijk} = \min\{T_{ijk}, C_{ijk}\}$ and $\nu_{ijk} = 1$ if $y_{ijk} = T_{ijk}$ and $\nu_{ijk} = 0$ otherwise.

Next, we discuss how to sample from the posterior distribution under the fitting prior. Let $y_{(1)} < y_{(2)} < \cdots < y_{(n^*)}$ be the n^* distinct failure and censoring times of the y_{ijk}'s. Write $y_{(l)} = y_{ijk_l}$ and ν_{ijk_l} denotes the corresponding censoring indicator for $l = 1, 2, \ldots, n^*$. Following Chen, Ibrahim, and Shao (2006), after integrating out $H_0(y)$, we have

$$L(\beta, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \gamma_1, \ldots, \gamma_K | D_{obs}, c_0) = \prod_{l=1}^{n^*} \left\{ \exp\left[c_0 H^*(y_{(l)}) \log\left(\frac{c_0 + A_{l+1}}{c_0 + A_l}\right)\right] \right.$$
$$\left. \times \prod_{(i,j,k) \in \mathcal{D}(y_{(l)})} \left[- c_0 h^*(y_{(l)}) \log\left(1 - \frac{\exp(\boldsymbol{x}_{ijk}'\beta + \boldsymbol{z}_{ijk}'\gamma_k + \mu_{jk})}{c_0 + A_l}\right)\right]^{\nu_{ijk_l}} \right\}, \quad (5.26)$$

where $\mathcal{D}(y_{(l)}) = \{l : \ y_{ijk} = y_{(l)}, \nu_{ijk} = 1\}$, $A_l = \sum_{(i,j,k)\in\mathcal{R}(y_{(l)})} \exp(x'_{ijk}\beta + z'_{ijk}\gamma_k + \mu_{jk})$, and $h^*(y) = \frac{dH^*(y)}{dy}$. Observe that

$$\left[- c_0 h^*(y_{(l)}) \log\left(1 - \frac{\exp(x'_{ijk}\beta + z'_{ijk}\gamma_k + \mu_{jk})}{c_0 + A_l}\right)\right]^{\nu_{ijk_l}} = \int_0^1 \int_0^\infty \exp\left\{\nu_{ijk_l}(x'_{ijk}\beta + z'_{ijk}\gamma_k + \mu_{jk})\right\}$$
$$\times \exp\left(- t_{ijk}\left[(1-\nu_{ijk_l}) + \nu_{ijk_l}\left\{c_0 + A_l - \omega_{ijk}\exp(x'_{ijk}\beta + z'_{ijk}\gamma_k + \mu_{jk})\right\}\right]\right) dt_{ijk} d\omega_{ijk} \quad (5.27)$$

$$\text{and} \quad \exp\left[c_0 h_l^* \log\left(\frac{c_0}{c_0 + A_l}\right)\right] = \int_0^\infty \exp(-h_i A_l) \frac{c_0^{c_0 h_l^*} h_i^{c_0 h_l^* - 1}}{\Gamma(c_0 h_l^*)} \exp(-c_0 h_i) dh_l, \quad (5.28)$$

where $h_{0l}^* = H^*(y_{(l)}) - H^*(y_{(l-1)})$. Using (5.26) to (5.28), we can develop a very efficient Markov chain Monte Carlo sampling algorithm via three sets of latent variables $\boldsymbol{\omega} = (\omega_{ijk})$, $\boldsymbol{t} = (t_{ijk})$, and $\boldsymbol{h} = (h_l)$. Once these two Monte Carlo samples are obtained, computing $\beta_s^{(n)}$ in (5.25) is straightforward.

Missing data and other considerations. Missing data are inevitable in clinical trials. In Bayesian design of for non-inferiority trials using these proposed meta-analytic models, we will carefully investigate the impact of missing data on the power and the type I error rate.

We will develop a simulation-based approach to quantify the effective sample size due to missing responses and/or covariates. In the proposed research, we will also examine theoretical and empirical connections between the proposed Bayesian SSD method and other existing criteria such as ACC, ALC, and WOC. The prior specification plays an important role in Bayesian SSD. The prior specification for the unknown variance and covariance parameters under the random effects generalized linear models or generalized linear mixed models becomes even more crucial. Based on our preliminary analysis here, the sampling priors should be much more informative than the fitting priors. In the proposed research, we will consider and extend the priors discussed in Gelman (2006) for the unknown variance and covariance parameters and we will also conduct various sensitivity analyses of the specification of fitting prior and sampling priors.

5.3 Meta-Analytic Methodology of Diagnostic Tests without a Gold Standard

Meta-Analytic methodology of two diagnostic tests without a gold standard using maximum likelihood and full Bayesian methods. estimating the accuracy of two imperfect diagnostic tests (A and B) in a meta-analysis or a multi-center clinical trial. According to convention, we focus on dichotomized test results as the outcome of interest. For study $i (i = 1, 2, \ldots, \text{I})$, let $P_{ijk} = \Pr(\text{A} = j, \text{B} = k)$ be the joint probability of test results and n_{ijk} be the corresponding observed count, $j, k = 0, 1$. Let π_i be the study-specific disease prevalence, and let $(Se_{iA}, Se_{iB}, Sp_{iA}, Sp_{iB})$ be the corresponding sensitivities and specificities for test A and test B. Under the assumption that the two tests are independent conditional on the true disease status, study-specific prevalence, sensitivities and specificities, we have the following relationship:

$$P_{i11} = \pi_i Se_{iA} Se_{iB} + (1-\pi_i)(1-Sp_{iA})(1-Sp_{iB}), P_{i10} = \pi_i Se_{iA}(1-Se_{iB}) + (1-\pi_i)(1-Sp_{iA})Sp_{iB}, \quad (5.29)$$
$$P_{i01} = \pi_i(1-Se_{iA})Se_{iB} + (1-\pi_i)Sp_{iA}(1-Sp_{iB}), P_{i00} = \pi_i(1-Se_{iA})(1-Se_{iB}) + (1-\pi_i)Sp_{iA}Sp_{iB}.$$

It is well known that if the conditional independence assumption is falsely assumed, parameter estimates can be biased (Vacek 1985; Torrance-Rynard and Walter 1997; Dendukuri and Joseph 2001). When the possibility of conditional dependence cannot be completely ruled out, as a sensitivity analysis to the conditional independence assumption, we extend the model in equation (5.29) to allow dependence. Specifically, we will incorporate the residual dependence of the two tests given the latent disease status and study-specific random effects by assuming homogeneous residual dependence across all studies. Let ρ_1 and ρ_0 denote the correlation of the two tests when the true disease status is positive and negative, respectively, equation (5.29) becomes (Vacek 1985; Shen, Wu, and Zelen 2001; Dendukuri and Joseph 2001),

$$\begin{aligned} P_{i11} &= \pi_i[Se_{iA}Se_{iB} + \delta_{1i}] + (1-\pi_i)[(1-Sp_{iA})(1-Sp_{iB}) + \delta_{0i}], \quad (5.30)\\ P_{i10} &= \pi_i[Se_{iA}(1-Se_{iB}) - \delta_{1i}] + (1-\pi_i)[(1-Sp_{iA})Sp_{iB} - \delta_{0i}]\\ P_{i01} &= \pi_i[(1-Se_{iA})Se_{iB} - \delta_{1i}] + (1-\pi_i)[Sp_{iA}(1-Sp_{iB}) - \delta_{0i}],\\ P_{i00} &= \pi_i[(1-Se_{iA})(1-Se_{iB}) + \delta_{1i}] + (1-\pi_i)[Sp_{iA}Sp_{iB} + \delta_{0i}], \end{aligned}$$

where $\delta_{1i} = \rho_1\sqrt{Se_{iA}Se_{iB}(1-Se_{iA})(1-Se_{iB})}$ and $\delta_{0i} = \rho_0\sqrt{Sp_{iA}Sp_{iB}(1-Sp_{iA})(1-Sp_{iB})}$ are the covariances between two tests among the diseased and non-diseased subjects in study i, respectively. The feasible range of correlations is determined by the sensitivities among diseased subjects and specificities among non-diseased subjects in each study. Specifically, the correlation coefficients ρ_1 and ρ_0 satisfy

$$\max_i\left\{-\sqrt{\frac{Se_{iA}Se_{iB}}{(1-Se_{iA})(1-Se_{iB})}},-\sqrt{\frac{(1-Se_{iA})(1-Se_{iB})}{Se_{iA}Se_{iB}}}\right\} \le \rho_1 \le \min_i\left\{\sqrt{\frac{Se_{iA}(1-Se_{iB})}{(1-Se_{iA})Se_{iB}}},\sqrt{\frac{(1-Se_{iA})Se_{iB}}{Se_{iA}(1-Se_{iB})}}\right\},$$

$$\max_i\left\{-\sqrt{\frac{Sp_{iA}Sp_{iB}}{(1-Sp_{iA})(1-Sp_{iB})}},-\sqrt{\frac{(1-Sp_{iA})(1-Sp_{iB})}{Sp_{iA}Sp_{iB}}}\right\} \le \rho_0 \le \min_i\left\{\sqrt{\frac{Sp_{iA}(1-Sp_{iB})}{(1-Sp_{iA})Sp_{iB}}},\sqrt{\frac{(1-Sp_{iA})Sp_{iB}}{Sp_{iA}(1-Sp_{iB})}}\right\}.$$

Although negative associations are possible, it seems more plausible that $\rho_i \ge 0 (i = 0, 1)$, which corresponds to positive dependence conditional on the latent disease status and study-specific random effects. If homogeneous conditional dependence between studies looks suspicious, methods allowing more complex dependent errors need to be considered, for example, by considering study-specific correlation coefficients ρ_{1i} and ρ_{0i} in equation (5.30). The homogeneous conditional dependence can also be specified by other parameterizations such as homogeneous odds ratio or relative risk (Böhning and Patilea 2008). In the presence of missing data due to partial testing, structure and notation for a study. we denote the probabilities of study i to fall in categories A and B by ω_{iA} and ω_{iB}.

Under the missing at random (MAR) assumption, the likelihood function can be factored into $L(\boldsymbol{\theta}_i, \vartheta_i) = L(\boldsymbol{\theta}_i) \times L(\vartheta_i)$, where $\boldsymbol{\theta}_i = (\pi_i, Se_{iA}, Se_{iB}, Sp_{iA}, Sp_{iB}, \rho_0, \rho_1)$ and $\vartheta_i = (\omega_{iA}, \omega_{iB})$. Assuming independence among subjects conditional on $\boldsymbol{\theta}$, the log-likelihood for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_I)$ is given by

$$\log L(\boldsymbol{\theta}|data) = \sum_i\{n_{i11}\log(P_{i11}) + n_{i10}\log(P_{i10}) + n_{i01}\log(P_{i01}) + n_{i00}\log(P_{i00}) +$$
$$n_{i1m}\log(P_{i11}+P_{i10}) + n_{i0m}\log(P_{i01}+P_{i00}) + n_{im1}\log(P_{i11}+P_{i01}) + n_{im0}\log(P_{i10}+P_{i00})\}, \quad (5.31)$$

where the relations among the components of $\boldsymbol{\theta}_i$ and P_{ijk} are summarized in (5.29) under conditional independence or (5.30) under conditional dependence. Between-study heterogeneity commonly exists in a meta-analysis since studies usually differ in their subject recruitment methods and laboratory techniques as well as arguably in overall study quality as reflected in the study protocol and adherence to the protocol. Thus, measurements within a study tend to be correlated beyond what would be anticipated for measurements between studies. To take into account of the potential between-study heterogeneity of the prevalence, sensitivity and specificity, we consider a random effects model. We introduce covariate vector X to model study-level covariates effect on prevalence. The generalized linear mixed effects model can then be specified as follows:

$$g(\pi_i|\varepsilon_i) = \mathbf{X}\boldsymbol{\eta} + \varepsilon_i, g(Se_{iA}|\mu_{iB}) = \alpha_A + \mu_{iA}, g(Se_{iB}|\mu_{iB}) = \alpha_B + \mu_{iB}, g(Se_{iA}|\mu_{iA}) = \beta_A + \nu_{iA}, \quad (5.32)$$
$$g(Sp_{iB}|\nu_{iB}) = \beta_B + \nu_{iB}, (\varepsilon_i, \mu_{iA}, \mu_{iB}, \nu_{iA}, \nu_{iB}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $g(\cdot|\cdot)$ is the link function such as the commonly used logit, probit or complementary log-log transformation functions. The diagonal elements of the variance-covariance matrix $\boldsymbol{\Sigma}$ capture the extent of heterogeneity of the parameters of interest across studies. If there is statistical or scientific evidence of homogeneity, the corresponding study-specific random effect(s) can be dropped from the model. We will adopt two approaches to make inference from the above random effects models. The first is a nonlinear mixed effects model (NLMM) (Davidian and Giltinan 1995; Vonesh and Chinchilli 1997; Molenberghs and Verbeke 2005); the second is a Bayesian hierarchical model (Carlin and Louis 2000; Gelman et al. 1995). Since these two approaches use different frameworks and different software, they can be considered complementary. In most instances, inferences obtained by Bayesian and frequentist methods agree when weak prior distributions are specified. However, the Bayesian framework is particularly attractive when suitable proper prior distributions can be constructed to incorporate known constraints and subject-matter knowledge on model parameters (Davidian and Giltinan 2003). Furthermore, the Bayesian framework provides for direct construction of $100(1-\alpha)\%$ equal tail and highest probability density (HPD) credible intervals of general functions of the estimated parameters without having to rely on asymptotic approximations. To avoid over-fitting the data with an excess of random effects, we used a

forward selection procedure based on information criteria. Specifically, Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) will be used as the guideline (Burnham and Anderson 1998) for NLMM, and the deviance information criterion (DIC) will be used for the Bayesian hierarchical models (Spiegelhalter et al. 2002). At each forward step, we added a random-effect component that provided the largest improvement based on the above model selection criteria.Furthermore, we propose a graphical method, the Kappa agreement plot, to quantitatively validate the conditional dependence assumption for each study based on the final model. This plot is obtained by plotting the model-based marginal agreement between the two tests for study i measured by the Kappa statistics (κ_i) with 95% confidence (or credible) intervals, which corrects the agreement that may occur by chance alone, against the observed marginal agreement between the two tests for study i. The model-based Kappa statistics for study i can be computed by

$$\kappa_i = \frac{P_{i11} + P_{i00} - (P_{i11} + P_{i10})(P_{i11} + P_{i01}) - (P_{i00} + P_{i10})(P_{i00} + P_{i01})}{1 - (P_{i11} + P_{i10})(P_{i11} + P_{i01}) - (P_{i00} + P_{i10})(P_{i00} + P_{i01})}. \tag{5.33}$$

If the model based 95% confidence (or credible) intervals include the observed Kappa statistics at close to the nominal rate, then we would lack evidence to reject the conditional independence assumption. We will reanalyze the meta-analysis data of 17 studies to evaluate the accuracy of microsatellite instability testing (MSI) in predicting Lynch syndrome (Chen, Watson, and Parmigiani 2005), and use simulations to investigate the importance of including appropriate random effects and the impact of overfitting, under-fitting and misfitting on model performance using both the nonlinear random effects models and Bayesian hierarchical models.

Meta-Analytic methodology of multiple diagnostic tests without a gold standard using full Bayesian methods. considered by Qu and colleagues for estimating the accuracies of multiple diagnostic tests in a single group (Qu, Tan, and Kutner 1996; Qu and Hadgu 1998), and propose statistical methods to estimate the accuracy of multiple imperfect diagnostic tests (≥ 3) in a meta-analysis or a multicenter clinical trial where heterogeneity is intrinsic among studies. Let the random variable T_j represent the classification based on the j^{th} diagnostic test ($j = 1, 2, \ldots, J$) with $T_j = 1$ if test positive and $T_j = 0$ if test negative. Let D represent the disease status with 1 denoting a case and 0 indicating a non-case. Let T_{ijk} represent the classification of the j^{th} test with a value 1 indicating test positive and 0 indicating test negative on the k^{th} subject ($k = 1, 2, \ldots, n_i$) in the i^{th} study ($i = 1, 2, \ldots, I$). For the relationship between imperfect measurements, in principle, we allow conditional dependent misclassification (i.e., the imperfect measurements can be correlated conditioning on the latent disease status). Specifically, the probability of disease in the i^{th} study given study-level covariate vector $\mathbf{X}$ and random effects ε_i are modeled through a generalized linear regression model as,

$$\pi_{ik}(\mathbf{X}) = P(D_{ik} = 1|\mathbf{X}, \varepsilon_i) = g^{-1}(\mathbf{X}\eta + \varepsilon_i), \tag{5.34}$$

where $g^{-1}(\cdot)$ is the inverse link function. The positive classification probability for the k^{th} subject in the i^{th} study by the j^{th} diagnostic test is assumed to dependent on the latent disease status D_{ik}, and a Gaussian latent variable Z capturing the correlation among J diagnostic tests for an individual and random effects μ_{ij} and ν_{ij} capturing the heterogeneity of sensitivities and specificities among studies for the j^{th} diagnostic test through a generalized linear regression model,

$$\begin{aligned} P(T_{ijk} = 1|D_{ik} = 1, Z = z_{ik}, \mu_{ij}) &= g^{-1}(\alpha_{1j} + \alpha_{2j} z_{ik} + \mu_{ij}), \\ P(T_{ijk} = 1|D_{ik} = 0, Z = z_{ik}, \nu_{ij}) &= g^{-1}(\beta_{1j} + \beta_{2j} z_{ik} + \nu_{ij}), \end{aligned} \tag{5.35}$$

where $z_{ik} \sim N(0, 1)$ capturing the conditional dependence at individual level, and $(\varepsilon_i, \boldsymbol{\mu_{ij}}, \boldsymbol{\nu_{ij}}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ capturing the between-study heterogeneity of prevalence, sensitivities and specificities, and $\boldsymbol{\mu_{ij}} = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{iJ})$, $\boldsymbol{\nu_{ij}} = (\nu_{i1}, \nu_{i2}, \cdots, \nu_{iJ})$. Given the unusual large dimension of random effects, new technical and powerful computational tools need to be developed. Conditioning on the study-level random effects $(\varepsilon_i, \boldsymbol{\mu_{ij}}, \boldsymbol{\nu_{ij}}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, the marginal probability of observing subjects with $T_1 = t_1, \ldots,$ and $T_J = t_J$ in the i^{th} study $P_i(\mathbf{T}, \mathbf{X})$ is

$$\begin{aligned} P_i(\mathbf{T}, \mathbf{X}) = \pi_{ik}(\mathbf{x}) \int_{-\infty}^{+\infty} \prod_{j=1}^{J} g^{-1}(\alpha_{1j} + \alpha_{2j} z_{ik} + \mu_{ij})^{t_{ijk}} \left\{1 - g^{-1}(\alpha_{1j} + \alpha_{2j} z_{ik} + \mu_{ij})\right\}^{1-t_{ijk}} d\Phi(z) + \\ [1 - \pi_{ik}(\mathbf{x})] \int_{-\infty}^{+\infty} \prod_{j=1}^{J} g^{-1}(\beta_{1j} + \beta_{2j} z_{ik} + \nu_{ij})^{t_{ijk}} \left\{1 - g^{-1}(\beta_{1j} + \beta_{2j} z_{ik} + \nu_{ij})\right\}^{1-t_{ijk}} d\Phi(z). \end{aligned} \tag{5.36}$$

Let $n_i(\mathbf{T},\mathbf{X})$ be the number of subjects with covariate vector $\mathbf{X}$ in the i^{th} study classified by the diagnostic test vector $\mathbf{T} = (t_1, t_2, \ldots, t_J)$. The conditional likelihood function for $(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma})$ with the study-level random effects $(\varepsilon_i,\boldsymbol{\mu_{ij}},\boldsymbol{\nu_{ij}})$ is the product of the contribution from each category, that is $L(data|\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}) = \prod_i \prod_{\mathbf{X}} P_i(\mathbf{T},\mathbf{X})^{n_i(\mathbf{T},\mathbf{X})}$. Let the prior joint distribution of $(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma})$ to be specified as $f(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma})$, then the posterior joint distribution $f(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}|data)$ is proportional to

$$f(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}|data) = \prod_i \prod_{\mathbf{X}} P_i(\mathbf{T},\mathbf{X})^{n_i(\mathbf{T},\mathbf{X})} f(\varepsilon_i,\boldsymbol{\mu_{ij}},\boldsymbol{\nu_{ij}}) f(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}). \tag{5.37}$$

We will consider commonly used link functions including logit, probit and complementary log-log functions and develop Markov chain Monte Carlo sampling methods to estimate the posterior distributions. We will analyze a multi-center data set from NCI Colorectal Cancer Family Study (data request approved as C-EX-1107-01 and C-EX-1107-01-A1) to evaluate the accuracy of 10 biomarkers (BAT25, BAT26, BAT40, BAT34C4, D5S346, D17S250, D18S55, D1097, ACTC and MycL) in predicting Lynch syndrome, and use extensive simulations to investigate the importance of including appropriate random effects and the impact of overfitting, under-fitting and misfitting on model performance using the Bayesian hierarchical models.

5.4 Regression Analysis of Rare Adverse Events for Post-marketing Safety Evaluation

Methodology for rare events. Recall that our definition of a rare event means that $p \to 0$ but $np \to \infty$ as $n \to \infty$. This means we are accruing more information as the sample size gets larger, and this usual means that the standard statistics are typically asymptotically normal, although not at the usual $\sqrt{n}$ rate. As mentioned previously, log-rank tests have been developed for this setting (Buyske et al., 2000), but very few other semiparametric approaches have been developed for this situation. As mentioned in Section 2.4, exact conditional permutation approaches have been developed for a few special logistic regression and log-rank settings (Hirji et al., 1987; Mehta and Patel, 2000; Heinze et al., 2003), but, unfortunately, these exact approaches can be very computationally intense, have a limited range of applicability, and are not suitable for sample size formula development. Nevertheless, we will carefully compare performance of our proposed large sample approach with exact approaches when possible as part of our overall performance assessment. However, our focus will primarily be on developing methodology and large sample results for several settings not covered by exact approaches, including supremum-type log-rank statistics (see, e.g., Eng and Kosorok, 2005), covariate-adjusted log-rank statistics (see, e.g., Kong and Slud, 1997; Li, 2001), and Cox proportional hazards regression for rare events, including also sample size formulas and robust inference techniques provided we have sufficient time. For some of this development, continuous time martingale central limit theorems will be sufficient. For the Cox model and robust inference, we will need to use the special empirical process central limit theorems for models changing with sample size mentioned in the preliminary results section.

An important complication is that the baseline hazard function approximates zero for rare events. This means that accuracy of estimators must be framed as a ratio rather than as a difference. Classical statistical analysis tools can still be employed but more care needs to be taken since a difference between quantities going to zero does not imply that the ratio of the quantities goes to one. Fortunately, both martingale and empirical process tools are flexible enough to adapt to this challenge. An advantage of rare event data is that the number of patients at risk for an event are very large. Consider, for example, the Nelson-Aalan estimator $\hat{\Lambda}_n$ mentioned above. The variance of $\hat{\Lambda}_n(t)/\Lambda(t)$ can be shown to be of order $[n\Lambda(t)]^{-1}$. In the rare event setting, $\Lambda(t) \to 0$ but $n\Lambda(t) \to \infty$ as $n \to \infty$, and thus the variance goes to zero. Hence $\hat{\Lambda}_n(t)/\Lambda(t) \to 1$ in probability, even though $\Lambda(t) \to 0$. While this is a slight over-simplification of the problem, this same general idea is applicable in some generality to the problem we will be considering and will serve as a guide to facilitate success. Sample size formulas will be developed along the lines of Eng and Kosorok (2005), and robust Cox model will follow the general strategy of Lin and Wei (1989) but through utilizing more delicate empirical process arguments.

Methodology for extremely rare events. Nelson-Aalan estimator setting, $n\Lambda(t) \to K_0(t) = \int_0^t k_0(s)ds$, as $n \to \infty$, where $0 < k_0(t) < \infty$, for $0 \le t \le \tau$ and some $\tau < \infty$ (and assuming Λ is suitably bounded and smooth over the interval $[0,\tau]$). In this situation, it can be shown that $N(t)$ converges in distribution uniformly to a Poisson process with integrated intensity $K_0(t)$. As mentioned in Section 2.4, some exact methods have been developed for the log-rank setting (Mehta and Patel, 2000; Heinze et al., 2003), but these have important computational and modeling limitations. The situation is much more complicated for models involving covariates such as the Cox model or for other semiparametric models. Consider, for example, the Cox score statistic $U_n(\beta,t)$ evaluated

at a given regression parameter value $\beta \in \mathbb{R}^d$ and at time $t \in [0, \tau]$. For large enough n, $Y(t) \to 1$, in probability, for all $t \in [0, \tau]$, because the number at risk is typically much larger than the number having the extremely rare event (by definition, essentially). Hence

$$U_n(\beta, t) = \sum_{i=1}^{n} \delta_i \left(Z_i - \frac{E[Ze^{\beta' Z}]}{E[e^{\beta' Z}]} \right) + o_P(1), \tag{5.38}$$

where the $o_P(1)$ goes to zero in probability uniformly over $t \in [0, \tau]$. It can be shown, using the techniques we will describe shortly, that $U_n(\beta, t)$ converges in distribution uniformly to the compound Poisson process $Z_0(t)$, where $Z_0(t) = \sum_{j=1}^{P_0(t)} \left[\tilde{Z}_j - Z_0(\beta) \right]$, $P_0(t)$ is a Poisson process with integrated intensity $K_0(t)$, $\tilde{Z}_1, \tilde{Z}_2, \ldots$ are an i.i.d. sequence of random variables having density $f_0(z)$ proportional to $e^{\beta_0' z} g(z)$, $Z_0(\beta)$ is the expectation of Z under the density proportional to $e^{\beta' z} g(z)$, and where $g(z)$ is the true distribution of the covariate in the population. While this limiting distribution is complicated, it is quite manageable for doing inference with.

The general structure and method needed to establish week convergence and to develop methods of inference appears to be quite similar across many rare event settings. As mentioned previously, the technical structure of the change-point problem shares surprisingly many features with the extremely rare event model estimation problem. We will employ the general strategy given in Kosorok and Song (2007) which strategy was also used to obtain the results in the above paragraph. Note that many of the techniques for change-point estimation have been around for decades, but, what is new about the approach given in Kosorok and Song is that a novel bootstrap procedure for confidence intervals in the compound Poisson setting is obtained. There are basically four or five steps to this procedure. The first step is to use empirical process entropy control methods to verify the proper rate of convergence (n instead of $\sqrt{n}$ in this case). The second step is to use a careful characteristic function argument to verify that all finite-dimensional distributions converge jointly. In the Cox model setting, this means that we have joint convergence of $U_n(\beta, t)$ evaluated at any finite set of times $t_1, \ldots, t_m$ and finite set of regression parameters $\beta_1, \ldots, \beta_k$. The third step is to establish asymptotic uniform tightness of the entire process. This will yield weak convergence to a Poisson, or compound Poisson, process. Sometimes a fourth step is needed at this point to bridge between this Poisson process and the actual estimator of interest. For example, for the Cox model problem, the quantity of interest is actually the solution of $U_n(\beta, \tau) = 0$. A delicate variant of the continuous mapping theorem may need to be invoked if the maximizer is non-unique as is the case for the change-point problem. The fifth step is to develop a modified bootstrap tool for inference.

One additional challenge arises when $n\Lambda(\tau) \to \infty$ very slowly, as $n \to \infty$, so that it may be difficult to classify the event type as rare or extremely rare. Fortunately, the limiting distributions for these two event types merge smoothly with each other. This follows from the simple observation that a Poisson process converges to a Gaussian process as the Poisson rate parameter increases. The issue will need to be explored in greater detail, but this merging property will be very helpful.

Methodology for adjudicated rare endpoints. the statistical analysis challenge is most acute when there are a number of un-adjudicated candidate events at the time of analysis. We will initially make the fairly reasonable assumption that all true events are observed as candidate events. If this is not the case, then the sampling and study design need to be assessed carefully and surveillance may need to be increased in its vigilance to reduce the chance of any false-negatives. Otherwise, some sort of estimate of false-negative rate will need to be obtained and the statistical model may need to be adjusted some. In addition to assuming that there are no false-negatives, we will also assume that the probability of conformation or refutation of a candidate event from adjudication is the same for both adjudicated and un-adjudicated endpoints after adjustment for observed confounders. This latter assumption (corresponding to Assumption II in Section 2.1 of Cook and Kosorok, 2004) can be checked using the procedure given in Section 2.5 of Cook and Kosorok. The basic idea of the method is to estimate the counting process for the true confirmed events with jumps of size 1 at all confirmed adjudicated events and to use jumps of size equal to the estimated probability of being a confirmed event for all un-adjudicated events. The probability estimated is computed from a fitted model based on only the adjudicated events. This method has been successfully applied to non-rare time-to-event data for the log-rank statistics, the Cox model and the Kaplan-Meier estimator. Delicate empirical process arguments were used because standard counting process techniques like martingale methods do not work when jump sizes are estimated.

For rare events, we expect that the techniques used in Cook and Kosorok (2004) can be extended along the lines described in Section **5.4**. For extremely rare events, new empirical process methodology will need to

be developed and evaluated using some of the approaches above. These limiting distributions will in general be quite complicated. However, the parametric bootstrap approach in Kosorok and Song (2007) is applicable in this instance, and thus a special Monte Carlo approach for inference can be developed along these lines.

Methodology for meta-analysis of rare endpoints. how to bring the studies onto a similar footing so that a single, cohesive model can be incorporate for valid scientific inference. Clearly, the approaches we have describe above are already complicated for rare events and even more complicated for extremely rare events. Thus it would appear that trying to perform meta-analysis in this context would be astonishingly difficult. However, if the differences between studies can be summarized by a low-dimensional fixed or random effect, the information about specific drug and risk-factor effects on a specific rare event is larger overall than it is for a single study. This means that estimates can be more precise and that asymptotic results can sometimes be more accurate than happens with single studies. In some case the number of observed extremely rare events in the meta-analysis becomes large enough that the more straightforward techniques for rare events are applicable. In order to utilize meta-analysis in this manner, very careful modeling is needed, as described in Shuster et al. (2007) for rare event meta-analysis studies in heart disease. Provided we can validate such models, the meta-analysis problem becomes essentially equivalent to statistical analysis of a large single study with a modest increase in the number of parameters that need to be estimated. This means that the methods developed above for both rare and extremely rare single studies can be applied.

We plan on exploring appropriate Bayesian approaches as well. We have not discussed this in more detail for a number reasons, including the issue that frequentist properties of Bayesian approaches for semiparametric models under the rare and extremely rare event setting have not been explored and would be somewhat more complicated to evaluate than the proposed frequentist approaches. On the other, Bayesian approaches have been shown in some instances to perform better than their frequentist counterparts for semiparametric inference (see, e.g., Cheng and Kosorok, 2008). We will explore the Bayesian approach in greater detail if we have time.

Evaluation by simulation study. scenarios will be used to evaluate the proposed methods for both rare and extremely rare events, including sequences of models that transition from the rare to the extremely rare case. We will utilize Core C for assistance in developing cod for these simulation studies. One special challenge for rare events—and especially for extremely rare events—is that the simulated studies may need to be extremely large in order to arrive at the right number of events. This can be very computationally expensive. Fortunately, the reverse time Markov model approach suggested in Frater et al. (1989) should significantly decrease the computational costs of such simulation studies. We will carefully evaluate both the accuracy of theoretical predictions as well as performance of the proposed methods under a broad range of simulation scenarios. We will use suitable adaptations of the simulation study design in Section 3 of Cook and Kosorok (2004).

Evaluation by data analysis. We will evaluate the proposed methods by applying the new methods in comparison with other approaches, such as naive approaches for non-rare events, to existing data from Cancer and other disease. The acquisition of such data will be accomplished through Core B. Through this core, we have access to data from clinical cancer studies conducted at both the Lineberger Comprehensive Cancer Center at UNC-CH and the Duke Comprehensive Cancer Center. We also have access to the full collection of completed Cancer and Leukemia Grade B (CALGB) clinical trials as well as limited access to the United Health Care claims data base as well as both the ACORN and SEER data. We have began evaluating rare events methods from a post-submission safety study of vaccine efficacy based on United Health Care claims data. This last data set involves both extremely rare events as well as adjudication and entails about 60,000 person years of follow-up for an adverse event incidence of about 1 per 2,000 person years. This means that about 30 events are observed. This is quite sufficient for evaluation of extremely rare event methodology since a Poisson random deviate with 30 expected events is quite close to being Gaussian distributed. We will perform data analyses to evaluate performance and computational properties of the proposed methods and to refine the software implementations until they are robustly usable by practitioners.

5.5 Identifying Optimal Individualized Therapies from Existing Clinical Trial Data Using Meta-Analysis, Utility Functions, Classification and Regression

Formulating the statistical problem. Work to date has lead to the development of a statistical foundation for the optimal dosing problems encountered in cancer and other chronic disease studies with competing priorities. In practice, one may have a set of treatments under consideration, which may be coded in a covariate vector X. This covariate may specify both the choice of treatment regime as well as the dose levels and treatment

duration and timing within that regime. The ultimate goal is to determine optimal choices of X, controlling for other patient specific risk factors, which may be coded in a covariate vector Y. The goal is to optimize patient outcomes Z by manipulating the controllable treatment X adjusting for the effects of Y. In general, the outcome Z and the treatment and risk factor vectors X and Y may be multivariate.

In IBCSG Trial VI, X would contain one component for whether the initial CMF was given for 3 months or 6 months and a second component for whether there was reintroduction of chemotherapy following the first cycles. Note that in other settings, for example, meta-analysis, where multiple studies are being combined, the components of X may capture dose levels, which may vary across study. As discussed below, the inferential issues may differ according to whether treatment is discrete or continuous. Patient risk factors, like ER status, age, number of positive nodes, would be included in Z. The outcome variable Z, say, would involve data related to the primary endpoints, such as OS, DFS, and QOL.

The basic idea is to formulate regression models which relate Z to (X, Y) and then optimize some criterion of interest, that is, a utility function, over $\mathbf{X}$ for a fixed value of $Y = y$. This optimal value of $\mathbf{X}$ is denoted by $\mathbf{X}^*(y)$. Of course, the optimal treatment will generally depend on the value of the patient specific risk factors. This would especially be the case if there are interactions between X and Y. This might occur in IBCSG VI, for example, if the effectiveness of treatments varies by country, in which case the residency of the patient might strongly influence the treatment decision. Such interactions can be formulated in the context of the regression models, or via more complex classification and learning techniques, such as CART, neural networks, and SVM (Hastie, Tibshirani, Friedman, 2001).

To be more concrete, we now consider IBCSG VI in greater detail. For DFS and OS, proportional hazards models might be fit. If only one of these outcomes is of interest, one might employ the appropriate overall or disease free survival probabilities at a time point of interest, like 10 years post treatment. Such analyses are commonly employed to determine treatment utility in cancer and other chronic disease studies, such as psychiatric disorders in the elderly, like Alzheimer's, as well as diabetes. Alternatively, the outcome might be a biological measurement, potentially measured longitudinally, such as CD4 counts in AIDS studies. prostate serum antigen titers for prostate cancer patients, and blood glucose levels in diabetics. Longitudinal regression models, either mixed effects models or generalized estimating equation approaches could be employed. In AIDS studies, treatment utility might be captured by mean number of CD4 counts at a given time point, or averaged over several time points. In IBCSG VI, the focus is not on biomarkers, but rather on quality of life. A linear regression model could be fit to QOL longitudinal data, with mean QOL maximized at a particular time point, or quality adjusted survival maximized over the entire life course. Finally, cost of care could be analyzed, either as a longitudinal measurement, or cumulative lifetime cost, with goal being the minimization of such costs. The difficulty is how to combine these endpoints in an overall evaluation of the treatments.

To combine models for the endpoints, we propose using a multiattribute utility function (Edwards, 1982). Such utility functions have been widely used in health policy applications, where economic evaluations are needed in understanding population level decision about resource allocation. Major successes have been realized in environmental risk assessment (Wilson and Crouch, 1987) and health care budgeting (Peacock, Richardson, Carter, Edwards, 2007). In the current setting, multiattribute functions are to be used as a means of optimizing individual as opposed to population level treatment policies across a spectrum of relevant endpoints. The scope is quite broad, permitting tradeoffs between multiple complex survival endpoints, biomarkers, quality of life, and costs of care. Adopting this approach enables a formal sensitivity analysis of the effect of changing the relative weights given to the individual endpoints in the composite utility; details are given below.

Modelling framework. Beginning quite generally, we suppose one has a model $f(\theta, x, y)$ which describes the relationship between some aspect of the distribution of Z and the covariates X and Y, where θ is some parameter. Note that because Z may be multidimensional, $f(\theta, x, y) = \{f_1(\theta_1, x, y), \ldots, f_p(\theta_p, x, y)\}$ may have dimension p greater than 1, where $\theta = (\theta_1, \ldots, \theta_p)$, and f_i and θ_i are the model and parameter for endpoint $i(= 1, \ldots, p)$. This permits models $f_1, \ldots, f_p$ which may be parametric, semiparametric, or nonparametric, depending on the dimension and structure of the parameter θ. While the parameter θ includes the parameters from all the models for the individual endpoints, the function f permits separate modelling of the different quantities with separate parameters, and does not require that they be artificially combined into composite endpoint. In IBCSG VI and other cancer studies, a proportional hazards model would ordinarily be fit for survival and disease free survival and f_1 and f_2 might define OS and DFS probabilities at a time point of interest. Moment or mixed effects models might be fit to longitudinal outcomes like QOL and cost of care, with

corresponding f_i's defined for mean values of these quantities at those time points.

The utility function is central to formally combining utility information in the different endpoints. Suppose one has a utility function $U(\theta, x, y)$ which is a known function, differentiable in each of its arguments, and is derived from the models $f_1, \ldots, f_p$. In the simplest case with $p = 1$ where overall survival at some time point τ is the only quantity of interest, the obvious choice for U is the survival function at that time point, where θ would be the parameters in the proportional hazards model. For multivariate f, it is natural to take U to be a weighted average of the utilities associated with the different quantities of interest, that is, $U = \sum_{i=1}^{p} w_i U_i$, where w_i are positive weights summing to 1 and U_i are single attribute utilities involving only $f_i, i = 1, \ldots, p$. More complex nonlinear multiattribute utilities may also be utilized (Olson, 2007). A limitation of such nonlinear utilities is difficulty in interpretation. Moreover, a key aspect in eliciting such utility is the determination of the weighting factors. A pragmatic approach is to view any given choice of w_i's as reflecting certain prior beliefs about the relative importance of different endpoints in decision making. Careful exploration of the impact of such weights on optimal treatment decisions is needed.

The optimal choice of X for given $Y = y$, $X^*(y)$ may now be defined as $X^*(y) = \text{argmin}_{x \in \mathcal{X}} \{U(\theta, x, y)\}$, where $\mathcal{X}$ defines the region of interest for the treatment variables in X. Of course, X^* will vary with y, so that the role of y in treatment decisions will be critical for individual patients. When X is categorical, $\mathcal{X}$ will contain a finite number of values, with the simplest case being that of two treatments, where $\mathcal{X}$ has two atoms. One may restrict $\mathcal{X}$ outside the natural support of X to limit treatment options in the inferential process. Such restrictions may be useful if the investigator knows a priori that certain treatments may not be practicable for logistical and/or ethical reasons. The main complicating factor in the definition of $X^*(y)$ is that θ is unknown and must be estimated from data. This poses inferential difficulties, which cannot be addressed using existing results.

Inferences for optimal treatments. Suppose one has an estimator of θ, $\hat{\theta}$, say. This may be achieved by fitting the models for $\hat{\theta}_1, \ldots, \hat{\theta}_p$, which may be accomplished by fitting them separately for each outcome. One might fit proportional hazards models for OS and DFS using standard partial likelihood techniques and GEE or likelihood analyses for longitudinal data (Diggle, Liang, Zeger, 1994). One might also calculate the parameter estimates by employing joint models, in which a full likelihood analysis is carried out simultaneously for all endpoints. The following results are very general and make weak, high level assumptions about $\hat{\theta}$. The rationale is that one can proceed with all of the usual analyses in order to understand the effects of X and Y on the outcomes. One may then use these results in a second stage utility driven analysis to determine optimal treatment assignments.

One may estimate U with $\hat{U}(x, y) = U(\hat{\theta}, x, y)$. If $\hat{\theta}$ is a "good" estimator for θ, then $\hat{U}$ should be a "good" estimator for U, where "good" remains to be defined rigorously. It seems natural to estimate $X^*(y)$ by $\hat{X}^*(y) = \text{argmin}_{x \in \mathcal{X}} \{\hat{U}(x, y)\}$. A closed form for $\hat{X}^*$ will not exist in general. If x is one dimensional, for example, dose level, then one may conduct a line search to find $\hat{X}^*$, which is very stable. For higher dimensional x, other numerical techniques may be needed. The Nelder-Mead algorithm may be reliably employed and does not require differentiation of U with respect to x, unlike gradient based methods. If U is differentiable in x, then Newton-Raphson type algorithm may be employed and may converge more quickly than the Nelder-Mead algorithm. The properties of $\hat{X}^*(y)$ will depend heavily on those of $\hat{\theta}$. It will be important to develop precise conditions under which $\hat{X}^*(y)$ consistently estimates $X^*(y)$ and under which the limiting sampling distribution of the estimate is normal. Furthermore, for the methods to be useful in practice, an assessment of sampling variability is required, including variance estimation.

To establish consistency of $\hat{X}^*$, we will assume that $\hat{\theta}$ is consistent in probability, uniformly. This enables one to show that $\hat{U}$ converges uniformly in probability to $U(\theta, x, y)$. Now, if one further assumes that X^* is unique for a given y, then the uniform convergence of $\hat{U}$ guarantees consistency of $\hat{X}^*(y)$ for $X^*(y)$ at a fixed value of $Y = y$. This result can be strengthened to establish the uniform consistency of $\hat{X}^*(y)$ for $X^*(y)$ for y in some set of interest. This is a strong result, which implies that regardless the patient specific risk factors, the optimal treatment configuration may be accurately estimated.

To establish asymptotic normality, additional conditions are needed. We will assume that as the size of the sample used to estimate θ, n, gets large, that $n^{1/2}(\hat{\theta} - \theta)$ is asymptotically mean normal with covariance Σ which may be consistently estimated by $\hat{\Sigma}$. Now, assuming $\mathcal{X}$ is compact subset, as would occur if X is continuous, like dose, and $X^*(y)$ is in the interior of $\mathcal{X}$, then a Taylor like expansion of the estimated utility function gives that for a fixed y, $n^{1/2}\{\hat{X}^*(y) - X^*(y)\}$ has approximately the same distribution as $[-\partial^2\{U(\theta, X^*(y), y)\}/\partial\theta\partial x]^{-1} \times$

$n^{1/2}(\hat{\theta} - \theta)$. It then follows from Slutsky's law that $n^{1/2}\{\hat{X}^*(y) - X^*(y)\}$ is also mean zero normal with variance $\Gamma = [-\partial^2\{U(\theta, X^*(y), y)\}/\partial\theta\partial]^{-1}\Sigma[-\partial\partial\{U(\theta, X^*(y), y)\}/\partial\theta\partial x]^{-1}$. The variance Γ may be estimated by $\hat{\Gamma}$ in which θ and Σ are replaced by $\hat{\theta}$ and $\hat{\Sigma}$. This involves no additional work beyond that needed to estimate θ and Σ in the original analyses. Weak convergence of the limit distribution to a tight Gaussian process may be further established under regularity conditions needed to ensure applicability of advance empirical process theory (van der Vaart and Wellner, 1996). Confidence intervals for $X^*(y)$ may be constructed accordingly using the plug-in variance estimates, with confidence bands possible by bootstrapping the data used to estimate θ.

A difficulty arises with categorical treatment covariate X, since even though U is differentiable, the utility is not "differentiable" at $X^*(y)$ when the support of X, $\mathcal{X}$ is discrete. The usual development of asymptotic theory for this scenario is unclear, since it relies heavily on differentiability in a local parameter space around the true value of the parameter being estimated. The development of variance estimators, confidence intervals and confidence bands is unclear. Moreover the use of the bootstrap in such irregular settings has not been theoretically justified and its deficiencies have been explored empirically with discrete parameter spaces (Newton, 1996).

To evaluate the effect of small changes in the weights w_i in the utility function, we recommend exploring changes in $\hat{X}^*(y)$ as the weights are varied. Such a sensitivity analysis may be useful in understanding how greater relative weights placed on certain endpoints may influence the optimal dose. This may perhaps be achieved with a few sets of weights, representing different perspectives on the decision making problem.

Numerical studies. Extensive simulation studies will be conducted to evaluate the performance of the inferential procedures in realistic samples. Data will be generated based on fitting models to IBCSG VI and other available cancer studies at the Lineberger Cancer Institute and at CALGB at Duke obtained through Core B. This will ensure that the simulations match scenarios which are commonly encountered in oncology. Small, moderate, and large sample sizes will be considered. The goal is to evaluate the bias and variance of the estimates, as well as the empirical coverage of the confidence interval procedures. Extensive analyses of IBCSG VI and other cancer studies at UNC and Duke will be undertaken to develop experience in applying the utility based methodology. The sensitivity of the resulting optimal treatments to the choice of utility function will be assessed, varying the weights to emphasize different endpoints. Such sensitivity analyses will be useful in highlighting the potential tradeoffs amongst survival, quality of life, and cost, the three main considerations in the decision making process.

5.6 Software Implementation and Dissemination and Timetable

Our overall software development goal is to develop user friendly software so that the proposed methods can be implemented easily, accurately and at a reasonable cost both by us and other investigators. Each of the proposed new methods will first undergo a developmental implementation until the procedure has been validated via simulation studies and data analyses. Core C will then assist with taking the implementation and developing it into a usable and robust software package appropriate for dissemination in both R and SAS formats. Guidebooks and web pages of instructions and examples will also be developed and the new software will be communicated through presentations and short courses at appropriate professional meetings. We will also identify several beta testers outside of our institutions to help refine the quality of the software implementations.

Each of the five aims will require about one year of development and theoretical work (which may involve simulations and data analyses to guide the theory), about 6 months of simulation and data analysis evaluation, and 6 months to 1 year of implementation and development. We will stagger the initiation of each phase so that they are a year apart, so that the theoretical work can be done sequentially. The theory for the adjudication component may take 1.5 years. This means that the results of the first phase will be completed by the middle of year 2, and that we should be able to complete the first three phases and start the fourth phase before the end of year 5 of the grant.

6 INCLUSION ENROLLMENT REPORT

N/A

7 BIBLIOGRAPHY AND REFERENCES CITED

Adcock, C. J. (1997). Sample size determination: a review. *The Statistician* **46**, 261–283.

Ahn, J., Peters, U., Albanes, D., Purdue, J. P., Abnet, C. C., Chatterjee, N., Horst, R. L., Hollis, B. W., Huang, W.-Y., Shikany, J. M., and Hayes, R. B., for the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Project Team (2008). Serum vitamin D concentration and prostate cancer risk: A nested case-control study. *Journal of the National Cancer Institute* **100**, 796–804.

Andersen, S. (1997). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* **145**, 290.

Aitkin, M. (1999). Meta-analysis by random effect modeling in generalized linear models. *Statistics in Medicine* **18**, 2343–2351.

Berlin, J. A., and Colditz, G. A. (1999). The role of meta-analysis in the regulatory process for foods, drugs and devices. *Journal of the American Medical Association* **28**, 830–834.

Biggerstaff, B. J., and Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16**, 753–768.

Brockwell, S. E., and Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine* **20**, 825–840.

Böhning, D., and Patilea, V. (2008), A capture-recapture approach for screening using two diagnostic tests with availability of disease status for the test positives only, *Journal of the American Statistical Association* **103**, 212-221.

Burnham, K.P., and Anderson, D.R. (1998). Model Selection and Inference: A Practical Information Theoretic Approach. Springer-Verlag: New York.

Burr, D., and Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association* **100**, 242–251.

Buyske, S., Fagerstrom, R., and Ying, Z. (2000). A class of weighted log-rank tests for survival data when the event is rare. *Journal of the American Statistical Association* **95**, 249–258.

Buzas, J. S., and Stefanski, L. A. (1996). A note on corrected-score estimation. *Statistics and Probability Letters* **28**, 1–8.

Cauley, J. A., Norton, L., Lippman, M. E., Eckert, S., Krueger, K. A., Purdie, D. W., Farrerons, J., Karasik, A., Mellstrom, D., Ng, K. W., Stepan, J. J., Powles, T. J., Morrow, M., Costa, A., Silfen, S. L., Walls, E. L., Schmitt, H., Muchmore, D. B., and Jordan, V. C. (2001). Continued breast cancer risk reduction in postmenopausal women treated with raloxifene: 4-year results from the MORE trial. *Breast Cancer Research and Treatment* **65**, 125–134.

Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology* **158**, 280–287.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2006). Posterior propriety and computation for the Cox regression model with applications to missing covariates. *Biometrika* **93**, 791–807.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2002). *Monte Carlo Methods in Bayesian Computation.* New York: Springer.

Chen, Q., Ibrahim, J. G., Chen, M.-H., and Senchaudhuri, P. (2008). Theory and inference for regression models with missing response and covariates. *Journal of Multivariate Analysis* **99**, 1302–1331.

Chen, S., Watson, P., and Parmigiani, G. (2005). Accuracy of MSI testing in predicting germline mutations of MSH2 and MLH1: a case study in Bayesian meta-analysis of diagnostic tests without a gold standard. *Biostatistics* **6**, 450–464.

Cheng, G., and Kosorok, M. R. (2008). Higher order semiparametric frequentist inference with the profile sampler. *Annals of Statistics* **36**, 1786–1818.

Chu, H., and Cole, S. R. (2006).

Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* **59**, 1331–1332.

Chu, H., and Guo, H. (2009). Letter to the Editor on "A unification of models for meta-analysis of diagnostic accuracy studies" by Harbord et al." *Biostatistics* **10**, 201–203.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10**, 101-129.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.

D'Agostino Sr., R. B., Massaro, J. M., and Sullivan, L. M. (2003). Non-inferiority trials: design concepts and issues – the ecounters of academic consultants in statistics. *Statistics in Medicine* **22**, 169–186.

Davidian, M. (2002). Dose calibration. In *Encyclopedia of Environmentrics*, New York: Wiley. pp. 562-565.

Davidian, M., and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data.* Boca Raton: Chapman and Hall/CRC.

Davidian, M., and Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: An over-view and update. *Journal of Agricultural Biological and Environmental Statistics* **8**, 387–419.

Dendukuri, N., and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* **57**, 158–167.

Diggle, P., Liang, K. Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Edwards, D. (1982). *Multiattribute Evaluation.* Beverly Hills: Sage.

Egger, M., Smith, G. D., and Altman, D. G. (2001). *Systematic Reviews in Health Care: Meta-analysis in Context.* Second Edition. BMJ Publishing Group.

Eng, K. H., and Kosorok, M. R. (2005). A sample size formula for the supremum log-rank statistic. *Biometrics* **61**, 86–91.

Fine, J. P., and Gelber, R. D. (2001). Joint regression analysis of survival and quality adjusted survival. *Biometrics* **57**, 376–382.

Fleming, T. R. (2008). Current issues in non-inferiority trials. *Statistics in Medicine* **27**, 317–332.

Fleming, T. R., and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Frater, M. R., Bitmead, R. R., Kennedy, R. A., and Anderson, B. D. O. (1989). Rare events and reverse-time models. *Proceedings of the 28th Conference on Decision and Control*, Tampa, FA.

Gart, J. J., and Buck, A. A. (1966). Comparison of a screening test and a reference test in Epidemiologic Studies II, a probabilistic model for comparison of diagnostic tests. *American Journal of Epidemiology* **83**, 593.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 515–534.

Glasziou, P. P., Simes, R. J. and Gelber, R. D. (1990). Quality adjusted survival analysis. *Statistics in Medicine* **9**, 1259–1276.

Gohagan, J. K., Prorok, P. C., Kramer, B. S., and Cornett, J. E. (1994). Prostate cancer screening in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial of the National Cancer Institute. *Journal of Urology* **152**, 1905–1909.

Harbord, R. M. et al (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* **8**, 239–251.

Hardy, R. J., and Thompson, S. G. (1996). A likehood approach to meta-analysis with random effects. *Statistics in Medicine* **15**, 619–629.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Staistical Learning*. New York: Springer.

Heinze, G., Gnant, M., and Schemper, M. (2003). Exact log-rank tests for unequal follow-up. *Biometrics* **59**, 1151–1157.

Higgns, J. P., and Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* **15**, 2733–2749.

Hirji, K. F., Mehta, C. R., and Patel, N. R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association* **82**, 1110–1117.

Hoskins, J. M., Goldberg, R. M., Qu, P, Ibrahim, JG, and McLeod, H. L. (2007). UGT1A1*28 and irinotecan-induced neutropenia: dose matters. *Journal of the National Cancer Institute* **99**, 1290–1295.

Huang, L., Chen, M.-H., and Ibrahim, J. G. (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics* **61**, 767–780.

Hui, S. L., and Walter, S. D. (1980). Estimating the error rates of diagnostic-tests. *Biometrics* **36**, 167–171.

Hung, H. M. J., Wang, S.-J., and O'Neill, R. T. (2007). Issues with statistical risks for testing methods in noninferiority trial without a placebo arm. *Journal of Biopharmaceutical Statistics* **17**, 201–213.

Hung, H. M. J., Wang, S.-J., and O'Neill, R. T. (2005). A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal* **47**, 28–36

Hung, H. M. J., Wang, S.-J., Tsong, Y., Lawrence, J., and O'Neill, R. T. (2003). Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* **22**, 213–225.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88**, 551–564.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer.

Ibrahim, J. G., Chen, M.-H., and Lipstiz S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics 55*, 591–596.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Series B* **61**, 173–190.

Inoue, L. Y. T., Berry, D. A., and Parmigiani, G. (2005). Relationship between Bayesian and frequentist sample size determination. *The American Statistician* **59**, 79–87.

International Breast Cancer Study Group (1996). Duration and reintroduction of adjuvant chemotherapy for node positive premenopausal breast cancer patients. *Journal of Clinical Oncology* **14**, 1885-1894.

Johnson, W. O., Gastwirth, J. L., and Pearson, L. M. (2001). Screening without a "gold standard": The Hui-Walter paradigm revisited. *American Journal of Epidemiology* **153**, 921–924.

Joseph, L., and Bélisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *The Statistician* **46**, 209–226.

Joseph, L., Gyorkos, T. W., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* **141**, 263–272.

Joseph, L., Wolfson, D. B., and Du Berger, R. (1995a). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician: Journal of the Institute of Statisticians* **44**, 143–154.

Joseph, L., Wolfson, D. B., and Du Berger, R. (1995b). Some comments on Bayesian sample size determination. *The Statistician: Journal of the Institute of Statisticians 44*, 167–171.

Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40**, 214–221.

Katsis, A., and Toman, B. (1999). Bayesian sample size calculations for binomial experiments. *Journal of Statistical Planning and Inference* **81**, 349–362.

Kieser, M., and Friede, T. (2007). Planning and analysis of three-arm non-inferiority trials with binary endpoints. *Statistics in Medicine* **26**, 253–273.

King, G., and Zeng, L. (2000). Logistic regression in rare events data. The Global Burden of Disease 2000 in Aging Populations, Research Paper No. 2. Harvard University.

Kleinman, K. P., and Ibrahim, J. G. (1998a). A semiparametric Bayesian approach to the random effects model. *Biometrics* **54**, 921–938.

Kleinman, K. P., and Ibrahim, J. G. (1998b). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine* **17**, 2579–2596.

Kong, F. H., and Slud, E. (1997). Robust covariate-adjusted logrank tests. *Biometrics* **84**, 847–862.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Kosorok, M. R., and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics* **35**, 957–989.

Lam, Y., and Lam, C. V. (1997). Bayesian double-sampling plans with normal distributions. *The Statistician* **46**, 193–207.

Li, Z. (2001). Covariate adjustment for non-parametric tests for censored survival data. *Statistics in Medicine* **20**, 1843–1853.

Li, Z., Zhou, S., Choubey, S., and Sievenpiper, C. (2007). Failure event prediction using the Cox proportional hazard model driven by frequent failure signatures. *IEEE Transactions* **39**, 303–315.

Lin, D. Y. (2003). Regression analysis of incomplete medical cost data. *Statistics in Medicine* **22**, 1181–2000.

Lin, D. Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53**, 419–430.

Lin, D. Y., and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84**, 1074–1078.

Lindley, D. V. (1997). The choice of sample size. *The Statistician* **46**, 129–138.

Lipsitz, S. R., and Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika* **83**, 916-922.

Lopes, H. F., Müller, P., and Rosner, G. L. (2003). Bayesian meta-analysis for longitudinal data models using multivariate mixture priors. *Biometrics* **59**,66–75.

Lucas, T., Beresford, J., and Vessey, M. (1993). Delay and completeness of cancer registration in a post-marketing surveillance study of cimetidine. *Pharmacoepidemiology and Drug Safety* **2**, 201–203.

Mehta, C. R., and Patel, N. R. (2000). *StatXact4 for Windows User Manual.* Cambridge, Massachusetts: Cytel Software.

M'Lan, C. E., Joseph, L., and Wolfson, D. B. (2008). Bayesian sample size determination for bionomial proportions. *Bayesian Analysis* **3**, 269–296.

M'Lan, C. E., Joseph, L., and Wolfson, D. B. (2006). Bayesian sample size determination for case-control studies. *Journal of the American Statistical Association* **101**, 760–772.

Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Mukherjee, D., Nissen, S. E., and Topol, E. J. (2001). Risk of cardiovascular events associated with selective COX-2 inhibitors. *Journal of the American Medical Association* **286**, 954–959.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, **65**, 331–366.

Murthy VH, Krumholz HM, Gross CP. (2004). Participation in cancer clinical trials: race-, sex-, and age-based disparities. *JAMA*, **291**, 2720–2726.

Newton, M. A. (1996). Bootstrapping phylogenies: large deviations and dispersion effects. *Biometrika* **83**, 315–328.

Normand, S. L. (1999). Tutorial in Biostatistics meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18**, 321–359.

Olson, D. L. (2007). Review of empirical studies in multiobjective mathematical programming: subject reflection of nonlinear utility and learning. *Decision Sciences* **23**, 1–20.

Peacock, S. J., Richardson, J. R. J., Carter, R., and Edwards, D. (2007). Priority setting in health care using multiattribute utility theory and programme budgeting and marginal analysis. *Social Science in Medicine* **64**, 897–910.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press.

Pham-Gia, T. On Bayesian analysis, Bayesian decision theory and the sample size problem. *The Statistician* **46**, 139–144.

Qu, P., Chu, H., Ibrahim, J. G., Peacock, J., Shen, X. J., Tepper, J., Sandler, R. S., and Keku, T. O. (2008). Statistical strategies to improve the efficiency of molecular studies of colorectal cancer prognosis. *British Journal of Cancer* **99**, 2001–2005.

Qu, Y., and Hadgu, A. (1998). A model for evaluating sensitivity and specificity for correlated diag-nostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association* **93**, 920–928.

Qu, Y., Tan, M., and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52**, 797–810.

Rahme, E., and Joseph, L. (1998). Exact sample size determination for binomial experiments. *Journal of Statistical Planning and Inference* **66**, 83–93.

Reitsma, J. B., et al (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* **58**, 982–990.

Riley, R. D., Thompson, J. R., and Abrams, K. R. (2008). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* **9**, 172–186.

Rothmann, M., Li, N., Chen, G., Chi, G. Y. H., Temple, R., and Tsou, H.-H. (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* **22**, 239–264.

Rubin, D. B., and Stern, H. S. (1998). Sample size determination using posterior predictive distributions. *Sankhyā, Series B* **60**, 161–175.

Rutter, C. A., and Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* **20**, 2865–2884.

Sargent, D.J., Zee, B. C., Milan, C., Torri, V., and Francini, G. (2000). Meta-analysis of individual-patient survival data using random-effect models. In *Meta-Analysis in Medicince and Health Policy* (Eds. D. K. Stangl and D. A. Berry). New York: Marcel Dekker, Inc. pp. 255–275.

Shen, Y., Wu, D. F., and Zelen, M. (2001). Testing the independence of two diagnostic tests. *Biometrics* **57**, 1009–1017.

Shuster, J. J., Jones, L. S., and Salmon, D. A. (2007). Fixed vs. random effects meta-analysis in rare event studies: The Rosiglitazone link with myocardial infarction and cardiac death. *Statistics in Medicine* **26**, 4375–4385.

Simon, R. (1999). Bayesian design and analysis of active control clinical trials. *Biometrics* **55**, 484–487.

Sinha, D., Ibrahim, J. G., and Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika* **90**, 629–641.

Smith, P. B., Benjamin, D. K. Jr., Murphy, M. D., Johann-Liang, R., Iyasu, S., Gould, B., Califf, R. M., Li, J. S., and Rodriguez, W. (2008). Safety monitoring of drugs receiving pediatric marketing exclusivity. *Pediatrics* **122**, e628–e633.

Song, R., Kosorok, M. R., and Fine, J. P. (2008). On asymptotically optimal tests under loss of identifiability in semiparametric models. *Annals of Statistics*, In press. NIHMSID: 75080.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 583-639.

Sutton, A. J., and Higgins, J. P. (2008). Recent developments in meta-analysis. *Statistics in Medicine* **27**, 625–650.

Taleb, N. N. (2008). The fourth quadrant: A map of the limits of statistics. *Edge* **257**, online publication at www.edge.org.

Temple, R. (1999). Meta-analysis and epidemiologic studies in drug development and postmarketing surveillance. *Journal of the American Medical Association* **281**, 841–844.

Tsiatis, A. A., and Davidian, M. (2004). Joint modeling of longitudinal and time to event data. *Statistica Sinica* **14**, 793–818.

Torrance-Rynard, V. L., and Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* **16**, 2157–2175.

Vacek, P. M. (1985). The Effect of Conditional Dependence on the Evaluation of Diagnostic-Tests. *Biometrics* **41**, 959–968.

van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* New York: Springer.

van Houwelingen, H. C., Arends. L. R., and Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* **21**, 589–624.

Vonesh, E. F., and Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* New York: Marcel Dekker.

Wang, F., and Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* **17**, 193–208.

Wang, S.-J., and Hung, H. M. J. (2003a). Assessing treatment efficacy in noninferiority trials. *Controlled Clinical Trials* **24**, 147–155.

Wang, S.-J., and Hung, H. M. J. (2003b). TACT method for non-inferiority testing in active controlled trials. *Statistics in Medicine* **22**, 227–238.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician* **46**, 185–191.

Whitehead A., and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* **10**, 1665–1677.

Wilson, R., and Crouch, E. A. (1987). Risk assessment and comparisons: an introduction. *Science* **5**, 236–267.

Xu, J., and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and time to events. *Applied Statistics* **50**, 375–387.

Zhou, X. H., Obuchowski, N. A., and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine.* New York: John Wiley and Sons.

Zwinderman, A. H., and Bossuyt, P. M. (2008). We should not pool diagnostic likelihood ratios in systematic reviews. *Statistics in Medicine* **27**, 687–697.

8 PROTECTION OF HUMAN SUBJECTS

Although the proposed research indirectly involves human subjects through the preparation, in Core B, of de-identified data sets from identifiable patient data sources, the investigators on Project 3 will have access only to the de-identified data. Thus, the investigators on Project 3 will have no access to any identifiable patient information.

9 INCLUSION OF WOMEN AND MINORITIES

The methods we develop will be applicable to studies with both women and minorities and also to studies which examine treatment differences adjusted for gender, ethnicity and race. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, many of the existing data sets to be studied and provided by Core B include women and minorities, although we will not be generating any new data involving human subjects.

10 TARGETED/PLANNED ENROLLMENT TABLE

N/A

11 INCLUSION OF CHILDREN

The methods we develop will be applicable to studies with children and also to studies which examine treatment differences adjusted for age. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, some of the existing data sets to be studied and provided by Core B may include children, although we will not be generating any new data involving human subjects.

12 VERTEBRATE ANIMALS

N/A

13 SELECT AGENT RESEARCH

N/A

14 MULTIPLE PD/PI LEADERSHIP PLAN

N/A

15 CONSORTIUM/CONTRACTUAL ARRANGEMENTS

If the present application is funded, the University of North Carolina at Chapel Hill will execute a subcontract with the consortium institution (North Carolina State University). The inter-institutional agreement will be written consistent with the NIH consortium agreement policy.

16 LETTERS OF SUPPORT - None

17 RESOURCE SHARING PLAN(S)

(a) Data sharing plan: The data-related resources generated by the proposed research consists of new statistical methodology, software packages for implementation of the methodology, and tutorials for the software. The statistical methodology will be shared through peer reviewed publications and national meetings and

through other standard means. All accepted publications will be deposited in PubMed Central in accordance with the NIH Public Access Policy. Summaries of the methodology, the software and tutorials will be shared through a public web site managed by Core A, while Core C will assist in preparation of the software and tutorials for dissemination. This project will use de-identified data prepared by Core B to test the methods and to create demonstrations of use of the methods to be included in tutorials. This project will not be involved in sharing of these data; this function will be addressed by Core B.

(b) Sharing model organisms: N/A

(c) GWAS: N/A

PROJECT 4

METHODS FOR PHARMACOGENOMICS AND INDIVIDUALIZED THERAPY TRIALS

Project Leader: Danyu Lin, PhD

PROJECT SUMMARY (See instructions):

The broad, long-term objectives of this research are the development of novel and high-impact statistical and computational tools for discovering genetic variants associated with inter-individual differences in the efficacy and toxicity of cancer medications and for optimizing drug therapy on the basis of each patient's genetic constitution. The specific aims include: (1) construction of robust and efficient statistical methods for assessing the effects of SNP genotypes and haplotypes on drug response with a variety of phenotypes (e.g., binary and continuous measures of efficacy and toxicity, right-censored survival time, interval-censored time to disease progression, and informatively censored PSA levels and adverse events); (2) development of statistical and data-mining techniques for predicting drug response based on high-dimensional, highly correlated genomic data and complex phenotypes; (3) investigation of statistical procedures for providing low-bias estimation of effect sizes with complex and highly multivariate genetic data for follow-up and confirmation studies; (4) exploration of a new form of machine learning for identifying candidate individualized therapies in both pre-clinical studies and clinical trials. All these aims have been motivated by the investigators' applied research experiences and address the most timely and important issues in pharmacogenomics and individualized therapy. The proposed solutions are built on sound statistical and data-mining principles. The theoretical properties of the new methods will be established rigorously via modern empirical process theory and other advanced mathematical arguments. Efficient and stable numerical algorithms will be devised to implement the new methods. Extensive simulation studies will be conducted to evaluate the operating characteristics of the new inferential and numerical procedures in realistic settings. Applications will be provided to a large number of cancer studies, most of which are carried out at Duke University and the University of North Carolina at Chapel Hill. Practical and user-friendly software will be developed and disseminated freely to the general public. Our research will change the ways pharmacogenomic studies and individualized therapy trials are designed and analyzed, which will lead to optimal treatments for patients in cancer and other diseases.

RELEVANCE (See instructions):

The proposed research will develop new statistical methods that will significantly improve the ways pharmacogenomic studies and individualized therapy trials are designed and analyzed. This will improve public health by hastening the discovery of better treatments for patients in cancer and in other diseases.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location			
Organizational Name: The University of North Carolina at Chapel Hill			
DUNS: 608195277			
Street 1: Office of Sponsored Research, CB #1350		Street 2: 104 Airport Dr., Suite 2200	
City: Chapel Hill	County: Orange		State: NC
Province:	Country: USA		Zip/Postal Code: 27599-1350
Project/Performance Site Congressional Districts: NC-004			

Additional Project/Performance Site Location			
Organizational Name: North Carolina State University			
DUNS: 042092122			
Street 1: Research Admin/ SPARCS		Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514	
City: Raleigh	County: Wake		State: NC
Province:	Country: USA		Zip/Postal Code: 27695-7514
Project/Performance Site Congressional Districts: NC-02			

Program Director/Principal Investigator (Last, First, Middle): Kosorok, Michael R., et al.

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location

Organizational Name: Duke University

DUNS: 044387793

Street 1: Hock Plaza | Street 2: Box 2716 Med Ct.

City: Durham | County: Durham | State: NC

Province: | Country: USA | Zip/Postal Code: 27705

Project/Performance Site Congressional Districts: NC-004

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Lin, Danyu	DANYU_LIN	UNC-CH	Project 4 Leader
Auman, J. Todd		UNC-CH	Co-Investigator
Bondell, Howard D.		NC State University	Co-Investigator
Febbo, Philip G.	FEBBO001	Duke University	Co-Investigator
Harpole, David H.	HARPO002	Duke University	Co-Investigator
Jung, Sin-Ho	Jung0005	Duke University	Project Co-Leader
Kosorok, Michael R.	Michael_Kosorok	UNC-CH	Project Co-Leader
Liu, Yufeng		UNC-CH	Co-Investigator
McLeod, Howard L.	Hmcleod	UNC-CH	Co-Investigator
Owzar, Kouros	KOWZAR	Duke University	Project Co-Leader
Pang, Herbert	Oxbert	Duke University	Co-Investigator
Tzeng, Jung-Ying	jytzeng	NC State University	Co-Investigator
Wang, Wei	wei_wang	UNC-CH	Co-Investigator
Wright, Fred A.	Fred_Wright	UNC-CH	Co-Investigator
Zeng, Donglin	Donglin_Zeng	UNC-CH	Co-Investigator
Zhang, H. Helen		NC State University	Co-Investigator

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project

Human Embryonic Stem Cells ☒ **No** ☐ **Yes**

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

RESEARCH PLAN

1 INTRODUCTION TO RESUBMISSION/REVISION APPLICATION - N/A

2 SPECIFIC AIMS

There is an enormous current interest in identifying genetic determinants of inter-individual differences in the efficacy and toxicity of cancer medications and in tailoring treatment regimens to each patient's genomic profile. The volume and complexity of data from these pharmacogenomic studies and individualized therapy trials pose unique statistical and computational challenges. The broad, long-term objectives of this research are to develop novel and high-impact statistical methods and computational tools for the designs and analysis of such cancer studies. We will focus on four specific aims:

Aim 1: Construction of robust and efficient statistical methods for assessing the effects of SNP genotypes and haplotypes on drug response. Our methods can handle any phenotypes, including binary and continuous efficacy and toxicity measures, right-censored survival time, interval-censored time to disease progression, and informatively censored PSA levels and adverse events, accommodate population stratification and clinical factors correlated with genetic variables, and allow association analysis at the SNP level (even for SNPs that are not on the genotyping chip), the haplotype level or the gene/pathway level.

Aim 2: Development of statistical and data-mining techniques for predicting drug response based on high-dimensional and highly correlated genomic data. We will develop efficient variable selection procedures for ultra-high dimensional SNP and gene expression data under a variety of parametric and semiparametric regression models for all possible measures of drug response, allowing a hierarchical structure in selecting main effects and interactions and the inclusion of genetic variables at a group level. We will also develop machine learning techniques for classification with variable selection capabilities.

Aim 3: Investigation of statistical procedures for providing low-bias estimation of effect sizes with complex and highly multivariate genetic data for follow-up and confirmation studies. We will explore a conditional likelihood approach for producing low-bias estimation of effect sizes for follow-up and confirmation of effects/predictors. We will also pursue methods for a large number of simultaneous tests and penalized regression techniques for clinical outcomes.

Aim 4: Exploration of machine learning techniques for identifying candidate individualized therapies in both pre-clinical and clinical studies. We will provide a unified framework that combines the discovery power of data mining with the stabilizing influence of statistical inference by creating a new form of machine learning, called "latent supervised learning", which balances the power and flexibility of data mining with the reproducibility of statistical inference. We will develop and validate latent supervised learning for use in both pre-clinical and clinical studies for discovery of candidate individualized therapies for cancer.

In all specific aims, we will establish the theoretical properties of the new methods via advanced mathematical tools, such as modern empirical process theory. We will devise efficient and stable numerical algorithms to implement the new methods. We will conduct extensive simulation studies to assess the operating characteristics of the proposed statistical and numerical methods in realistic settings. We will apply the new methods to a variety of cancer studies, including several ongoing clinical trials in the Cancer and Leukemia Group B. We will develop practical and user-friendly software and actively disseminate it to the broad scientific community. The results of this research have the potential to significantly enhance our understanding of the genetic basis of inter-individual variability in drug response and in discovering effective new individualized therapies to improve the quality and longevity of cancer patients.

3 BACKGROUND AND SIGNIFICANCE

There is tremendous variation in the way patients in cancer and other diseases respond to medications, in terms of host toxicity and treatment efficacy. Although such heterogeneity is potentially attributed to clinical factors (e.g., pathogenesis and severity of the disease, drug interactions, and the patient's age, nutritional status, renal and liver function, and concomitant illnesses), inherited differences in the metabolism and disposition of drugs and genetic polymorphisms in the targets of drug therapy (e.g., receptors) can have even greater influence on the efficacy and toxicity of medications (e.g., Evans and Relling, 1999; Evans and McLeod, 2003). Figure 1

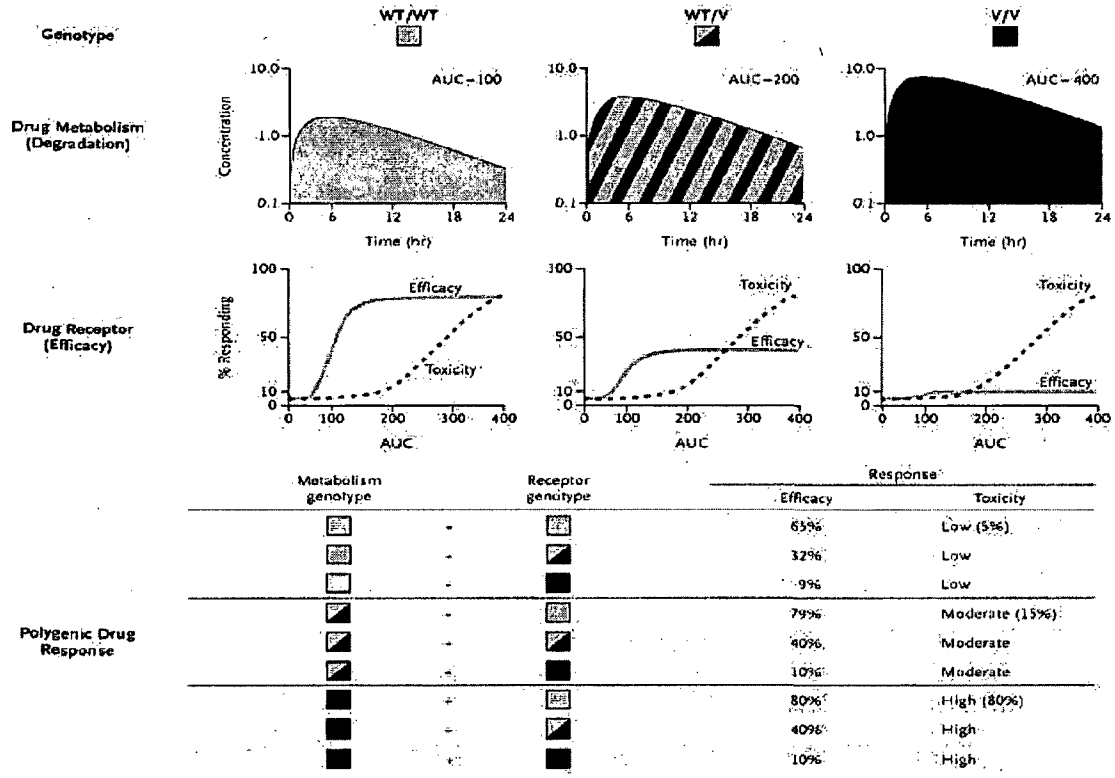


Figure 1: Polygenic determinants of drug response. The potential effects of two genetic polymorphisms are illustrated, one involving a drug-metabolizing enzyme (top) and the second involving a drug receptor (middle), depicting differences in drug clearance (or the area under the plasma concentration-time curve [AUC]) and receptor sensitivity in patients who are homozygous for the wild-type allele (WT/WT), are heterozygous for one wild-type and one variant (V) allele (WT/V), or have two variant alleles (V/V) for the two polymorphisms. At the bottom are shown the nine potential combinations of drug-metabolism and drug-receptor genotypes and the corresponding drug-response phenotypes calculated from data at the top, yielding therapeutic indexes (efficacy:toxicity ratios) ranging from 13 (65 percent:5 percent) to 0.125 (10 percent:80 percent). Reproduced from Evans and McLeod (2003).

illustrates the potential effects of two genetic polymorphisms (one involving a drug-metabolizing enzyme and one involving a drug receptor) on drug response. In cancer, irinotecan-treated patients who are homozygous for the UGT1A1*28 allele have greater risk of hematologic toxic effects than patients who have one or two copies of the wild-type allele (Hoskins et al., 2007), and postmenopausal women taking tamoxifen who have an inherited deficiency in the CYP2D6 gene have increased risk of early breast cancer compared to women without the deficiency (ScienceDaily, 2008).

The study of genetic variation in treatment response is called pharmacogenetics or pharmacogenomics. Pharmacogenetics targets one or at most a few genes, while pharmacogenomics considers the entire genome; however, the two terms have been used interchangeably. Thanks to recent advances in human genome research (e.g., The International Human Genome Sequencing Consortium, 2001; The International HapMap Consortium, 2005) and high-throughput genotying technologies (e.g., Altshuler et al., 2008), the field is now shifting from candidate-gene studies to genome-wide association studies (GWAS), which survey the entire human genome with high-density genotying platforms containing 0.5-1 million single nucleotide polymorphisms (SNPs). There is also a rapid increase in sequence studies, as well as gene expression profiling and proteomic studies.

The genetic polymorphisms that have been established to influence the metabolism and disposition of med-

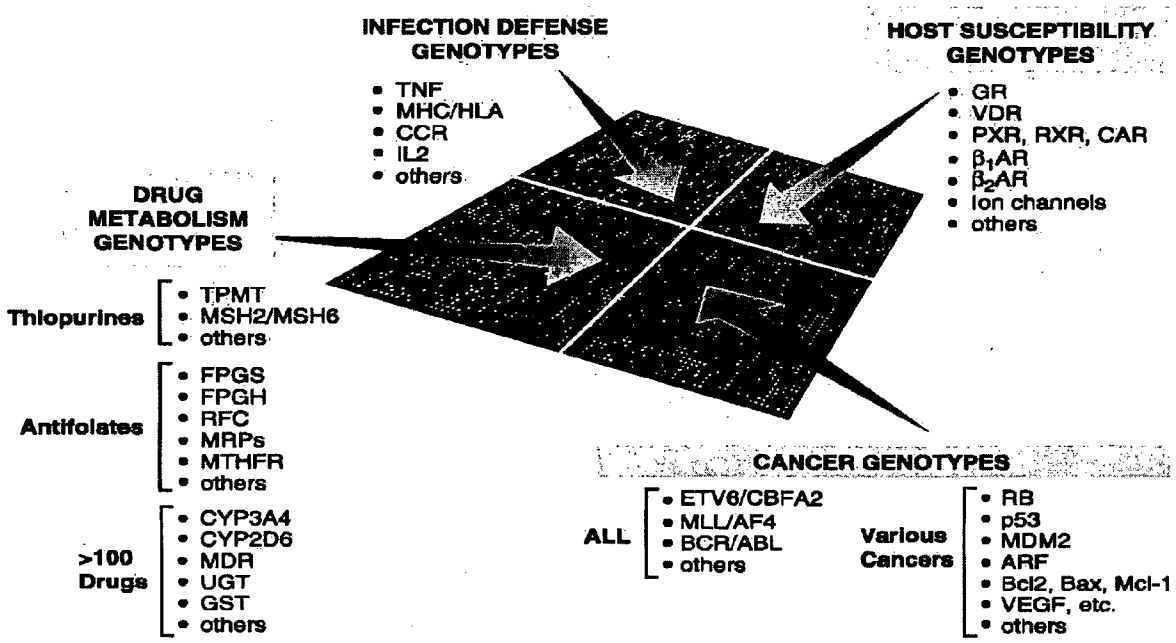


Figure 2: Molecular diagnostics of pharmacogenomic traits. DNA arrays are being made for automated, high-throughput detection of functionally important mutations in genes that are important determinants of drug effects, such as drug-metabolizing enzymes, drug targets (receptors), disease pathogenesis, and other polymorphic genes that influence an individual's susceptibility to drug toxicities or environmental exposures (such as pathogens, carcinogens, and others). This figure exemplifies components of a potential diagnostic DNA array for genes that could influence a patient's response to chemotherapy for acute lymphoblastic leukemia, including genes that determine drug metabolism, disease sensitivity, and the risk of adverse effects of treatment (e.g., cardiovascular or endocrine toxicities, and infections). Reproduced from Evans and Relling (1999).

ications and the targets of drug therapy can be tested for each patient, and such diagnosis can then become the blueprint for individualizing treatments (e.g., Evans and Relling, 1999; Evans and McLeod, 2003). Figure 2 shows how various genes could be genotyped to guide the selection and dosing of chemotherapy for a patient with acute lymphoblastic leukemia. In the near future, it will be technologically and economically feasible to determine the entire DNA sequence for each individual. The potential use of this vast information in individualizing treatments is only limited by our knowledge of the genetic basis for drug disposition and response.

In 2006, the University of North Carolina at Chapel Hill (UNC-CH) established the Institute for Pharmacogenomics and Individualized Therapy (IPIT). IPIT was formed as a collaborative effort of the Schools of Pharmacy, Medicine, Public Health, the Lineberger Comprehensive Cancer Center and the Carolina Center for Genome Sciences. It has launched a number of projects to create effective and precise treatment options for patients suffering from a wide range of conditions, with the focus on creating better tools to assist physicians in medical decision making on the basis of the genetic profile of the individual patient. The institute is directed by the world-renowned pharmacogeneticist Dr. Howard McLeod, who is an investigator of this project. Many of our statistical investigators, particularly Drs. Kosorok and Lin, have already established collaborative relationships with Dr. McLeod.

Pharmacogenomics has been incorporated into an increasing number of clinical trials. This is certainly the case with the clinical trials in the Cancer and Leukemia Group B (CALGB). Currently, CALGB has 108 active trials, 40 of which are treatment trials. Most treatment trials compare standard-chemotherapy to standard-chemotherapy + a biologic agent or another chemo agent. The primary efficacy measures are overall survival and disease progression. For prostate cancer (e.g., protocols 90203 and 90401), PSA levels are also of interest. The toxicity measures of interest include neutropenia, hypertension, neuropathy, and protoneuria. Many CALGB treatment trials have incorporated pharmacogenomic components. Although the pharmacogenomic companion

studies were previously designed to examine a few candidate genes, most of them have switched to GWAS. Gene expression and protein array data are collected in several trials. The CALGB Statistical Center is housed at Duke University, with Dr. Stephen George as the director. Dr. Kouros Owzar, director of the bioinformatics unit, and Dr. Sin-Ho Jung, director of the biostatistics unit, have taken primary responsibilities for the design and analysis of many of the trials. Dr. Howard McLeod is the co-chair of the CALGB Pharmacology and Experimental Therapeutics Committee and is directly involved in 21 pharmacogenomic companion studies.

The data from pharmacogenomic studies and individualized therapy trials are substantially different from those of traditional clinical trials. Indeed, new types of data are being generated every year because of evolving technologies. The volume and complexity of the data have presented enormous statistical and computational challenges. The method development has not been able to keep pace with the collection of the data! In this project, we focus on four specific aims addressing some of the most timely and important issues in pharmacogenomics and individualized therapy. Our research will change the ways pharmacogenomic studies and individualized therapy trials are designed and analyzed, which will lead to optimal treatments for patients in cancer and other diseases.

3.1 Assessing the Effects of SNP Genotypes and Haplotypes on Drug Response

SNPs are DNA sequence variations that occur when a single nucleotide in the genome sequence is altered. SNPs make up about 90% of all human genetic variation and have a major impact on disease susceptibility and drug response. Thanks to the availabilities of dense SNP maps across the human genome and precipitous drops in genotyping costs, SNP-based association studies have gained great popularity in pharmacogenomics. An increasing number of these studies are GWAS.

Missing data present a major challenge in genetic association studies. An important form of missing data arises in the analysis of haplotype effects. A haplotype is a specific sequence of nucleotides on the same chromosome of a subject. Because haplotypes incorporate the linkage disequilibrium information (i.e., correlation structure) of multiple SNPs, the use of haplotypes tends to yield more efficient analysis of association than the use of individual SNPs, especially when the causal SNPs are not directly measured or when multiple mutations occur on the same chromosome (e.g., Judson et al., 2000; Bader, 2001; Schaid, 2004). Unfortunately, current genotyping technologies do not separate a subject's two homologous chromosomes, so that we can only observe the combination of the two haplotypes, which is referred to as the (unphased) genotype.

Missing data are also encountered in the analysis of the effects of individual SNPs. Even with high-quality genotyping, some study subjects will have missing genotypes at certain SNP sites because of assay failures. Genotype data may also be missing by design to reduce genotyping costs. An extreme form of missing data arises when investigators are interested in untyped SNPs, i.e., the SNPs that are not even on the genotyping chip used in the study and are thus missing on all study subjects. Conducting association analysis at untyped SNPs can facilitate the selection of SNPs to be genotyped in follow-up studies and enable investigators to compare or combine results from multiple studies with different genotyping chips.

A number of methods have been proposed to assess haplotype-disease association based on unphased genotype data (e.g., Schaid et al., 2002; Zhao et al., 2003; Epstein and Satten, 2003; Stram et al., 2003; Lake et al., 2003; Lin et al., 2005; Spinka et al., 2005; Lin and Zeng, 2006). In addition, several methods have been developed to analyze untyped SNPs (Nicolae, 2006; Marchini et al., 2007; Lin et al., 2008). The existing literature has been focused primarily on case-control studies. Pharmacogenomic studies differ from case-control studies in several ways. First, the phenotypes tends to be more complex. Survival time is subject to right censoring, whereas time to disease progression is subject to interval censoring in that the progression can only be determined to lie between successive visits; PSA levels and adverse events, as well time to disease progression, can be informatively censored in that the observations on the phenotypes of interest are terminated when the patients die or withdraw from the trial for health-related reasons. Secondly, pharmacogenomics is focused on the interactions of genetic markers and treatments in drug response rather than the (main) effects of genetic markers on disease. Thirdly, there is a strong interest in estimation and prediction, as opposed to hypothesis testing.

In this research, we provide a unified framework for assessing the role of individual SNPs (including untyped SNPs) or their haplotypes in drug response. The effects of genetic markers, treatments and clinical variables (e.g., body size and disease severity) on drug response are formulated through flexible regression models that incorporate appropriate genetic mechanisms and interactions. We construct appropriate likelihoods for

all possible phenotypes (including right-censored survival time, interval-censored time to disease progression, and informatively censored PSA levels and adverse events). We establish the theoretical properties of the maximum likelihood estimators by appealing to modern asymptotic techniques, and develop efficient and stable numerical algorithms to implement the corresponding inference procedures. We also construct test statistics that maximize power under all possible modes of inheritance and derive power/sample size formulas.

There is a heightened recent interest in the use of gene-based methods, which combine information from all SNPs in a gene, pathway or certain region of interest. A major challenge in combining information from multiple SNPs is to ensure that opposite effects would not cancel, which would lead to a loss of power. Indeed, many existing methods, including weighted-sums of genotypes (Wang and Elston, 2007), random-effects methods (Goeman et al., 2004; 2006), and U-statistics (Schaid et al., 2005), would suffer from power loss when different SNPs act in opposite directions (Chapman and Wittaker, 2008; Wei et al., 2008). Furthermore, most methods assume additive effects of SNPs in a gene and cannot handle gene-gene and gene-environment interactions.

In this research, we will use gene-trait similarity regression to aggregate information from SNPs that are in the same gene/pathway. We seek to construct effective statistical procedures that (1) perform joint analysis of all SNPs (possibly non-additive) within a gene/pathway while balancing information and degrees of freedom; (2) incorporate main genetic effects, gene-gene interactions, and gene-environment interactions; and (3) accommodate a broad range of phenotype (quantitative, binary or censored).

The two approaches taken in this aim (i.e., haplotype regression and gene-trait similarity regression) are complementary to each other. Similarity regression tends to be more powerful than single SNP tests and haplotype tests, but it can only detect a global signal. On the other hand, haplotype regression methods can be used to to pinpoint the specific variants that cause the global significance, but may be too "refined" for screening purposes. The combination of the two approaches provides a comprehensive framework to study genetic association at the single SNP level, haplotype level or gene-based level. We will compare the two approaches through extensive simulation studies and apply them to the aforementioned CALGB studies.

3.2 Predicting Drug Response Based on High-Dimensional Genomic Data

The methods to be developed in Aim 1 and indeed virtually all existing methods are focused on one SNP or a few SNPs at a time and do not take full advantages of genomewide data. Although some methods use genomewide information to infer haplotypes and untyped SNPs, the association analysis itself is only done one SNP or a few SNPs at a time. In Aim 2, we will explore methods that relate genomewide data to drug response.

There are strong reasons for considering all the markers or at least a large subset of them simultaneously. The predictive power of a single marker tends to be fairly low. We can improve the accuracy of prediction substantially by utilizing a large number of important markers. The marginal effects of SNPs (i.e., the effect of each SNP on drug response when it is considered alone) may be quite different from their joint effects: (1) a SNP that is not related to drug response but is correlated with the causal SNP tends to be marginally associated with drug response; (2) some SNPs may have weak marginal effects but strong joint effects.

It is very challenging to decide which set of markers should be included in the joint analysis because the number of SNPs in a GWAS is much larger than the sample size. This is commonly referred to as the "small n, large p" problem. A major challenge in this problem is that the number and extent of spurious associations between predictors and response increase rapidly with increasing p.

There is a large body of literature on variable selection methods, including bridge regression (Frank and Friedman, 1993), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), and adaptive lasso (Zou, 2006). However, these methods are designed for a moderate number of predictors (i.e, tens or hundreds). For ultra-high p, these methods may be computationally infeasible and statistically inaccurate.

Recently, Fan and Lv (2008) developed the so-called sure independence screening (SIS) strategy for high dimensional statistical modelling. The idea is to first reduce the dimension from a very large scale to a moderate scale that is below sample size by univariate correlation learning, and then select important predictors by a moderate-scale variable selection method, such as LASSO or SCAD. Fan and Lv (2008) proved that the SIS possesses the sure screening property in that all important predictors survive after variable screening with probability tending to 1 as n approaches ∞. For moderate sample sizes, Fan and Lv (2008) suggested to iterate the SIS procedure so as to capture predictors that are marginally uncorrelated with response.

Fan and Lv's work is confined to linear regression of a continuous response. In pharmacogenomic studies, we are more interested in binary phenotypes and potentially censored event times. Extension to such phenotypes is not trivial because residuals and prediction errors are not well-defined. The assumption of multivariate normal predictors for the sure screening property of the SIS used by Fan and Lv is obviously not satisfied with SNP data. In addition, the high correlations among genetic variants present unique challenges. Finally, the existing literature is focused on selecting main effects whereas we are primarily interested in interactions.

In this research, we will develop variable selection methods for high-dimensional and highly correlated genomic features (such as SNPs and gene expressions). We will extend the iterative SIS (ISIS) idea to all possible phenotypes, particularly binary phenotypes and censored event times. Our methods will be targeted at selecting important interactions and thus identifying genetic variants predictive of drug response. To that end, we will introduce a hierarchical ISIS procedure to simultaneously select main effects and interactions. We will explore the selection of genetic variables at a group level so as to reduce collinearity and capture the joint effects of genetic variables in the same gene/pathway. We will investigate a novel cross-validation criterion for censored event times based on prediction intervals.

For binary phenotypes, one may apply machine learning techniques such as support vector machine (SVM) (Vapnik, 1995). Compared to the above regression approach, SVM is targeted at the simpler problem of classifying subjects without producing estimated probabilities and does not require regression modelling; therefore, SVM may provide more accurate and more robust classification. It would be unwise to classify the patients by using all genetic variables because of accumulation of noise and reduction of interpretability. In this research, we will incorporate variable selection into SVM to improve classification accuracy and enhance interpretability. We will extend the aforementioned ideas of ISIS, hierarchical structure, and group variable selection to SVM.

3.3 Providing Low-Bias Estimation of Effect Sizes With Complex and Highly Multivariate Data

In comparison with other areas of applied and methodological statistics, clinical trial design and analysis have developed in a relatively conservative manner. The insistence on randomization (in Phase III trials), the intent-to-treat principle, and the prospective declaration of primary and secondary endpoints have all resulted from hard lessons on the subtle biases that can arise from post hoc analysis decisions. A large, expensive clinical trial may provide the only data to guide the design of a future trial or more focused study. Thus, a considerable burden falls on the statistical analyst to fully consider the sampling scheme by which the data arose, and (importantly) the extent to which the data itself was used to guide the analysis. For traditional designs involving a small number of parameters, such as the effect of a new treatment therapy vs. control and perhaps a few key clinical variables, it is feasible to report all parameter estimates directly, as the estimates are of value regardless of apparent statistical significance. However, this approach is not reasonable when the number of variables is extremely large, as occurs in pharmacogenomic studies in which hundreds of thousands of genotypes or expression measurements may be gathered and related to drug response.

In GWAS, significance bias (i.e., the winner's curse) is receiving considerable attention (Zöllner and Pritchard, 2007; Ghosh et al., 2008; Zhong and Prentice, 2008) as a source of inflation of disease risk estimates. Significance bias arises in any setting in which a parameter estimate is obtained only when an associated test is declared statistically significant. A mathematically similar source of bias arises in the sequential analysis of clinical trials with data-dependent stopping rules, and has been widely studied (Whitehead, 1986). However, the effect of significance testing itself as a source of bias in clinical trials is not well-recognized, and is often conflated with the issue of *publication* bias (Dickersin et al., 1987; Veitch 2005).

The availability of large numbers of -omics measurements is likely to place increasing emphasis on *prediction* of outcomes using such measurements. Microarray-based prediction rules have been used in a variety of cancers, including prognosis of breast cancer (Glas et al., 2006; Oh et al., 2006), colon cancer (Garman et al., 2008), and melanoma (Mandruzzato et al., 2006). Such prognostic rules are important in the identification of potential tumor subtypes (which may offer differing responses by treatment), as well as efficient trial design in which only patients at high risk of recurrence should be included.

In pharmacogenetics and individualized therapy trials, future trial design will be heavily influenced by current studies in which the effects of numerous genotypes, potential patient subgroups, and data from other -omics platforms have been subject to prior significance testing. The failure to properly consider significance bias can result in improper trial design and analysis. While it is possible to handle these data in a complex Bayesian framework, a simpler frequentist conditional likelihood approach holds considerable promise in accounting for

significance bias in the design and analysis of clinical trials with complex data types and prediction schemes. However, several obstacles remain in the use of the conditional likelihood approach in clinical trials. One difficulty arises when hypothesis testing is performed for a primary parameter of interest, but additional inference is desired for a secondary parameter. One simple example arises when SNPs (other other -omics features) are tested for significant interaction with treatment in determining trial outcome, and inference is desired for the main effect of SNP on response.

The consideration of bias in estimating multiple effect parameters also provides a new perspective on prediction of outcome. The conditional likelihood approach can be restated in a manner that is very similar to established penalized regression procedures. This perspective is used to propose a new regression procedure that explicitly considers significance bias in the effects of multiple predictors, and is highly suited to the challenges of clinical trials using pharmacogenomics data.

Much of the methodology proposed here is relatively simple, and it is worth asking why the issue of significance bias has not been more prominently discussed in applied statistics (although it has been considered sporadically). Part of the answer may lie in the fact that only recently has it become routine to perform $10^5 \sim 10^6$ hypothesis tests, for which significance bias becomes an important issue. In addition, the -omics literature has until recently been mainly concerned with testing, not estimation. We anticipate that the analysis of clinical trials will be a very favorable additional proving ground for the methods described here.

3.4 Identifying Candidate Individualized Therapies in Pre-Clinical and Clinical Studies

A very exciting recent development in cancer treatment research is the concept of individualized therapies which are treatments tailored to individual characteristics (e.g., Huang et al., 2003; Ramaswamy and Perou, 2003). A discovery that has the potential to yield individualized therapies in breast cancer appeared in the December 13, 2008 issue of the ScienceDaily (ScienceDaily, 2008), which showed that postmenopausal women taking tamoxifen who have an inherited deficiency in the CYP2D6 gene have about a fourfold increased risk of early breast cancer compared to women without the deficiency. Another recent example of a potential individualized therapy is given by Paik et al. (2006), who described a biomarker which is a weighted sum of 21 gene expression levels that appears to predict the efficacy of chemotherapy on the 10-year recurrence of distant metastases.

As opposed to traditional drug discovery, which attempts to find the best single treatment for a large group of patients, an individualized therapy rule has the clear advantage of providing the best treatment options according to individual needs. A second advantage is the capacity of individualized therapy studies to discover effective drugs that may be missed by a clinical trial. For example, if treatment A is half as effective as treatment B for patients whose values of a biomarker are greater than its median value but is twice as effective as treatment B for patients whose values of the biomarker are less than its median, then the treatment differences will be completely undetected when all patients are lumped together. However, if we adopt the optimal individualized therapy rule, we will discover the very important latent structure that treatment A is quite effective when given to some people but withheld from others. Thus, individualized therapy studies are not just a nuanced improvement over standard clinical trials but a paradigm shift in methods for discovering effective treatments.

Appropriate statistical methods are needed for both the discovery and evaluation of biomarkers and potentially latent candidate individualized therapies based on those biomarkers. Generally speaking, the form that individualized therapies can take is a list of rules which assign different treatments according to different values of patient biomarkers based on genomic data or other prognostic factors. Finding individualized therapies that work for most patients is a very daunting task with essentially two fundamental problems. The first problem is that candidate individualized therapies are very difficult to find (they can be latent as argued above) and their identifications typically require intensive bioinformatic data mining. The second problem is that the candidate individualized therapies discovered from data mining are often not reproducible. The issue is that multiple testing is generally not controlled sufficiently during data mining and spurious results are not only possible but inevitable (e.g., Reid et al., 2005; Simon, 2005).

In this research, we will develop analytical tools for finding biomarkers from high dimensional data that significantly affect drug response. There are many bioinformatic tools for high dimensional association studies in medicine but most tools fall into two basic categories which often delineate two stages of analysis, high throughput screening and machine learning. The machine learning stage is frequently in the form of hierarchical clustering which is a type of unsupervised learning (e.g., Ma et al., 2004; Bertucci et al., 2005; Pawitan et al., 2005; Troester et al., 2006). Generally speaking, both approaches attempt to find important relationships

between a feature space X of genetic or other prognostic factors and an outcome Y. High throughput screening is widely used in the first stage of microarray analysis and in other platforms for selecting an initial set of genes that are associated with some phenotype of interest. Machine learning is also a very popular approach in medical and life science research (e.g., Cios et al., 2007), but the approach usually requires the outcome Y to be in the form of a dichotomous (or low-dimensional polytomous) classification (e.g., responders versus non-responders) unless unsupervised approaches are used. This limitation makes machine learning hard to use in some situations but it also allows for identification of more complex biomarker interactions with clinical outcomes. Thus there are trade-offs between the two approaches as well as a tremendous unmet need for new statistical methods that appropriately balance the weaknesses and strengths of data mining and statistical inference in a manner attuned to the task of identifying potentially reproducible candidate individualized therapies.

Consider the following pre-clinical in-vitro study, which our group is currently working on, of pharmacogenomic factors that influence colorectal cancer sensitivity to different chemotherapeutic drugs. Fresh colorectal cancer samples ($n = 100$) are obtained from patients who have consented to have their excess tissue kept for future research by the UNC-CH Tissue Procurement Facility. A portion of the excess tumor tissue will be seeded into culture flasks to establish extant cultures. The extant culture system can grow enough cells for up to 12 different drugs and drug combinations to be assessed, with most cultures providing enough cells for 6–8 drug assessments. After a sufficient quantity of cells have been cultured, the cells will be treated with multiple concentrations of the different drugs and drug combinations. The drugs and drug combinations chosen come exclusively from approved treatments for colorectal cancer patients or from current investigational trials for colorectal cancer. Chemosensitivity for each drug and drug combination will be compared across samples by assessing the EC50s (effective concentrations of 50% cell kill) for each drug.

Microarray-based genome-wide gene expression profiling will be performed on RNA isolated from the tumor samples in an effort to identify gene expression alterations that characterize drug sensitivity. Genome-wide profiling of DNA will also be conducted on the tumor samples in an effort to identify genetic variants that are associated with drug sensitivity. In order to utilize this rich data in an effective manner, it is important to have a valid statistical methodology that can at the very least classify the feature space (gene expression and profiling data) according to distinct chemosensitivity characteristics. To date, there exists no demonstrably valid and reproducible method of data analysis to accomplish this.

As a second example, consider a hypothetical cancer clinical trial where high dimensional patient biomarker data (such as gene-expression data) is assessed and patients are randomized to two or more treatment options. This kind of study can be built into a standard clinical trial framework, and is thus not only scientifically valuable but also feasible. However, in order for such a study to be meaningful, it is crucial to have a statistically valid method of analysis that handles the high-dimensional feature spaces and yields reproducible results of a potentially latent association. The goal of the proposed research is to validly analyze data of this kind, as well as data of the type described in the colorectal tumor example, in order to identify candidate individualized therapies which have a reasonable potential of being reproducible. Our target is a statistically valid method of inference involving a model with sufficient parsimony to assure both internal and external validity. Such an approach will still require a validation process, but the number of false leads and spurious results will be greatly reduced.

4 PRELIMINARY STUDIES

4.1 Investigators

We have assembled a group of investigators from the three institutions who have the relevant expertise and experience to carry out the proposed research. Dr. Danyu Lin has published extensively in the areas of survival analysis, clinical trials, and genetic association studies. Dr. Fred Wright's research is focused on statistical genomics and bioinformatics methods. Dr. Michael Kosorok's expertise includes high dimensional data analysis, microarrays, semiparametric inference, and applications of empirical process theory to statistical learning in biomedicine. Dr. Donglin Zeng's research interests include semiparametric models, empirical process theory, high-dimensional data analysis, and genetic association studies. Dr. Howard L. McLeod is internationally recognized for his expertise in pharmacogenomics and individualized therapy and has over 300 publications. Dr. James Todd Auma is specialized in pharmacology, toxicology, colorectal cancer biology, and the use of genomic profiling to investigate the molecular mechanisms underlying drug response. Dr. Yufeng Liu's expertise includes

statistical machine learning, high dimensional data and genomics. Dr. Wei Wang's expertise includes biomedical computing, high dimensional computing, machine learning and genomics. Dr. Sin-Ho Jung has published many methodological papers on survival analysis and clinical trials. Dr. Kourous Owzar's research interests include pharmacogenomics and survival analysis. Dr. Jung-Ying Tzeng's research is focused on genetic epidemiology and statistical genetics. Dr. Hao H. Zhang is an expert in nonparametrics and smoothing, variable selection, statistical machine learning, and high dimensional data analysis.

We describe below some of our prior work and ongoing studies that are relevant to the four specific aims.

4.2 Assessing the Effects of SNP Genotypes and Haplotypes on Drug Response

We have been at the forefront of developing statistical methods to detect haplotype-disease associations in cross-sectional, case-control and cohort studies (e.g., Lin et al., 2005; Lin and Zeng, 2006; Zeng et al., 2006) and to analyze untyped SNPs in case-control studies (Lin et al., 2008). Our software interface HAPSTAT and SNPMStat have been downloaded by more than 100 researchers and used in several real studies.

There are two major limitations with our work, and indeed with all existing literature. First, little attention has been paid to the complex phenotypes encountered in pharmacogenomic studies. Second, genetic and environmental factors are assumed to be independent. The independence assumption fails in many pharmacogenomic applications. For example, certain genes may influence both clinical variables (e.g., body size and disease severity) and drug response. The assumption is also violated when the environmental factors pertain to the covariates that are used to adjust for unmeasured confounding due to population substructure. The proposed research will remove such limitations and yield valid and efficient methods for detecting the effects of haplotypes and untyped SNPs on drug response.

Tzeng et al. (2003ab) developed a unified case-control test based on haplotype sharing for genome association screening. Recently, Tzeng et al. (2009) extended this work to general regression models. In addition, they united haplotype sharing methods and haplotype random-effects methods via gene-trait similarity regression. They showed that testing for zero coefficient in similarity regression is equivalent to testing for zero genetic variance component in the random-effects model. In addition, the score statistics under the two models share similar quadratic forms and incorporate genetic information solely through haplotype similarity. This gene-trait similarity regression model serves as the foundation for the gene-based analysis proposed in this research.

We have published extensively on multiple testing and sample size calculations. Lin (2005, 2006) provided an efficient Monte Carlo approach to assessing genomewide statistical significance for correlated test statistics. Jung and Jang (2006) showed how to accurately control the false discovery rate (FDR) for correlated test statistics. We have also published multiple testing methods to discover genomic markers that are correlated with survival endpoints using a nonparametric approach (Jung et al., 2005b) and a semi-parametric approach (Owzar et al., 2007). Jung (2005) and Jung et al. (2005a) discussed sample size calculations for FDR-based multiple testing methods. Jung et al. (2008) developed sample size formulas for the log-rank tests comparing one control arm and several experiment arms. Jung et al. (2009) derived a sample size formula to discover genes under blocked one-way ANOVA settings.

4.3 Predicting Drug Response Based on High-Dimensional Genomic Data

We have considerable experience in variable selection and model building in various contexts. Lu and Zhang (2007), Zhang and Lu (2007) and Johnson et al. (2008) studied moderate-scale variable selection under semiparametric regression models for censored event time data, while Xiao et al. (2008) considered variable selection for semiparametric linear mixed models in longitudinal studies. Zhang (2006) developed a general regularization framework to conduct simultaneous classification and variable selection for kernel support vector machines. Tang and Zhang (2006) developed the proximal multiclass support vector machines. Zhang et al. (2008) proposed a new shrinkage method based on the supnorm penalty for variable selection in multiclass SVMs. Zou and Zhang (2008) studied the adaptive elastic-net when the dimension of data diverges with the sample size.

We have conducted some simulation studies on the ISIS procedure with the LASSO penalty for a binary phenotype. Our limited results show that this approach is computationally feasible for GWAS and is capable of identifying important SNPs. We have also studied the SVM procedure with the SCAD penalty for micro-array data and found the approach promising.

4.4 Providing Low-Bias Estimation of Effect Sizes With Complex and Highly Multivariate Data

Our approach to handling significance bias has been recently described in the context of genetic association testing (Ghosh et al., 2008), but is quite generally applicable. Suppose that we have a model in which a (scalar) parameter β governs the relationship between any kind of predictor or experimental treatment and response, and for which $\beta = 0$ corresponds to the null hypothesis. Note that it does not matter that β may be only one parameter among many that are subject to testing, provided that the tests are performed separately. For example, in a cancer clinical trial, tests involving primary treatment effects and numerous tests of genotypes might be all be performed using the same dataset, perhaps with differing significance thresholds to account for the numerous comparisons performed for the -omics hypotheses. With an estimated parameter $\widehat{\beta}$ and standard error estimate $\widehat{\mathrm{SE}}(\widehat{\beta})$, the Wald statistic $T = \widehat{\beta}/\widehat{\mathrm{SE}}(\widehat{\beta})$ may be used for testing the null hypothesis. However, we may be interested in performing inference only when the test statistic is significant (i.e. $|T| \geq c$ for some fixed c). We will refer to this event as "significance selection," and under such selection $\widehat{\beta}$ may be a very poor estimate of β, with high bias and variability. We currently ignore the higher-order distinctions between T and alternative test statistics such as the likelihood ratio and score statistics, and place relatively "standard" requirements on the consistency of $\widehat{\beta}$ and $\widehat{\mathrm{SE}}(\widehat{\beta})$. Thus, the approach to follow applies to any finite-parameter model estimation typically used in clinical trial analysis. We do not currently consider the potential complications that might arise in semiparametric inference, or settings in which estimation departs from $\sqrt{n}$ consistency.

To perform improved inference, we have introduced an approximate conditional likelihood for the quantity $\mu = \beta/\widehat{\mathrm{SE}}(\widehat{\beta})$, noting that the random variable T is approximately normal with mean μ and unit variance prior to significance selection. To account for significance selection, we use the approximate conditional likelihood

$$L_c(\mu) = \frac{\phi(t-\mu)}{P_\mu(|T| > c)}$$

to represent the evidence for μ (Ghosh et al., 2008), where t is the observed value of T, and ϕ is the standard normal density. In this setting, a number of point estimates of μ are reasonable, with different strengths. It can be shown that no uniformly unbiased estimate of μ exists, and we have explored both the maximum conditional MLE $\widetilde{\mu} = \mathrm{argmax}_\mu L_c(\mu)$ and the estimate $\widetilde{\widetilde{\mu}} = \int \mu L_c(\mu) d\mu / \int L_c(\mu) d\mu$, which has desirable overall mean-squared error properties. A true confidence interval procedure for μ is available, after considering the significance selection, by inverting a test procedure for each μ. Once a point estimate and confidence interval are obtained, we convert back to the β scale using $\beta = \mu\, \widehat{\mathrm{SE}}(\widehat{\beta})$. Simulation studies demonstrated that the corresponding point estimates ($\widetilde{\beta}$ and $\widetilde{\widetilde{\beta}}$) have vastly superior bias and mean-squared error performance to $\widehat{\beta}$, and that the confidence procedure performs well in terms of β. In addition, we have shown via simulations that a standard confidence interval procedure proposed by Zhong and Prentice (2008) does not have correct confidence coverage, even with large sample sizes.

If the test statistic T is highly significant, then the above procedure produces estimates that are similar to $\widehat{\beta}$. The largest impact of the proposed procedure occurs when the test statistic is just barely significant. In such a situation, the modified estimate of β (whether $\widetilde{\beta}$ or $\widetilde{\widetilde{\beta}}$) tends to be considerably shrunk toward zero, effectively counteracting the significance bias. Although the approach here is described for a single parameter, we emphasize that it applies equally well when the approach is used repeatedly for different parameters β, as in single-SNP testing in genome scans. Indeed, the effects of significance bias are most extreme under such situations, because c must be large in order to provide effective error control.

4.5 Identifying Candidate Individualized Therapies in Pre-Clinical and Clinical Studies

We have started development of a new machine learning technique, "latent supervised learning," which directly addresses the challenges described in Section 3.4 for finding candidate individualized therapies. The basic idea is to first use unsupervised learning and related approaches to reduce the dimension of the feature space X and then classify X into two or more groups based on the distribution of an outcome Y under the assumption that there are latent classes that completely determine the distribution of Y. The latent classes can be thought of as "phenotypes" characterized by groups with different distributions for Y. What makes this approach powerful is that the groups do not have to be ordered according to Y. For example, the groups could have different

variances of Y but have the same mean. Importantly, the groups could represent different patterns of treatment response, or chemosensitivity, as in the colorectal cancer sample study mentioned in Section 3.4. To thoroughly capture this, we allow Y to also depend on an additional covariate vector Z which defines the distribution of Y. For example, Z could be two dimensional, with the first dimension being treatment indicator and the second dimension being age. If there are two different phenotypes defined by X but characterized by different relationships between Y and Z, then our proposed model will capture the phenotypes.

We conducted a small simulation study using a preliminary version of our algorithm with both X and Z being two-dimensional and Y being a continuous outcome. We set X to be bivariate standard normal with two phenotypes defined by whether or not $X^{\mathrm{T}}w > \gamma$, where $w = (w_1, w_2) = (-0.71, 0.71)$ and $\gamma = 0$. For simplicity, we set $Z = X$ and set $Y = \beta_{j1}X_1 + \beta_{j2}X_2 + \epsilon$, where $j = 1$ corresponded to $X^{\mathrm{T}}w > \gamma$ and $j = 2$ corresponded to $X^{\mathrm{T}}w \leq \gamma$, and ϵ is standard normal. We also set $(\beta_{11}, \beta_{12}) = (2, 1)$ and $(\beta_{21}, \beta_{22}) = (-2, -1)$. What is interesting about this example is that if one ignores the phenotype structure, the overall effect of X will wash out and no regression effect will be detected. Hence the structure is latent. We applied our method to 30 simulated data sets of size 200 each, and the means of the parameter estimates and standard errors are given in the table below. The method appears to work quite well overall, with all of the mean estimates being quite close to the true values and the standard errors being small. While there is clearly much work to be done before this method is ready for use in discovery of individualized therapies, the simulation studies show that the basic concept is feasible and appears to be internally valid.

Parameter	w_1	w_2	γ	β_{11}	β_{12}	β_{21}	β_{22}
True value	-0.71	0.71	0.00	2.00	1.00	-2.00	-1.00
Mean estimate	-0.74	0.67	-0.02	2.01	1.00	-2.03	-1.03
Standard error	0.03	0.03	0.05	0.17	0.11	0.15	0.19

5 RESEARCH DESIGN AND METHODS

In this section, we describe the key ideas and techniques to be used to accomplish the specific aims of the proposed research. We keep our description fairly non-technical and adopt the same notation for each aim.

5.1 Assessing the Effects of SNP Genotypes and Haplotypes on Drug Response

5.1.1 Notation and Assumptions

We consider a set of (correlated) SNPs. We may have a direct interest in the haplotypes of these SNPs or wish to use the haplotype distribution to infer the unknown value of one SNP from the observed values of the other SNPs. Let H and G denote the diplotype (i.e., the pair of haplotypes on the two homologous chromosomes) and genotype, respectively. We write $H = (h, h')$ if the diplotype consists of h and h', in which case $G = h + h'$. We allow the values in G to be missing at random. Note that H cannot be determined with certainty on the basis of G if the two constituent haplotypes differ at more than one position or if any SNP genotype is missing.

Let Y and X denote, respectively, the phenotype or trait of interest and the environmental factors or covariates. The phenotype may be quantitative or qualitative; it may also be an event time that is right or interval-censored. The covariates may consist of treatments, clinical variables and principal components used to adjust for population stratification (Price et al., 2006). Principal components and certain clinical variables (e.g., body size and disease severity) may be correlated with H. For quantitative and qualitative traits, the effects of X and H on Y are characterized by the conditional density of $Y = y$ given $X = x$ and $H = (h, h')$, denoted by $P_{\alpha,\beta,\xi}(y|x, (h, h'))$, where α, β and ξ pertain to intercept(s), regression parameters, and nuisance parameters (e.g., variance and overdispersion parameters), respectively. The regression effects are specified through the design vector $\mathcal{Z}(X, H)$, which is a vector-function of X and H. For example, if we are interested in the additive genetic effect of a risk haplotype h^* and its interactions with X, then we may specify

$$\mathcal{Z}(x, (h, h')) = \begin{bmatrix} I(h = h^*) + I(h' = h^*) \\ x \\ \{I(h = h^*) + I(h' = h^*)\}x \end{bmatrix},$$

where $I(\cdot)$ is the indicator function. The ability to incorporate interactions is critically important in pharmacogenomics. For the dominant and recessive models, we replace $I(h = h^*) + I(h' = h^*)$ by $I(h = h^* \text{ or } h' = h^*)$ and $I(h = h' = h^*)$, respectively. If we are interested in the additive effect of a particular SNP, then we replace $I(h = h^*) + I(h' = h^*)$ by the value of $(h + h')$ at that SNP position; dominant and recessive effects are defined similarly. When the phenotype of interest Y pertains to an event time, it is natural to formulate the effects of X and H on Y through the Cox (1972) proportional hazards model: $\lambda(t|X, H) = \lambda_0(t)e^{\beta^{\mathrm{T}} Z(X,H)}$, where $\lambda(\cdot|X, H)$ is the conditional hazard function of Y given X and H, $\lambda_0(\cdot)$ is an unspecified baseline hazard function, and β is the set of log hazard ratio parameters.

Let K be the total number of haplotypes that exist in the population. For $k = 1, \ldots, K$, we denote the kth haplotype by h_k. Define $\pi_{kl} = \Pr(H = (h_k, h_l))$ and $\pi_k = \Pr(h = h_k)$, $k, l = 1, \ldots, K$. Under Hardy-Weinberg equilibrium (HWE), $\pi_{kl} = \pi_k \pi_l$ $(k, l = 1, \ldots, K)$. We may allow Hardy-Weinberg disequilibrium (HWD) by incorporating an inbreeding coefficient as in Lin and Zeng (2006). Denote the probability function of H by $P_\gamma(\cdot)$, where γ consists of $\pi = (\pi_1, \ldots, \pi_K)^{\mathrm{T}}$ under HWE and of π and the inbreeding coefficient under HWD.

We formulate the dependence of X on H through the conditional density function $P(X|H)$, which is decomposed as $P(X_1|H, X_2) \times P(X_2)$, where X_1 is correlated with H, and X_2 is independent of H. If X_1 and X_2 are independent (which is the case when X_2 pertains to treatments), then $P(X_1|H, X_2)$ reduces to $P(X_1|H)$. We characterize the dependence between X_1 and H (and possibly X_2) through the general odds-ratio function (Chen, 2004) while leaving the marginal distributions of X_1 and X_2 completely unspecified. We denote $P(X_1|H, X_2)$ under such parametrization as $P_{\eta,F_1}(X_1|H, X_2)$, where η is the set of odds ratio parameters, and F_1 is the conditional distribution function of X_1 given some fixed values of H and X_2.

5.1.2 Data Structures and Likelihood Functions

Quantitative and Qualitative Phenotypes. For a clinical trial with n patients, the data consist of (Y_i, X_i, G_i) $(i = 1, \ldots, n)$. The phenotype Y can be binary or continuous, and possibly multivariate. As mentioned in Section 5.1.1, the conditional density of Y given X and H is given by $P_{\alpha,\beta,\xi}(Y|X, H)$, which can be formulated by generalized linear models (McCullagh and Nelder, 1989) for univariate phenotypes and by generalized linear mixed models (Diggle et al., 2002, Ch. 9) for multivariate phenotypes. Write $\theta = (\alpha, \beta, \xi, \gamma, \eta, F_1)$. The likelihood function for θ is proportional to

$$L(\theta) = \prod_{i=1}^{n} \sum_{H \in \mathcal{S}(G_i)} P_{\alpha,\beta,\xi}(Y_i|X_i, H) P_\gamma(H) P_{\eta,F_1}(X_{1i}|H, X_{2i}),$$

where $\mathcal{S}(G)$ denotes the set of diplotypes that are compatible with genotype G.

Event Times. If Y pertains to survival time or disease-free survival time, then Y is subject to right censoring; if Y pertains to time to disease progression, then it is subject to interval censoring. For right-censored event times, we observe $(\widetilde{Y}_i, \Delta_i)$ instead of Y_i, where $\widetilde{Y}_i = \min(Y_i, C_i)$, $\Delta_i = I(Y_i \le C_i)$, and C_i is the censoring time of the ith patient; for interval-censored event times, we observe (L_i, U_i), where L_i and U_i are the left and right examination times between which disease progression occurs. In either case, we formulate the effects of X and H on Y through the aforementioned proportional hazards model. Write $\theta = (\beta, \gamma, \eta, F_1, \Lambda_0)$, where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. The likelihood functions for θ based on right-censored and interval-censored data are

$$L(\theta) = \prod_{i=1}^{n} \sum_{H \in \mathcal{S}(G_i)} \left\{\lambda_0(\widetilde{Y}_i) e^{\beta^{\mathrm{T}} Z(X_i,H)}\right\}^{\Delta_i} \exp\left\{-\Lambda_0(\widetilde{Y}_i) e^{\beta^{\mathrm{T}} Z(X_i,H)}\right\} P_\gamma(H) P_{\eta,F_1}(X_{1i}|H, X_{2i}),$$

and

$$L(\theta) = \prod_{i=1}^{n} \sum_{H \in \mathcal{S}(G_i)} \left[\exp\left\{-\Lambda_0(L_i) e^{\beta^{\mathrm{T}} Z(X_i,H)}\right\} - \exp\left\{-\Lambda_0(U_i) e^{\beta^{\mathrm{T}} Z(X_i,H)}\right\}\right] P_\gamma(H) P_{\eta,F_1}(X_{1i}|H, X_{2i}),$$

respectively.

Informative Censoring. When the phenotype of primary interest Y is subject to informative censoring by time T, we takes the joint modeling approach. In particular, if Y pertains to the repeated measures of the PSA level and T pertains to time to informative drop-out (such as death and voluntary withdrawal), we formulate the conditional density function of Y through the generalized linear mixed model $P_{\alpha,\beta,\xi}(y|x,(h,h');b)$ and the conditional hazard function of T through the frailty model $\lambda(t|X,H;b) = \lambda_0(t)e^{\widetilde{\beta}^{\mathrm{T}}Z(X,H)+(\zeta\circ b)^{\mathrm{T}}\widetilde{X}}$, where b is a set of random effects with density function $P_\phi(b)$ that captures the dependence between Y and T, $\widetilde{X}$ is a subset of X, and $\zeta \circ b$ denotes the component-wise product of ζ and b. Write $\theta = (\alpha,\beta,\xi,\widetilde{\beta},\zeta,\phi,\gamma,\eta,F_1,\Lambda_0)$. Then the likelihood function for θ takes the form

$$L(\theta) = \prod_{i=1}^{n}\int_b \sum_{H\in\mathcal{S}(G_i)} \left\{\lambda_0(\widetilde{Y}_i)e^{\widetilde{\beta}^{\mathrm{T}}Z(X_i,H)+(\zeta\circ b)^{\mathrm{T}}\widetilde{X}_i}\right\}^{\Delta_i} \exp\left\{-\Lambda_0(\widetilde{Y}_i)e^{\widetilde{\beta}^{\mathrm{T}}Z(X_i,H)+(\zeta\circ b)^{\mathrm{T}}\widetilde{X}_i}\right\}$$

$$\times P_{\alpha,\beta,\xi}(Y_i|X_i,H;b)P_\gamma(H)P_{\eta,F_1}(X_{1i}|H,X_{2i})P_\phi(b)db,$$

where $\widetilde{Y}_i = \min(T_i,C_i)$, and $\Delta_i = I(T_i \le C_i)$.

Untyped SNPs. When one of the SNPs in G is untyped, i.e., missing on all study subjects, the haplotype distribution π cannot be estimated from the study data alone. Fortunately, external databases, such as the HapMap, can be used to estimate π. Let $L_E(\pi)$ denote the likelihood function for π based on the external sample. The forms of $L_E(\pi)$ for trios and unrelated individuals are given in Lin et al. (2008). The likelihood function for θ that combines the study data and the external data is $L_C(\theta) \equiv L(\theta)L_E(\pi)$.

5.1.3 Computation and Inference

The likelihood functions involve both finite- and infinite-dimensional parameters, posing tremendous computational and theoretical challenges. We adopt the nonparametric maximum likelihood estimation (NPMLE) approach. In this approach, the distribution function $F_1(\cdot)$ and the cumulative baseline function $\Lambda_0(\cdot)$ are treated as right-continuous functions with jumps at the observed values of X_1 and the observed event times, respectively. We will construct appropriate EM algorithms to carry out the maximization. The resulting NPMLEs are expected to be consistent, asymptotically normal, and asymptotically efficient, and the limiting covariance matrix of the NPMLE of β can be consistently estimated by inverting the information matrix for all parameters (including the jump sizes of F_1 and Λ_0) or by using the profile likelihood function (Murphy and van der Vaart, 2000). Likelihood-based procedures (such as Wald, score and likelihood-ratio statistics) can be used to make inference about individual components of β. We will establish the desired asymptotic properties by appealing to modern empirical process theory and semiparametric efficiency theory (Bickel et al., 1993; van der Vaart and Wellner, 1996; Kosorok, 2008).

5.1.4 Mode of Inheritance

Let T_l denotes the standard-normal test statistic under the lth mode of inheritance, where $l = 1,2,3$ under the additive, dominant and recessive models, respectively. We consider the maximum test statistic $Q = \max(|T_1|,|T_2|,|T_3|)$. The correlation matrix for (T_1,T_2,T_3) can be estimated upon expressing the numerator of each T_l as a sum of independent terms through the efficient score function (Lin, 2005; 2006). Given the (estimated) correlation matrix, the critical value for Q can be obtained by numerical integration.

5.1.5 Power/Sample Size Calculations

We can apply the usual formulas for calculating the power and sample size for standard-normal test statistics. For the maximum test statistic, the power calculation involves evaluating $\Pr(Q > c)$ through numerical integration. The power depends on allele frequencies/haplotype distribution, mode of inheritance, effect sizes and type I error. To correct for multiple testing, the type I error for each test is set at a very low level, in the order of 10^{-7}.

5.1.6 Gene-Trait Similarity Regression

For patient i, let Y_i be a discrete trait or a normal trait (after appropriate transformation), X_i be a vector of covariates, and H_i be the haplotype design vector. For patients i and j $(i < j)$, let Z_{ij} be the trait similarity defined as the cross product of the covariate-adjusted means, and S_{ij} be the haplotype similarity. Note that S_{ij} can be quantified directly from unphased genotypes if a phase-independent similarity metric is used (Tzeng et al., 2003b). The gene-trait similarity regression model for genetic main effect takes the form of $E(Z_{ij} \mid X, H) = bS_{ij}$, and the genetic effect can be detected by testing the null hypothesis of $b = 0$. Note that the proposed regression has a zero intercept because the effects of covariates have been adjusted through Z_{ij}. The similarity regression model incorporating the interactions between genes A and B takes the form $E(Z_{ij} \mid X, H) = b_A S_{ij}^A + b_B S_{ij}^B + b_{AB} S_{ij}^A S_{ij}^B$, where S_{ij}^A and S_{ij}^B pertain to the genetic similarity between patients i and j in genes A and B, respectively, and b_{AB} represents the gene-gene interaction. For studying gene-treatment interactions, we consider a two-armed trial with treatment indicator X_i for the ith patient. By the arguments of Elston et al. (2000), $X_i X_j$ represents the treatment effects in the similarity regression if the treatment effects are not adjusted through Z_{ij}. Thus, the regression model that incorporates the interaction takes the form of $E(Z_{ij} \mid X, H) = bS_{ij} + dX_i X_j S_{ij}$, and the gene-treatment interaction can be detected by testing the null hypothesis of $d = 0$.

For a given similarity regression model of interest, we can specify an equivalent haplotype random-effects model, in which the marginal trait covariance (with the haplotype random effects integrated out) can be partitioned into the genetic variance components and the interaction variance components. Testing for the zero regression coefficient (i.e., $b_{AB} = 0$ or $d = 0$) is equivalent to testing for the zero interaction variance component under the random-effects model, whose test statistic can be derived along the lines of Tzeng and Zhang (2007). For normal traits, we will construct the score function based on the restricted maximum likelihood (REML) function and derive the asymptotic distribution of the score statistic under the null hypothesis. We expect that the score statistic does not have an asymptotic normal distribution and the variation in the second term of the score is negligible relative to the first term. We propose to use the first term as the test statistic and expect it to follow a weighted χ^2 distribution, which can be approximated by a scaled χ^2 distribution.

The extension from normal traits to other traits can be achieved by considering the score function of the marginal likelihood function, which is in parallel to the REML function of the normal trait. For right-censored event times, we plan to construct the score statistic using the marginal partial likelihood (Prentice and Self, 1985). The results can be extended to multi-armed trials, in which case there are multiple interaction variance components and the test statistic will be a weighted sum of the score statistics.

5.1.7 Numerical Studies

We will apply the new methods to various real studies, including the CALGB studies mentioned in Section 3. GWAS SNP data have already been collected for CALGB 80303, a randomized study in pancreatic cancer. For CALGB 40101, a randomized study in breast cancer, over 2000 specimens have been collected, 1000 of which have been sent to the genotyping lab. For CALGB 90401, a randomized study in prostate cancer, over 900 specimens have been collected and will be genotyped in 2009. This study will provide repeated measures such as PSA levels in addition to the usual efficacy and toxicity outcome data.

We will conduct extensive simulation studies to assess the operating characteristics of the new methods in realistic settings. The simulations will be designed to mimic the CALGB studies. We will consider all possible phenotypes and various forms of genetic effects (including the effects of single SNPs, haplotypes and genes). We will consider clinical factors and principal components that are potentially correlated with SNP markers.

5.2 Predicting Drug Response Based on High-Dimensional Genomic Data

5.2.1 Iterative Sure Independence Screening

The data consist of (Y_i, Z_i) $(i = 1, \ldots, n)$, where Y_i is the phenotype of the ith patient and Z_i is the corresponding p-vector of predictors. The phenotype can be quantitative or qualitative. (Censored event times will be discussed later.) The predictors may consist of genetic variables (e.g., SNP markers or gene expression levels),

treatments, clinical variables (e.g., age, body size, and disease severity) and principal components used to adjust for population stratification, as well as interactions between genetic variables and treatments. All predictors are standardized by their sample standard derivations. We relate Y to Z through a generalized linear model. It is expected that only a small number of predictors contribute to the phenotype, such that most of the regression coefficients are zero. We wish to identify the subset of important predictors (whose regression coefficients are non-zero) so as to improve the accuracy of estimation/prediction and the interpretability of the model.

We first reduce the dimensionality of predictors from p to d through univariate correlation learning, where d is smaller than n, e.g., $d = n/\log n$ or $n-1$. Specifically, we perform univariate regression on each predictor and select the d predictors that are the most significant among the p predictors. If the predictors are SNP markers, then the screening can be done very efficiently with the Pearson chi-squared statistics for binary phenotypes and t statistics for continuous phenotypes.

Given the set of d predictors, we select a small subset through a moderate-scale variable selection procedure based on penalized likelihood. The penalized likelihood function takes the form

$$\widetilde{l}(\beta) = l(\beta) + \sum_{j=1}^{d} q_{\lambda_j}(|\beta_j|),$$

where $l(\beta)$ is the minus log-likelihood function based on the d predictors, $\beta = (\beta_1, \ldots \beta_d)^{\mathrm{T}}$, and $q_{\lambda_j}(\cdot)$ is a penalty function indexed by a regularization parameter λ_j. We will consider several penalty functions: (a) the LASSO penalty $q_\lambda(|\theta|) = \lambda|\theta|$ (Tibshirani, 1996), (b) the SCAD penalty

$$q_\lambda(|\theta|) = \begin{cases} \lambda|\theta| & \text{if} \quad |\theta| \le \lambda, \\ -\frac{(|\theta|^2 - 2a\lambda|\theta| + \lambda^2)}{2(a-1)} & \text{if} \quad \lambda < |\theta| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if} \quad |\theta| > a\lambda, \end{cases}$$

where a is a constant that is commonly set to 3.7 (Fan and Li, 2001), and (c) the adaptive LASSO penalty $q_\lambda(|\beta_j|) = \lambda\omega_j|\beta_j|$, where ω_j is a known weight (Zou, 2006). Imposing different magnitudes of penalty on predictors allows us to incorporate prior information or restrictions. Typically, we set $\lambda_j = c_j\lambda$, where λ is a common unknown tuning parameter and c_j is a pre-specified constant. For example, we may force the intercept and treatments in the model by setting the corresponding c_j to 0 while setting the c_j to 1 for all other predictors. Given the λ_j, we minimize $\widetilde{l}(\beta)$ by using local quadratic approximation (Fan and Li, 2001) or local linear approximation (Zou and Li, 2008). To speed up computation, we will explore the use of the coordinate decent technique (Friedman et al., 2007), which is tantamount to minimizing $\widetilde{l}(\beta)$ in a componentwise manner. We will choose the regularization parameters by cross-validation (Tibshirani, 1996; Fan and Li, 2001) or Bayesian information criterion (BIC) (Wang et al., 2007).

To reduce the chances of omitting important predictors, we apply the iterative sure independence screening (ISIS) procedure, which works as follows. In the first step, we select a subset of d_1 predictors $\mathcal{S}_1$ out of the p potential predictors using the SIS-based variable selection method described in the previous two paragraphs. In the next step, we include the estimated linear combination of those d_1 predictors (i.e., $\widehat{\beta}_1 Z_{1i} + \ldots + \widehat{\beta}_{d_1} Z_{d_1,i}$) as an offset in the model and apply the same variable selection method to the remaining $(p - d_1)$ predictors to obtain a subset of d_2 predictors $\mathcal{S}_2$. (Inclusion of the estimated linear combination in the model can significantly weaken the priority of those unimportant predictors that are highly associated with response through their correlations with the predictors in $\mathcal{S}_1$ while increasing the priority of those important predictors that are marginally weakly associated with response purely due to the presence of the predictors in $\mathcal{S}_1$.) We repeat this process until we obtain K subsets $\mathcal{S}_1, \ldots, \mathcal{S}_K$, whose union $\mathcal{S} = \cup_{k=1}^{K} \mathcal{S}_k$ has a size less than n. We then apply a moderate-scale variable selection method to the predictors in $\mathcal{S}$.

5.2.2 Hierarchical Iterative Sure Independence Screening

In pharmacogenomic studies, we are particularly interested in the interactions between genetic variables and treatments. It is natural to include the main effects of those variables whose interactions are selected. We will

extend the ISIS procedure to achieve this hierarchical structure. For simplicity of description, suppose that we have a two-armed trial with d SNP markers. Let R_i be the treatment indicator for the ith patient, and let G_{ji} be the genotype score for the jth SNP on the ith patient. To impose the hierarchical structure, we parametrize the linear combination of predictors as

$$\beta_0 + \sum_{j=1}^{d} \beta_j G_{ji} + \tau R_i + \sum_{j=1}^{d} \beta_j \gamma_j R_i G_{ji},$$

and impose the shrinkage penalty on the β_j's and γ_j's while imposing no penalty on β_0 and τ.

5.2.3 Group Selection

Genetic variables tend to be highly correlated. We may improve variable selection by grouping the genetic variables first. One possibility is to group the SNPs in the same gene or same biological pathway. Another possibility is to use a clustering algorithm to group SNPs into several clusters and then apply the principal components analysis to the SNPs within each cluster to construct weakly correlated predictors. By selecting the groups rather than the individual SNPs, we may avoid the collinearity of predictors and take advantage of the joint information of the SNPs in the same gene/pathway. To perform variable selection at the group level, we penalize $|\sum_j \beta_j^2|^{1/2}$ rather than the individual $\beta_j's$, where the summation is taken over all the predictors in the same group.

5.2.4 Censored Data

For potentially right-censored event times, we use the proportional hazards model (Cox, 1972) instead of a generalized linear model and replace $l(\beta)$ by the partial likelihood function. The univariate correlation screening can be done efficiently with the log-rank statistic. For interval-censored event times and informatively censored repeated measures, we replace $l(\beta)$ by the nonparametric likelihood functions described in Section 5.1.2. To speed up computation, we may parametrize the baseline hazard function and use the corresponding parametric likelihood. (Nonparametric likelihood can be used in the final step.)

5.2.5 Prediction Accuracy

For quantitative or qualitative phenotypes, we may assess the accuracy of prediction by comparing the observed and model-predicted values of the phenotype for each patient. This strategy does not work for censored data since the event times are unknown for the censored observations. We will explore a novel idea based on prediction interval. A prediction interval is a time interval which is expected to cover the event time T with a certain probability. A $100(1-\alpha)\%$ prediction interval associated with a set of predictors z is defined by $(t_{\alpha/2}, t_{1-\alpha/2})$, where t_α is the 100α percentile of $\widehat{S}(t|z)$, the estimated conditional survival function of T given z. To accommodate censoring, we calculate the prediction interval for the observation time $\widetilde{T} = \min(T, C)$, where C is the censoring time. Let $S_c(t|z)$ be the conditional survival function of C given z, which is estimated by $\widehat{S}_c(t|z)$ under a proportional hazards model or nonparametrically. Then a $100(1-\alpha)\%$ prediction interval for $\widetilde{T}$, denoted as (l, u), is defined by $1 - \widehat{S}_c(l|z)\widehat{S}(l|z) = \alpha/2$ and $\widehat{S}_c(u|z)\widehat{S}(u|z) = \alpha/2$. We may measure the accuracy of prediction by the proportion of the observation times that lie inside their prediction intervals. The evaluation is carried out via cross-validation.

5.2.6 Support Vector Machines

We consider the data structure described in Section 5.2.1 and code the binary phenotype as 1 vs -1. After reducing the number of predictors from p to d via univariate correlation learning, we solve the following opti-

mization problem:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left\{ 1 - Y_i \left(\beta_0 + \sum_{j=1}^{d} \beta_j Z_{ji} \right) \right\}_+ + \sum_{j=1}^{d} q_{\lambda_j}(|\beta_j|),$$

where $\beta = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$, and $a_+ = \max(0, a)$. The first term is called the hinge loss function.

To impose the hierarchical rule in selecting the main effects and interactions, we adopt the notation of Section 5.2.2 and propose to solve

$$\min_{\beta_0,\beta,\tau,\gamma} \sum_{i=1}^{n} \left\{ 1 - Y_i \left(\beta_0 + \sum_{j=1}^{d} \beta_j G_{ji} + \tau R_i + \sum_{j=1}^{d} \beta_j \gamma_j R_i G_{ji} \right) \right\}_+ + \sum_{j=1}^{d} \left\{ q_\lambda(|\beta_j|) + q_{\widetilde{\lambda}}(|\gamma_j|) \right\},$$

where λ and $\widetilde{\lambda}$ control the shrinkages of the main effects and interactions, respectively. We are particularly interested in the SCAD penalty because of its oracle property (Fan and Li, 1996). The SCAD penalty is a quadratic spline function with two knots λ and $a\lambda$, along with continuous first-order derivatives. Because the SCAD penalty is not convex, it is not easy to minimize the above function numerically. We propose to use local linear approximation (Zou and Li, 2008), which is tantamount to solving a LASSO-type problem. We will choose the tuning parameters λ and $\widetilde{\lambda}$ by cross-validation.

Suppose now that we use certain biological knowledge or empirical evidence to divide a set of d predictors into M groups with d_m predictors in the mth group such that $d_1 + \ldots + d_M = d$. For $m = 1, \ldots, M$, let W_{mi} be the set of d_m predictors in the mth group and let θ_m be the corresponding d_m-vector of coefficients. We propose to solve

$$\min_{\beta_0,\theta_1,\ldots,\theta_M} \sum_{i=1}^{n} \left\{ 1 - Y_i \left(\beta_0 + \sum_{m=1}^{M} \theta_m^{\mathrm{T}} W_{mi} \right) \right\}_+ + \sum_{m=1}^{M} q_\lambda(\|\theta_m\|),$$

where $\|\theta_m\| = (\theta_m^{\mathrm{T}} \theta_m)^{1/2}$. By penalizing the L_2-norm of θ_m, the mth group of variables will be selected or removed simultaneously. This selection is invariant under group-wise orthogonal transformations. Due to the special structure of $\|\theta_m\|$, we propose to use local quadratic approximation (Fan and Li, 1996) to solve the above minimization problem, which can be implemented via quadratic programming.

5.2.7 Theoretical Properties

We speculate that the univariate correlation learning has the sure screening property and that the ISIS procedure with the SCAD penalty has the oracle property (i.e., selecting the true model with probability 1 as n approaches ∞) in our settings. We also anticipate the selection consistency of the SVMs equipped with certain penalties to incorporate the hierarchical or grouping structure of the model. We will establish these theoretical properties by using advanced mathematical arguments (e.g., Brouwer fixed-point theorem, Bernstein inequality, random matrix theory, and empirical process theory).

5.2.8 Numerical Studies

We will apply the new methods to a variety of cancer studies, including the CALGB clinical trials. It should be noted that CALGB genomics studies are not restricted to SNP arrays. CALGB is also carrying out genomics studies with RNA microarrays, aCGH arrays and DASL protein arrays. As CALGB is conducting a diverse set of large genomics studies in many major disease sites and can provide professionally managed phenotypic data, it will be an invaluable resource to the proposed research.

We will conduct extensive simulation studies to evaluate the new methods. The simulations will be designed to mimic the CALGB studies. It is critically important to simulate realistic genomewide data. For SNP data, we will explore the genome simulators of Liang et al. (2007) and Li and Li (2008). We will simulate 500,000 SNPs and assign causal SNPs in a variety of manners. For binary phenotypes, we will compare the performance of logistic regression and SVM.

5.3 Providing Low-Bias Estimation of Effect Sizes With Complex and Highly Multivariate Data

There are two scientific goals for this aim. First, we will develop methods of inference for evaluating effect sizes for secondary parameters when they are selected for inference based on a test for the first parameter. For example, suppose that we are interested in gene×treatment interactions, but only for those genes with significant main effects, or perhaps vice versa. Other important examples include the selection of -omics features as biomarkers for intermediate phenotypes (such as PSA level in prostate cancer), where the relationship of the intermediate phenotype to an important clinical outcome (such as recurrent prostate cancer) has also been measured. A second goal of this aim is to use the conditional likelihood framework as a guide toward simultaneous estimation of numerous parameters. The latter work includes development of a new penalized regression procedure in order to produce improved prediction of clinical outcome from numerous predictors.

5.3.1 Correlated Parameter Estimates Under Significance Selection

We adopt the notation of Section 4.3. We assume that the researcher uses a model with primary effect β_1, which is subject to significance testing with the standard-normal test statistic $T_1 = \widehat{\beta}_1/\widehat{\mathrm{SE}}(\widehat{\beta}_1)$. Here, the term "primary" signifies only that it forms the basis for testing in the current framework, and does not imply that the parameter is necessarily related to the primary endpoints in the clinical trial. In addition, we expect that the procedure may be performed multiple times, with a different parameter serving the primary role in each instance. Let $\boldsymbol{\beta}_{-1} = (\beta_2, \ldots, \beta_p)^{\mathrm{T}}$ denote the vector of true secondary effect parameters. We wish to estimate $\boldsymbol{\beta}_{-1}$ only when the first test is significant, i.e., $|T_1| > c$ for a value c chosen appropriately to control error in manner appropriate to the problem. It is helpful to keep in mind that p will typically be small compared to the sample size (e.g., the vector β may refer to main effects and interactions produced by a single SNP or other -omics feature). However, the entire inference procedure described below may be performed hundreds of thousands of times. Thus c will often be large in order to control family-wise error rates. In such extreme instances, the correlation structure across multiple tests is less consequential than when performing few tests, and in fact is often near the Bonferroni bound, and at any rate can often be reasonably specified using analytic considerations or permutation analysis.

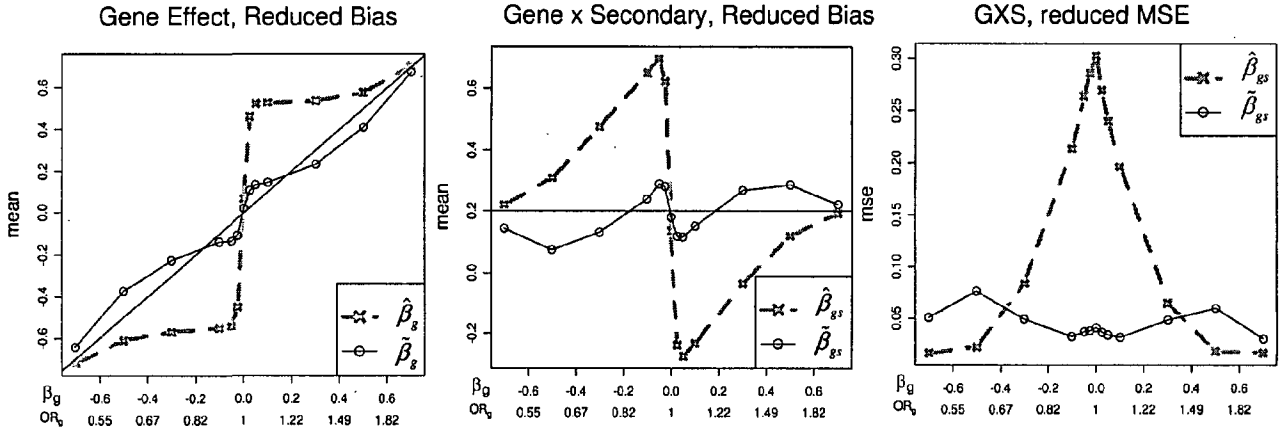
We will refer to $\widehat{\boldsymbol{\beta}}_{-1}$ as the naïve estimator, as it is obtained from standard statistical procedures without acknowledging selection of the first feature based on significance. The problem of estimating $\widehat{\boldsymbol{\beta}}_{-1}$ can be restated as mean-parameter estimation for truncated multivariate normal distribution with known variance-covariance matrix. Define $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \boldsymbol{\beta}_{-1} \end{pmatrix}$, $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\boldsymbol{\beta}}_{-1} \end{pmatrix}$, and $\mathbf{T} = \begin{pmatrix} T_1 \\ \mathbf{T}_{-1} \end{pmatrix}$, where $\mathbf{T}_{-1} = (T_2, \ldots, T_p)^{\mathrm{T}}$, and $T_i = \frac{\widehat{\beta}_i}{\widehat{\mathrm{SE}}(\widehat{\beta}_i)}$. Also, write $\mathbf{R} = \mathrm{corr}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} 1 & \boldsymbol{\rho}^T \\ \boldsymbol{\rho} & \mathbf{R}_{22} \end{pmatrix}$. By the arguments of Ghosh et al. (2008), $\mathbf{T}$ is approximately multivariate normal with mean $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \boldsymbol{\mu}_{-1} \end{pmatrix}$ and covariance matrix $\mathbf{R}$, where $\boldsymbol{\mu}_{-1} = (\mu_2, \ldots, \mu_p)^{\mathrm{T}}$, and $\mu_i = \frac{\beta_i}{\widehat{\mathrm{SE}}(\widehat{\beta}_i)}$.

We will see below that the bias in the secondary parameter vector depends importantly on the correlation $\boldsymbol{\rho}$ of $\widehat{\beta}_1$ with the terms in $\widehat{\boldsymbol{\beta}}_{-1}$. We assume here that $\mathbf{R}$ is estimated with sufficient accuracy, and treat it as known for the purposes of exposition. In practice, the estimation of $\mathbf{R}$ will devolve to the specifics of the particular problem, but in most cases is straightforward. For example, if the primary parameter and the secondary parameters appear in a common regression model, then the correlation of parameter estimates is implicit in the information matrix from the regression. We will also derive applicable results when a test is performed on a predictor and an intermediate phenotype, but not on the clinical outcome directly. Here again $\mathbf{R}$ may be estimated, provided that the correlation between the intermediate phenotype and the clinical outcome is also estimated, perhaps from a different study. We denote the naïve estimate of μ as $\widehat{\mu} = \mathbf{t}$, and the expectation can be shown analytically to be $E_{\mu}(\mathbf{T}| \, |T_1| > c) = \boldsymbol{\mu} + \begin{pmatrix} 1 \\ \boldsymbol{\rho} \end{pmatrix} \frac{\phi(c-\mu_1)-\phi(c+\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}$, where ϕ and Φ are the density and the cumulative distribution function of a standard normal, respectively. It is clear from the equation that the bias incurred in naïve estimation of $\boldsymbol{\mu}_{-1}$ is $\boldsymbol{\rho}$ times the bias in μ_1. In the special case of the null hypothesis $\mu_1 = 0$, the naïve estimates are unbiased. However, the variance of the estimates are extremely large in this situation, and so even in this case the shrinkage estimation described below is extremely valuable. The corresponding conditional

likelihood is

$$L_c(\mu) = p_\mu(\mathbf{t}| \, |T_1| > c) = \frac{p(\mathbf{t}; \mu, \mathbf{R})}{\Phi(-c-\mu_1) + \Phi(-c+\mu_1)},$$

where the numerator is the multivariate standard normal density. As in the single parameter problem, a number of competing estimators of μ are reasonable. Due to space constraints, we show here only the conditional m.l.e. $\widetilde{\mu} = \text{argmax}_\mu L_c(\mu)$, which again is converted to $\widehat{\beta}$ by multiplication with the estimated standard error vector. The plot below shows the results of simulations using the bias correction procedure for a primary parameter and two secondary parameters. The particular example uses logistic regression, which is often applied in followup clinical trial analysis, and sometimes in the main analysis (Steele and Wang, 2006). However, the key features will hold regardless of the statistical test. Here we are using a dichotomous genotype (g, with relative frequency 0.56 for $g = 0$ and 0.44 for $g = 1$) as a primary predictor (testing $\beta_g = 0$) of a dichotomous response Y (500 patients each for $Y = 0$ and $Y = 1$). In addition, a secondary dichotomous predictor (s, frequency 0.5 each for $s = 0$ and $s = 1$) is considered that is independent of g. The full model is $\text{logit}\{\Pr(Y = 1)\} = \beta_0 + \beta_g g + \beta_s s + \beta_{g,s}(g \times s)$. It is important to recognize that in this multiple logistic model the parameter estimates are correlated, even though g and s are not. Using a stringent choice of $c = 5$ to account for the fact that the genotype predictor is just one among perhaps several hundred thousand typed markers, we show in the figure below the results for a variety of primary log-odds ratios β_g, with 1,000 significant simulated datasets to obtain each point in the curves, and choosing β_0 to maintain a constant prevalence $\Pr(Y = 1) = 0.01$. We specified constant log-odds ratios $\beta_s = 0.2$ and $\beta_{g,s} = 0.3$. Note that the proposed estimation procedure provides greatly reduced bias in estimation of both β_g and $\beta_{g,s}$ (first two panels), along with greatly improved mean-squared error for $\beta_{g,s}$ for most values of β_g (third panel). Results for the estimation of s effects are similarly improved (not shown). Note that this is a simple example of very general phenomena, and similar results would hold if we had instead used the interaction null hypothesis $\beta_{g,s} = 0$ for the primary test.



Despite the successful use of the approximate conditional likelihood in this setting, several challenges remain. The performance of various point estimates of β_{-1} remain to be fully investigated. The creation of confidence intervals for the individual components of β_{-1} is also no longer straightforward, and this will be the subject of active investigation. We will investigate the possibility of obtaining pivotal quantities, but the crucial dependence on μ_1 suggests that no true confidence procedure is possible. We will investigate profile likelihood procedures in order to obtain approximate confidence intervals. Additionally, there are several issues involving applications to clinical trials, and specifically tests of interactions of treatment with -omics observations, that remain to be worked out. The estimation of $\mathbf{R}$ is straightforward in regression models with simultaneous fitting

of predictors, but is not entirely straightforward when β_1 and β_{-1} arise from separate estimation procedures, as occurs when the primary test involves a comparison to an intermediate phenotype/biomarker, while β_{-1} describes the relationship of predictors to a different clinical outcome.

5.3.2 Large-Scale Simultaneous Tests and Clinical Outcome Prediction

With the increasing complexity of biomarker assays, surrogate endpoints, and the multitude of possible treatment regimens in individualized therapy trials, it is inevitable that the analysis of clinical trials move toward the systematic handling of datasets in which the number of tests/predictors is much larger than the sample size (i.e., the "small n, large p" problem). The previous sub-aim finessed this issue by envisioning a series of tests performed separately, and the multiplicity of tests is handled by stringency in the threshold c. However, we must recognize a possible connection between this work and alternate approaches in which the test statistics themselves are penalized. A popular example includes the SAM statistics (Tusher et al., 2001), which are typically applied to gene expression datasets and which penalize lowly expressed genes more than highly expressed genes. The success of SAM (cited over 3000 times) derives in part from its improved false discovery rate (FDR) properties, in which genes that are measured more accurately are given greater weight. SAM-like statistics take the form $T' = \widehat{\beta}/(\widehat{\mathrm{SE}}(\widehat{\beta}) + s_0)$ for an s_0 suitably chosen for favorable FDR control.

A natural question arises - after applying significance testing for multiple hypotheses in which they are penalized (e.g. using the SAM penalty), how should one approach estimating the β parameters for those declared significant? The answer is entirely unknown, and yet will be important in order to resolve the conflict between methods designed to reduce the FDR and the growing need for proper estimation in the highly multivariate setting. We will investigate simple models for the signal-to-noise ratio of the -omics technology, and how that may be used in a conditional likelihood approach, but applied to T' instead of T.

Another issue that we anticipate will become important in the truly multivariate setting is the fact each μ as defined in the previous section is not truly a parameter. It is a random variable, because the standard error of the corresponding β is estimated from the data. A deeper understanding of the estimation procedure will benefit from a more rigorous approach, in which the dependence on the standard error accuracy is explored in more detail. Similarly, when p is large in the previous sub-aim (but still less than the sample size n), we may be reliant on a relatively poor estimate of the coefficient correlation matrix $\mathbf{R}$. We will investigate the consequences of this reliance for large p, and seek more robust approaches.

We have observed that the conditional likelihood described above has many properties in common with penalized regression techniques. Here we wish to perform accurate prediction of a clinical outcome based on p predictors, where $p >> n$, and we are not performing significance testing *per se*. Starting with a classic normal regression model $\mathbf{Y} = \mathbf{XB} + \epsilon$, $\mathbf{Y}$ may be a continuous measure of clinical response, and $\mathbf{X}$ may be an $n \times (p+1)$ matrix of -omics predictors and clinical variables. Here the danger of overfitting relationships to clinical outcome is severe, and the entire inferential and prediction process must be handled carefully. We consider for a moment the statistical comparison of a single predictor X_j to Y (i.e., an isolated test of only the two variables). If the test statistic for the corresponding coefficient β_j is not significant, then we may in essence view the estimate as having been thresholded to zero. If, however, the test statistic is significant, we may be subject to significance bias as discussed above. Returning to the full regression setting, it is reasonable to investigate whether our use of the conditional likelihood has a penalized likelihood interpretation, in which coefficients that are not "significant" are thresholded to zero, while other coefficients are shrunk toward zero to varying degrees.

Use of the conditional likelihood for the estimation of a single regression coefficient may be viewed as the application of a penalty term applied to μ, in which the log-likelihood for μ is penalized by the log of the denominator in the conditional likelihood, or $\log(P_\mu(|T| > c))$. If μ is far from zero, then the penalty is likely to have little effect, while if μ is small then the penalty is likely to result in substantial shrinkage of the coefficient. The conditional likelihood also has a Bayesian interpretation, in which $P_\mu(|T| > c)$ represents an improper prior for μ. In this manner, the prior may be compared to similar Bayesian interpretations of the one-predictor version of other penalized regression methods such as LASSO (double exponential prior) and ridge regression (normal prior).

We propose to extend these ideas into a fully-fledged method called *coefficient test regression* for penalized

regression using constructions similar to the conditional likelihood. The approach presents key challenges in handling multiple predictors, and the operating characteristics must be investigated. A rough outline of our current thinking follows. We perform stepwise regression or other multiple regression approach for a quantitative response to limit the number of non-zero coefficients to fewer than the sample size, requiring that the Wald statistics for all coefficients exceed c, where c need not be specified in advance, and indeed we expect that the optimal c will depend on the data. Then we obtain shrunken coefficient estimates using the conditional likelihood approach, and use these new estimates for prediction. Cross-validation is used to select the threshold c that provides the least error in prediction. We expect that this approach will work in a straightforward manner for linear regression, in which coefficient estimates are unbiased even if additional important predictors have not been included in the model. However, nonlinear regression approaches can give biased estimates of regression coefficients if some predictors are not included. This phenomenon can occur even for generalized linear models such as logistic regression. This source of bias may need to be considered explicitly, as we consider extending our approach to generalized linear models, because of the iterative nature of including only a small number of eventual predictors in the model.

5.3.3 Numerical Studies

Each of the proposed new methods will undergo a period of intense methodological development and testing and validation with simulated and existing datasets. Simulated datasets will include the use of actual gene expression datasets that have had mean effects removed in order to provide realistic residual correlation. These residuals can then be used in a specified alternative model as performed in Hu et al. (2005). For SNP genotype datasets, we have developed simulation methods through the HAP-SAMPLE approach, which re-uses HapMap data in order to simulate realistic case-control data (www.hapsample.org, Wright et al., 2007). HAP-SAMPLE can be used to simulate SNP×treatment interactions with a simple modification. We will also work with Core B to access data from clinical cancer studies at both the UNC-CH LCCC and Duke Comprehensive Cancer Center, and associated CALGB trials. While these data are accrued and undergoing data management, we also have immediate access to public datasets on cancer clinical trials and gene expression. Several such datasets are available on ArrayExpress and GEO, including datasets on multiple myeloma (Mulligan et al., 2007, ArrayExpress E-GEOD-9782), breast cancer (Bonnefoi et al., 2007, E-GEOD-6861), and glioblastoma (E-GEOD-7696).

5.4 Identifying Candidate Individualized Therapies in Pre-Clinical and Clinical Studies

We propose to combine data mining and statistical inference into a single unified procedure that ameliorates the multiple testing problem while allowing sufficient flexibility to achieve most of the discovery power found in data mining approaches. The basic idea is a new form of machine learning, "latent supervised learning," that balances the power and flexibility of data mining with the reproducibility of statistical inference. Preliminary studies of a very simple scenario, as described in Section 3.4, indicate that this procedure has promise to find the hidden but important structure that is of interest in the search for individualized therapies. The goal of the proposed research is to fully develop latent supervised learning (LSL) for use in pre-clinical studies, such as the chemosensitivity study mentioned in Section 3.4, as well as clinical studies for finding candidate individualized therapies for cancer.

5.4.1 Proposed Model and Algorithm

We assume that interest focuses on finding subgroups in a feature space X characterized by differences in the distribution of a clinical outcome of interest Y possibly modulated through a covariate Z. As a simple hypothetical example in the individualized therapy context, Z could be a dichotomous treatment indicator of whether the patient is receiving treatment A ($Z = 0$) or B ($Z = 1$), Y could be survival time, and X could be gene expression data for a biopsied tumor from the patient. We wish to know whether and how X can be best partitioned into meaningful groups which are heterogeneous in terms of the relationship between Y and Z. Specifically, we assume that there exists some linear classifier w with $\|w\| = 1$, a scalar cut-point γ, and a

classification $C \in \{0,1\}$ such that $C = I(w^{\mathrm{T}}X - \gamma > 0)$ and the conditional expectation or conditional distribution of Y given Z has a parametric form $g(Z;\theta_C)$ or $g(Y,Z;\theta_C)$, where θ_C, $C = 0,1$, is low dimensional, and $\theta_0 \neq \theta_1$. Estimation is based on a sample of size n of the form (X_i, Z_i, Y_i) $(i = 1,\ldots,n)$.

To reduce the dimension of X, some form of unsupervised learning and/or principal components or sparse principal components (Zou et al., 2004) could be used first. The procedure does not have to be restricted to linear classifiers if X is replaced by functionals of X before analysis. In this context, our linear classifier $w^{\mathrm{T}}x - \gamma$ is, in fact, non-linear. For example, X could be expanded to include all pair-wise products and squares of terms in addition to the original observations and sparse principal components could be used to reduce dimension. This would allow both linear and quadratic classifiers. Without loss of generality, however, we will denote the final form of the feature space as X, even if there are functionals of the original data and dimension reduction involved.

Our final model then involves a hyperplane $w^{\mathrm{T}}x - \gamma$ that classifies observations into two groups within which the model of the relationship between Y and Z is the same but the parameters are quite different. In the context of individualized therapies, this new model combines classification with assessment of treatment effect and thus can reduce the number of classification steps involved in identifying candidate individualized therapies. We will also develop rigorous statistical inference methods for this which will assure internal validity. The relative parsimony of the model will also assure external validity (reproducibility). Although this approach will still require a clinical validation process, the number of false leads and spurious results should be greatly reduced.

The parsimony of this model becomes clear when we compare it to a few other related models. For illustration, consider the latent supervised learning model where we assume that given $C = c$, $Y = \theta_c^{\mathrm{T}}Z + \epsilon$, where ϵ is independent with mean zero and finite but unknown variance, and has an otherwise unknown distribution, and that $\theta_0 \neq \theta_1$. In this case, $g(Z;\theta_c) = \theta_c^{\mathrm{T}}Z$. Note that this is a semiparametric model since the distribution of ϵ is mostly unspecified. The effect of X on (Y,Z) is completely captured by the hyperplane $w^{\mathrm{T}}x - \gamma$ and the difference between θ_0 and θ_1. We really cannot get more parsimonious without simplifying either the form of g or the form of the classifier. We will now demonstrate how competing models can potentially be considerably more complex. For example, one step up in complexity is the varying coefficient model $E(Y|Z = z, X = x) = \theta^{\mathrm{T}}(w^{\mathrm{T}}x)z$, where $\theta(u)$ is constrained to be monotone in u or to have some other smoothness requirement. This is a form of projection pursuit (Friedman and Stuetzle, 1981). We can easily make the model even more complicated by allowing w to be a matrix instead of only a vector. Greater complexity is also possible, but the model quickly becomes very difficult to interpret. Thus, the proposed latent supervised learning approach is essentially the most parsimonious that is achievable while still maintaining sufficient flexibility for finding candidate individualized therapies.

For estimation, we minimize an objective function which is computed by first dividing the data into two groups according to whether $\delta(X;w,\gamma) = I(w^{\mathrm{T}}X - \gamma > 0)$ is 0 or 1. Withing each group, we minimize the sum of some objective function $M(Y,Z;\theta)$. Applying this to all of the data, we estimate the parameters by minimizing

$$\mathcal{L}(w,\gamma,\theta_0,\theta_1) = \sum_{i=1}^{n} [\{1 - \delta(X_i;w,\gamma)\}M(Y_i,Z_i;\theta_0) + \delta(X_i;w,\gamma)M(Y_i,Z_i;\theta_1)]$$

over w, γ, θ_0 and θ_1, under the restriction $\|w\| = 1$ and possibly other restrictions. For example, M could take the form $(Y - Z^{\mathrm{T}}\theta)^2$ for least squares, $|Y - Z^{\mathrm{T}}\theta|$ for least absolute deviation, or the negative log-likelihood for maximum likelihood estimation. The flexibility of this approach goes beyond the usual likelihood based approach to include methods more robust to data contamination such as least absolute deviation. As with all high dimensional procedures, however, care must be taken with both selecting the tuning parameters of the model and restricting the dimension of X to avoid over fitting. For the simulation study mentioned in Section 3.4, we used least squares. In general, we will use the notation $\widehat{w}_n$, $\widehat{\gamma}_n$, $\widehat{\theta}_{0n}$ and $\widehat{\theta}_{1n}$ to denote the maximizers of $\mathcal{L}(w,\gamma,\theta_0,\theta_1)$.

5.4.2 Computational Methodology

Because the objective function $\mathcal{L}(w,\gamma,\theta_0,\theta_1)$ is in general not convex, maximization is quite difficult. In the special case where M has the least absolute deviation form, the optimization problem can be reduced to a

line search on γ combined with mixed-linear programming (e.g., Nemhauser and Wolsey, 1988). Unfortunately, even with excellent commercial software, the computational intensity requires days of computation even for modest sized samples. It appears that a direct form of optimization is needed. The approach we propose is to enumerate all partitions of the data into two subsets that are possible through the hyperplanes $w^{\mathrm{T}}x - \gamma$, under some sort of reasonableness constraint such as the requirement that at least $1 \leq m_0 \leq n/2$ observations must be in both subsets. For example, setting $m_0 = 5$ or $m_0 = 10$ would be reasonable. This is just to ensure that the results are reasonable. We also carefully track the subset generation to avoid redundancy. For each partition so generated, we maximize $\mathcal{L}(w, \gamma, \theta_0, \theta_1)$ for a pair (w, γ) consistent with the partition, i.e., the partition is defined through $\delta(X_i; w, \gamma)$.

The challenge is then reduced to efficient enumeration of a collection of hyperplanes that creates all possible reasonable partitions. Note that there are several values of (w, γ) that will generate the same partition, but for large sample sizes, these values are all quite close together. Thus, the non-uniqueness does not pose a real difficulty. Returning to the enumeration problem, let $X_n = \{x_1, \ldots, x_n\}$ be a fixed data set of size n of d-dimensional feature vectors. Note that for a given w, the data can be ranked by the values $w^{\mathrm{T}}x_1, \ldots, w^{\mathrm{T}}x_n$, and then the different values of γ yield different slices through this ranking. Hence, if we can enumerate all possible rankings induced by w and then check only those slices that yield partitions not previously generated, we would be able to efficiently enumerate all possible unique partitions caused by hyperplanes. One way to accomplish this is by enumerating values of w that lead to unique rankings of the data, considering only those values of γ at the midpoint between the sorted values of $w^{\mathrm{T}}x_1, \ldots, w^{\mathrm{T}}x_n$, while omitting partitions already examined. This requires a combination of efficient computation and efficient hash table construction.

Interestingly, this basic problem has been of interest for quite some time, and sharp bounds on both the number of possible unique vectors w leading to unique orderings and the number of unique partitions caused by dividing hyperplanes have been derived. The first number is the "number of linearly inducible orderings" (Cover, 1967), while the second number is the "number of partitioning hyperplanes" (Cover, 1965; Anthony, 2004). Cover (1967) showed that the number of linearly inducible orderings is bounded above by $Q(n, d)$, where $Q(n, d)$ is defined recursively as $Q(n, d) = Q(n-1, d) + (n-1)Q(n-1, d-1)$ with $Q(n, 1) = 2$ and $Q(2, d) = 2$. Note that $Q(n, 1) = 2$, $Q(n, 2) = n(n-1)$, and, in general, $Q(n, d) = O(n^d)$. Cover (1965) showed that the number of partitioning hyperplanes is at most $\sum_{i=0}^{d} \binom{n-1}{i}$. Anthony (2004) extended the result to give a bound on the number of partitions of the data achieved by k parallel hyperplanes which is applicable to the setting where more the two phenotypes are expected. The challenge, of course, is how to use these results to yield a computationally efficient algorithm for enumerating partitioning hyperplanes.

The proof of the result in Cover (1967) shows that the unique w vectors are connected to the structure of all pairwise difference vectors $x_i - x_j$ $(i \neq j)$. After careful analysis, we were able to exploit this difference to come up with an algorithm for generating a set of vectors w which generates all possible linearly inducible orderings without any duplication of orderings. For any $w \in \mathbb{R}^d$ with $\|w\| = 1$, let $R(w)$ be the ranking of the observations $w^{\mathrm{T}}x_1, \ldots, w^{\mathrm{T}}x_n$. Let W be a finite set of w's. We say W has "no duplication" if $R(w_1) \neq R(w_2)$ for any two distinct $w_1, w_2 \in W$. In addition, W is "complete" if for any $w \in \mathbb{R}^d$ with $\|w\| = 1$, there exists some $\widetilde{w} \in W$ such that $R(w) = R(\widetilde{w})$. Our computational goal is then to find a W which is complete with no duplication.

We now describe our proposed approach to finding such a W for the special case when $d = 2$. For every distinct pair of points $x_i, x_j \in X_n$, calculate the angle θ_{ij} between the line $x_j - x_i$ and the vector $(0, 1)$. This can be done by taking the arctan of the slope of the line $x_j - x_i$. Add both the angle $\pi/2 + \theta_{ij}$ and $3\pi/2 + \theta_{ij}$ (subtracting off 2π if the total $> 2\pi$) and save in a set T_n. Do this for all $1 \leq j < i \leq n$ to obtain T_n with no more than $n(n-1)$ distinct elements (there may be fewer) all in the range $(0, \pi]$. Now sort the elements of T_n, and denote the resulting ordered distinct elements $t_1, \ldots, t_k$, where $k \leq n(n-1)$. Now compute $u_j = (t_j + t_{j+1})/2$, for $j = 1, \ldots, k-1$, and compute $u_k = (t_k + t_1 + 2\pi)/2$, where we take $u_k = u_k - 2\pi$ if $u_k > 2\pi$, and call the resulting collection U. Now let W be the set of vectors of the form $(\sin u, \cos u)$ running over $u \in U$. We will prove that the set W so constructed is complete with no duplication. The approach for general d is actually quite similar to the case of $d = 2$ but requires dividing hyperspheres into regions rather than circles into segments.

To generate all hyperplanes, we iterate through all values of $w \in W$, and, for each such w, we only need to check those values of γ that lie between the sorted values of $w^{\mathrm{T}}x_1, \ldots, w^{\mathrm{T}}x_n$. To make this fully efficient, a hash table needs to be generated that avoids computations for partitioning hyperplanes already constructed.

Preliminary assessments of this general algorithm show that it is promising and efficient.

5.4.3 Statistical Theory and Inference

We will study the large-sample properties of the proposed estimation procedure, including consistency and distributional convergence for all parameters, using empirical process methods (Kosorok, 2008; van der Vaart and Wellner, 1996). Estimation for the (w, γ) component is quite similar to the problem of finding the change-point in change-point regression (Kosorok and Song, 2007; Section 14.5.1 of Kosorok, 2008). The $\widehat{\gamma}_n$ parameter, in particular, has the non-regular property that $n(\widehat{\gamma}_n - \gamma)$ converges to a compound Poisson process when the model is correctly specified and that other convergence rates (besides n) and limiting distributions are possible when the model is incorrectly specified (e.g., Banerjee and McKeague, 2007). Empirical processes are absolutely needed here to obtain valid results. The other parameter estimators have different limiting distributions that can also be determined through empirical process techniques. Unfortunately, standard approaches to inference, including the bootstrap, are not applicable here and new approaches to inference are needed. It appears that a modified bootstrap, of the kind described in Kosorok and Song (2007), will work for latent supervised learning.

The limiting distributions of the estimators are quite different if $\theta_0 = \theta_1$, and having a valid test of the null hypothesis $H_0 : \theta_0 = \theta_1$ is crucial. Unfortunately, the model occurring under the null hypothesis is not identifiable. Fortunately, our group has recently developed a theoretically valid hypothesis testing procedure for this general non-identifiability setting (Song et al., 2008), and the approach appears to be applicable to the latent supervised learning setting described above. We have extensive experience in applying empirical process techniques and other statistical and computational techniques to problems such as this, and we believe that our general approach will be successful.

5.4.4 Numerical Studies

We will utilize Core C to assist us in coding a careful and thorough simulation study of the internal validity and reproducibility of our proposed approach. We will simulate data sets with a range of sample sizes from small to large to evaluate the theoretical predictions as well as the performance for realistic sample sizes in the range of moderate sized Phase III clinical trials, e.g. 100, 200, 500, 1000, 2000 and 5000, as well as smaller sizes such as 50 and 20. We also consider a range of models motivated by the colorectal tumor example mentioned in Section 3.4 as well as motivated by clinical trial data sets in Core B. We will also compare the performance of our approach with existing approaches for identifying candidate therapies such as those approaches mentioned in Section 3.4.

We will also evaluate our approach on existing clinical trial data sets obtained through Core B which have high dimensional prognostic data as well as the usual clinical outcomes. The selection of these data sets will need to wait until the methodology has been developed far enough for data analysis beyond simulations. The 100 colon data set mentioned in Section 3.4 will probably become available in the fifth year of the grant, at which point we will use that data to evaluate the new methods. Part of the challenge is that sample size requirements will not be known until we have completed the statistical theory work on this problem which will occur in approximately year three. However, we will utilize careful and appropriate sample size analysis in our selection of data sets to analyze.

5.5 Software Implementation and Dissemination

Each of the proposed new methods will first undergo a developmental implementation until the procedure has been validated via simulation studies and real data applications. Core C will then assist with making the implementation into a usable and robust software package for dissemination in appropriate formats, including R and SAS. Guidebooks and web pages of instructions and examples will also be developed and the new software will be communicated through presentations and short courses at appropriate professional meetings. We will also identify several beta testers outside of our institutions to help refine the quality of the software implementation.

We are experienced in producing JAVA code with direct application (FastMap, Gatti et al., 2008a), as well as a useful interface (SAFE-GUI, Gatti et al., 2008b) to underlying code in R (Barry, Nobel and Wright,

2005). SAFE-GUI offers a good example of code development that relies on existing genomic annotation in R/Bioconductor, and thus expressly avoids needing to "reinvent the wheel" in maintaining genomic annotation.

5.6 Timetable

For each of the new methods to be developed in each aim, there will be four phases of research and development: (1) methodological phase in which we construct new statistical and computational methods, establish their theoretical properties, and implement them in research code; (2) simulation phase where we assess the performance of the proposed methods through extensive simulation studies; (3) data analysis phase where we apply the new methods to real cancer studies; and (4) software development and dissemination phase where we develop our software, test it in practical settings, and disseminate the software in a manner useful and accessible for practitioners. Generally speaking, most of the activities in phase 1 will take place over the first three years, phase 2 will expand over years 2-4, phase 3 will begin near the end of year four, and phase 4 will commence in year five.

6 INCLUSION ENROLLMENT REPORT

N/A

7 BIBLIOGRAPHY AND REFERENCES CITED

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science* **322**, 881–888.

Bader, J. S. (2001). The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics* **2**, 11–24.

Banerjee, M., and McKeague, I. W. (2007). Confidence sets for split points in decision trees. *Annals of Statistics* **35**, 543–574.

Barry, W. T., Nobel, A. B., and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21**, 1943–1949.

Bertucci, F., Finetti, P., Rougemont, J., Charafe-Jauffret, E., Cervera, N., Tarpin, C., Nguyen, C., Xerri, L., Houlgatte, R., Jacquemier, J., Viens, P., and Birnbaum, D. (2005). Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer Research* **65**, 2170–2178.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: Johns Hopkins University Press.

Bonnefoi, H., Potti, A., Delorenzi, M., Mauriac, L., Campone, M., Tubiana-Hulin, M., Petit, T., Rouanet, P., Jassem, J., Blot, E., Becette, V., Farmer, P., Andr, S., Chaitanya, A., Mukherjee, S., Cameron, D., Bergh, J., Nevins, J., Iggo, R. (2007). Validation in a breast cancer randomized clinical trial (EORTC 10994/BIG 00-01) of regimen-specific gene signatures that predict pathological complete response to neo-adjuvant chemotherapy. *Lancet Oncology* **8**, 1071–1078.

Chapman, J., and Whittaker, J. (2008). Analysis of multiple SNPs in a candidate gene or region. *Genetic Epidemiology* **32**, 560–566.

Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association* **99**, 1176–1189.

Cios, K. J., Kurgan, L. A., and Reformat, M. (2007). Machine learning in the life sciences. *Engineering in Medicine and Biology Magazine, IEEE* **26**, 14–16.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., and Smith, H., Jr. (1987). Publication bias and clinical trials. *Controlled Clinical Trials***8**, 343–353.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd Ed. Oxford: Oxford University Press.

Domenici, J. S., Bruinsma, T. J., Warmoes, M. O., Bernards, R., Wessels, L. F., and Van't Veer, L. J. (2006). Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics.* **7**, 278.

Elston, R. C., Buxbaum, S., Jacobs, K. B., and Olson, J. M. (2000). Haseman and Elston revisited. *Genetic Epidemiology* **19**, 1–17.

Epstein, M. P., and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* **73**, 1316–1329.

Evans, W. E., and Relling, M. V. (1999). Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491.

Evans, W. E., and McLeod, H. L. (2003). Pharmacogenomics–drug disposition, drug targets, and side effects. *New England Journal of Medicine* **348**, 538–549.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B* **70**, 849–911.

Frank, I. E., and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.

Friedman, J. H., and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817–823.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302–332.

Garman, K. S., Acharya, C. R., Edelman, E., Grade, M., Gaedcke, J., Sud, S., Barry, W., Diehl, A. M., Provenzale, D., Ginsburg, G. S., Ghadimi, B. M., Ried, T., Nevins, J. R., Mukherjee, S., Hsu, D., and Potti, A. (2008). A genomic approach to colon cancer risk stratification yields biologic insights into therapeutic opportunities. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 19432–19437.

Gatti, D. M., Shabalin, A. A., Lam, T. C., Wright, F. A., Rusyn, I., and Nobel, A. B. (2008). FastMap: fast eQTL mapping in homozygous populations. *Bioinformatics*,Epub Date: 2008/12/19.

Gatti, D. M., Sypa, M., Rusyn, I., Wright, F. A. and. Barry, W. T. (2008). SAFEGUI: resampling-based tests of categorical significance in gene expression data made easy. *Bioinformatics*, Epub Date: 2008/12/23.

Ghosh, A., Zou, F., and Wright, F. A. (2008). Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *American Journal of Human Genetics.* **82**, 1064–1074.

Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93–99.

Goeman, J. J., van de Geer, S. A., and van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society, Series B* **68**, 477–493.

Hoskins, J. M., Goldberg, R. M., Qu, P., Ibrahim, J. G., and McLeod, H. L. (2007). UGT1A1*28 genotype and irinotecan-induced neutropenia: dose matters. *Journal of the National Cancer Institute* **99**, 1290–1295.

Hu, J., Zou, F., and Wright, F. A. (2005). Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics* **21**, 3264–3272.

Huang, E., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R., and Huang, A. T. (2003). Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008) Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 672–680.

Judson, R., Stephens, J. C., and Windemuth, A. (2000). The predictive power of haplotypes in clinical response. *Pharmacogenomics* **1**, 15-26.

Jung, S. H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–3104.

Jung, S. H., Bang, H., and Young, S. (2005a). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics* **6**, 157–169.

Jung, S. H., Owzar, K., and George, S. L. (2005b). A multiple testing procedure to associate gene expression levels with survival. *Statistics in Medicine* **24**, 3077–3088.

Jung, S. H. and Jang, W. (2006). How accurately can we control the FDR in analyzing microarray data? *Bioinformatics* **22**, 1730–1736.

Jung, S. H., Kim, C., and Chow, S. C. (2008). Sample size calculation for the log-rank tests for multi-arm trials with a common control. *Journal of Korean Statistical Society* **37**, 11–22.

Jung, S. H., Sohn, I., Feng, L., Leppert, P. C., and George, S. L. (2009). Analysis and sample size calculation for micorarray experiments with blocked one-way design. *BMC Bioinfomatics*, Under review.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Kosorok, M. R., and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics* **35**, 957–989.

Lake, S. L., Lyon, H., Tantisira, K., Silverman, E. K., Weiss, S. T., Laird, N. M., and Schaid, D. J. (2003). Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity* **55**, 56–65.

Li, C., and Li, M. (2008). GWAsimulator: a rapid whole-genome simulation program. *Bioinfomatics* **24**, 140–142.

Liang, L., Zöllner, S., and Abecasis, G. R. (2007). GENOME: a rapid coalescent-based whole genome simulator. *Bioinfomatics* **23**, 1565–1567.

Lin, D. Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21**, 781–787.

Lin, D. Y. (2006). Evaluating statistical significance in two-stage genomewide association studies. *The American Journal of Human Genetics* **78**, 505–509.

Lin, D. Y., Zeng, D., and Millikan, R. (2005). Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genetic Epidemiology* **29**, 299–312.

Lin, D. Y., and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* **101**, 89–118.

Lin, D. Y., Hu, Y., and Huang, B. E. (2008). Simple and efficient analysis of disease association with missing genotype data. *American Journal of Human Genetics* **82**, 444–452.

Lu, W. and Zhang, H. H. (2007). Variable selection for proportional odds model. *Statistics in Medicine* **26**, 3771-3781.

Ma, X.-J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J. T., Tran, Y., Tran, D., Tassin, A., Amon, P., Wang, W., Wang, W., Enright, E., Stecker, K., Estepa-Sabal, E., Smith, B., Younger, J., Balis, U., Michaelson, J., Bhan, A., Habin, K., Baer, T. M., Brugge, J., Haber, D. A., Erlander, M .G., and Sgroi, D. C. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **5**, 607–616.

Mandruzzato, S., Callegaro, A., Turcatel, G., Francescato, S., Montesco, M. C., Chiarion-Sileni, V., Mocellin, S., Rossi, C. R., Bicciato, S., Wang, E., Marincola, F. M., and Zanovello, P. (2006). A gene expression signature associated with survival in metastatic melanoma. *Journal of Translational Medicine*. **4**, 50.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Ed. New York: Chapman and Hall.

Mulligan, G., Mitsiades, C., Bryant, B., Zhan, F., Chng, W. J., Roels, S., Koenig, E., Fergus, A., Huang, Y., Richardson, P., Trepicchio, W. L., Broyl, A., Sonneveld, P., Shaughnessy, J.D., Jr., Bergsagel, P. L., Schenkein, D., Esseltine, D. L., Boral, A., and Anderson, K. C. (2007). Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood* **109**, 3177–3188.

Murphy, S. A., and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95**, 449–465.

Nemhauser, G. L., and Wolsey, L. A. (1988). *Integer and Combinatorial Optimization*. New York: Wiley.

Nicolae, D. L. (2006). Testing untyped alleles (TUNA)–applications to genome-wide association studies. *Genetic Epidemiology* **30**, 718–727.

Oh, D. S., Troester, M. A., Usary, J., Hu, Z., He, X., Fan, C., Wu, J., Carey, L. A., and Perou, C. M. (2006). Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *Journal of Clinical Oncology* **24**, 1656–1664.

Owzar, K., Jung, S. H., and Sen, P. K. (2007). A copula approach for detecting prognostic genes associated with survival outcome in microarray studies. *Biometrics* **63**, 1089–1098.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F. L., Watson, D., Bryant, J., Costantino, J. P., Geyer, C. E. Jr., Wickerham, D. L., and Wolmark, N. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *Journal of Clinical Oncology* **24**, 3726–3734.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E. T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedrén, S., and Bergh, J. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research* **7**, R953–R964.

Prentice, R. L., and Self, S. G. (1985). Contribution to the discussion of the paper by D. G. Clayton and J. Cuzick. *Journal of the Royal Statistical Society, Series A* **148**, 112–113.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

Ramaswamy, S., and Perou, C. M. (2003). Commentary. *Lancet* **361**, 1576–1577.

Reid, J. F., Lusa, L., De Cecco, L., Coradini, D., Veneroni, S., Grazia Daidone, M., Gariboldi, M., Pierotti, M. A. (2005). Limits of predictive models using microarray data for breast cancer clinical treatment outcome. *Journal of the National Cancer Institute* **97**, 927–930.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology* **27**, 348–364.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

Schaid, D. J., McDonnell, S. K., Hebbring, S. J., Cunningham, J. M., and Thibodeau, S. N. (2005). Nonparametric tests of association of multiple genes with human disease. *American Journal of Human Genetics* **76**, 780–793.

ScienceDaily (2008). Web address: http://www.sciencedaily.com/releases/2008/12/081213130059.htm.

Simon, R. (2005). Development and validation of therapeutically relevant multi-gene biomarker classifiers. *Journal of the National Cancer Institute* **97**, 866–867.

Song, R., Kosorok, M. R., Fine, J. P. (2008). On asymptotically optimal tests under loss of identifiability in semiparametric models. *Annals of Statistics*, In press (NIHMS 75080).

Spinka, C., Carroll, R. J., and Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* **29**, 108–127.

Steele, F., and Wang, D. (2006). *Clinical Trials: A Practical Guide to Design, Analysis, and Reporting*, 273–278. London: Remedica.

Stram, D. O., Pearce, C. L., Bretsky, P., Freedman, M., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E., and Thomas, D. C. (2003). Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity* **55**, 179–190.

Tang, Y. and Zhang, H. H. (2006). Multiclass proximal support vector machines. *Journal of Computational and Graphical Statistics* **15**, 339–355.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

The International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Troester, M. A., Herschkowitz, J. I., Oh, D. S., He, X., Hoadley, K. A., Barbier, C. S., Perou, C. M. (2006). Gene expression patterns associated with p53 status in breast cancer. *BMC Cancer* **6**, 276.

Tzeng, J. Y., Byerley, W., Devlin, B., Roeder, K., and Wasserman, L. (2003a). Outlier detection and false discovery rates for whole-genome DNA matching. *Journal of the America Statistical Association* **98**, 236–246.

Tzeng, J. Y., Devlin, B., Wasserman, L., and Roeder, K. (2003b). On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *American Journal of Human Genetics* **72**, 891–902.

Tzeng, J. Y., and Zhang, D. (2007). Haplotype-based association analysis via variance component score test. *American Journal of Human Genetics* **81**, 927–938.

Tzeng, J. Y., Zhang, D., Chang, S. M., Thomas, D., and Davidian, M. (2009). Gene-trait similarity regression for multimarker-based association analysis. *Biometrics*, In press. NIHMSID: 90025.

van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications in Statistics*. New York: Springer-Verlag.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Veitch, E. (2005). Tackling publication bias in clinical trial reporting. PLoS announces the launch of a new online journal. *PLoS Med* **2**, e367.

Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Wang, T., and Elston, R. C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *American Journal of Human Genetics* **80**, 353–360.

Wei, Z., Li, M., Rebbeck, T., and Li, H. (2008). U-statistics-based tests for multiple genes in genetic association studies. *Annals of Human Genetics* **72**, 821–833.

Whitehead, J. (1986). On the bias of maximum-likelihood-estimation following a sequential test. *Biometrika* **73**, 573–581.

Wright, F. A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., Barry, W. T., de Villena, F. P., Sullivan, P. F., Wilhelmsen, K. C., and Zou, F. (2007). Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics* **23**, 2581–2588.

Xiao, N., Zhang, D., and Zhang, H. H. (2008). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, Tentatively accepted.

Zeng, D., Lin, D. Y., Avery, C. L., North, K. E., and Bray, M. S. (2006). Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics* **7**, 486–502.

Zhang, H. H. (2006). Variable selection for support vector machines via smoothing spline ANOVA. *Statistica Sinica* **16**, 659–674.

Zhang, H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

Zhang, H. H., Liu, Y., Wu, Y., and Zhu, J. (2008). Variable selection for multicategory SVM via sup-norm regularization. *Electronic Journal of Statistics* **2**, 149–167.

Zhao, L. P., Li, S. S., and Khalid, N. (2003). A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics* **72**, 1231–1250.

Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621–634.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265–286.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36**, 1509–1533.

Zou, H. and Zhang , H. (2008). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, To appear.

8 PROTECTION OF HUMAN SUBJECTS

Although the proposed research indirectly involves human subjects through the preparation, in Core B, of de-identified data sets from identifiable patient data sources, the investigators on Project 4 will have access only to the de-identified data. Thus, the investigators on Project 4 will have no access to any identifiable patient information.

9 INCLUSION OF WOMEN AND MINORITIES

The methods we develop will be applicable to studies with both women and minorities and also to studies which examine treatment differences adjusted for gender, ethnicity and race. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, many of the existing data sets to be studied and provided by Core B include women and minorities, although we will not be generating any new data involving human subjects.

10 TARGETED/PLANNED ENROLLMENT TABLE

N/A

11 INCLUSION OF CHILDREN

The methods we develop will be applicable to studies with children and also to studies which examine treatment differences adjusted for age. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, some of the existing data sets to be studied and provided by Core B may include children, although we will not be generating any new data involving human subjects.

12 VERTEBRATE ANIMALS

N/A

13 SELECT AGENT RESEARCH

N/A

14 MULTIPLE PD/PI LEADERSHIP PLAN

N/A

15 CONSORTIUM/CONTRACTUAL ARRANGEMENTS

If the present application is funded, the University of North Carolina at Chapel Hill will execute subcontracts with the consortium institutions (Duke University and North Carolina State University). These inter-institutional agreements will be written consistent with the NIH consortium agreement policy.

16 LETTERS OF SUPPORT - None

17 RESOURCE SHARING PLAN(S)

(a) Data sharing plan: The data-related resources generated by the proposed research consists of new statistical methodology, software packages for implementation of the methodology, and tutorials for the software. The statistical methodology will be shared through peer reviewed publications and national meetings and

through other standard means. All accepted publications will be deposited in PubMed Central in accordance with the NIH Public Access Policy. Summaries of the methodology, the software and tutorials will be shared through a public web site managed by Core A, while Core C will assist in preparation of the software and tutorials for dissemination. This project will use de-identified data prepared by Core B to test the methods and to create demonstrations of use of the methods to be included in tutorials. This project will not be involved in sharing of these data; this function will be addressed by Core B.

(b) Sharing model organisms: N/A

(c) GWAS: N/A

PROJECT 5

METHODS FOR DISCOVERY AND ANALYSIS OF DYNAMIC TREATMENT REGIMES

Project Leader: Anastasios A. Tsiatis, PhD

PROJECT SUMMARY (See instructions):

Treatment of cancer is an ongoing process during which clinicians make a series of therapeutic decisions over the course of the disease. However, while there is increasing interest in identifying the overall strategy of sequential decisions leading to the most beneficial clinical outcomes, where those decisions may be predicated on complex information on the patient up to that point, current cancer clinical trials evaluate only the therapeutic options available at a single decision point, mostly in a "one-size-fits-all" manner. Attempts to synthesize information from several isolated trials conducted at different milestones in the disease are problematic, because the best treatment at any one decision point may not be best when placed in the context of the entire decision process owing to possible delayed effects of past treatments on the efficacy of future treatments. Considering cancer treatment strategies as dynamic treatment regimes, which are formal algorithms for sequential decision-making that use accrued information on the patient at each decision point in an evidence-based manner to determine the next step of treatment, along with analytical reinforcement learning methods from computer science that provide a principled framework for identifying the optimal such regime, offers the potential to revolutionize how cancer treatment is viewed and effect a paradigm shift in the design and conduct of cancer clinical trials. The four specific aims of this project seek to catalyze this advance by studying these issues for the first time in the cancer treatment context. The first aim will evaluate various learning methods to establish the best techniques for use in developing optimal dynamic treatment regimes for cancer, and the second will focus on a specific version of this methodology when clinicians are interested in finding the best regime among a particular set of regimes. The third aim will develop new methods for making formal statistical inference on regimes developed based on data, which have been heretofore unavailable owing to the theoretical complexity of the problem. In the fourth aim, methods for design of so-called sequentially randomized trials for the specific purpose of developing dynamic treatment regimes, including determination of sample sizes that will ensure identification of the best regimes from among those in the trial, will be developed. Coupling trial design with learning methods for analysis, a new model, the clinical reinforcement trial, will be developed and applied to designing studies to identify optimal regimes for non-small cell lung cancer and other cancers. Collectively, these aims will result in high-impact, new methodology that will allow individualization of the therapy to the patient over time.

RELEVANCE (See instructions):

Although treatment of cancer involves a series of therapeutic decisions over time, cancer clinical trials evaluate treatments only at specific decision points, and hence the best treatment in such a trial may not be best when placed in the context of the overall decision-making process. This research will study cancer treatment formally as an overall, individualized strategy so that the entire series of decisions leading to the best outcomes can be determined, promoting a paradigm shift in the way cancer therapies are evaluated.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location		
Organizational Name: The University of North Carolina at Chapel Hill		
DUNS: 608195277		
Street 1: Office of Sponsored Research, CB #1350	Street 2: 104 Airport Dr., Suite 2200	
City: Chapel Hill	County: Orange	State: NC
Province:	Country: USA	Zip/Postal Code: 27599-1350
Project/Performance Site Congressional Districts: NC-004		

Additional Project/Performance Site Location		
Organizational Name: North Carolina State University		
DUNS: 042092122		
Street 1: Research Admin/ SPARCS	Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514	
City: Raleigh	County: Wake	State: NC
Province:	Country: USA	Zip/Postal Code: 27695-7514
Project/Performance Site Congressional Districts: NC-02		

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Tsiatis, Anastasios	butch_tsiatis	NC State University	Project 5 Leader
Bondell, Howard D.		NC State University	Co-Investigator
Boos, Dennis D.	dennis_boos	NC State University	Co-Investigator
Davidian, Marie	davidian	NC State University	Project Co-Leader
Kosorok, Michael R.	Michael_Kosorok	UNC-CH	Project Co-Leader
Socinski, Mark A.		UNC-CH	Co-Investigator
Stefanski, Leonard A.		NC State University	Co-Investigator
Zhang, H. Helen		NC State University	Co-Investigator

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project

Human Embryonic Stem Cells ☒ No ☐ Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

RESEARCH PLAN

1 INTRODUCTION TO RESUBMISSION/REVISION APPLICATION- N/A

2 SPECIFIC AIMS

In clinical practice, treatment of cancer is a dynamic process involving a series of therapeutic decisions over time. However, most cancer clinical trials focus on effects of treatments given at a single decision point in the course of the disease, e.g., the selection of a first-line chemotherapeutic option for patients with Stage IIIB/IV non-small cell lung cancer. Conclusions on the best overall strategy over the series of key decision points in the disease are consequently cobbled together from the results of many such single-decision studies, and, due in part to the possibility that the treatment given at one point in time may have delayed effects on the efficacy of future treatment, may be misleading and, indeed, deleterious. In some chronic disease/disorder areas, notably behavioral disorders and infectious diseases, there has been a growing recognition that this myopic point of view may not result in patients receiving the best sequence of treatments and that the entire sequential decision-making process must be studied as a whole in order to identify strategies that are the most beneficial.

This perspective has led to considerable recent interest in methodology for developing and studying *dynamic treatment regimes*. A dynamic treatment regime is a set of sequential decision rules dictating at each decision point the selection of the next treatment for a patient based on information on the patient, including measures of disease progression, biomarkers, and previous treatment, up to that point, thereby individualizing each step of treatment to the patient. With more than one option at each decision point and numerous possibilities for synthesizing the available information at each into decision rules, many regimes may be conceived, and identifying the optimal regime, that leading to the most benefit if followed over the course of the disease by the population of cancer patients, presents many challenges. We propose four specific aims that will lead to advances in methodology for discovering and evaluating dynamic treatment regimes:

Aim 1: To develop and evaluate learning methods for optimal dynamic treatment regimes. Because of the complexity of the problem, standard statistical methods are not useful for identification of the optimal regime from data. *Reinforcement learning methods* from computer science, adapted to incorporate statistical inference, are a promising and powerful approach to this problem. However, although they have been used extensively in areas such as artificial intelligence and robotics, only limited research has been conducted on their use for developing optimal treatment strategies for human diseases and disorders, and no work has been done to evaluate their feasibility and application in developing optimal regimes for cancer. We will carry out the first, comprehensive study of competing learning methods in the context of cancer research.

Aim 2: To develop methods for identifying optimal dynamic treatment regimes from a restricted, feasible set. A key challenge in identification of the optimal dynamic treatment regime is that, with many decision points, treatment options, and high-dimensional patient information, the number of possible regimes can be enormous. An alternative, practical approach would be to restrict the candidate regimes to a smaller, feasible set based on considerations including current clinical practice, cost, and complexity. We will develop methods for estimating population mean outcome for regimes within a feasible set and for identifying the best regime within the set.

Aim 3: To develop and evaluate inferential methods for dynamic treatment regimes. Methods for making inference on optimal dynamic treatment regimes derived from the learning techniques in Aim 1 pose a significant challenge in that parameters in the statistical models that characterize these regimes are often constrained to lie on the boundary of the parameter space. Standard inferential approaches, including bootstrap methods, break down under these conditions, and a fundamentally new statistical framework is needed. We will develop methods for constructing hypothesis tests and confidence intervals for optimal dynamic treatment regimes.

Aim 4: To develop methods for the design of sequentially randomized trials for dynamic treatment regimes. Sequentially randomized clinical trials, where subjects are involved in multiple randomizations to therapeutic options at each of several decision points, have been advocated for evaluating specific dynamic treatment regimes and for developing optimal regimes. Numerous sequentially randomized cancer trials have been conducted, demonstrating the feasibility of carrying out such studies in practice, although they have not had the goal of investigating the dynamic treatment regimes embedded in them. We will develop a new model for cancer clinical trials, *clinical reinforcement trials*, which involve sequential randomization, allow for a contin-

uum of treatment options, and have the goal of developing optimal regimes using learning techniques. We will apply these first to non-small cell lung cancer and generalize to other cancers. A key challenge in the design of sequentially randomized studies for deducing optimal regimes is that, as the number of decision points and treatment options at each grows, the greater the sample size requirements can be. We will develop new approaches to evaluating the properties of these designs that will enable determination of numbers of decision points and treatment options that can be studied with sufficient precision using realistic sample sizes.

The overarching goal of this project is to catalyze a paradigm shift in the way cancer therapies are conceived and evaluated that has the potential to make evidence-based, individualized treatment strategies a reality.

3 BACKGROUND AND SIGNIFICANCE

3.1 Dynamic Treatment Regimes

Treatment of cancer is an ongoing process. For example, a patient with untreated acute promyelocytic leukemia might first receive concurrent tretinoin and chemotherapy. If the patient responds to this treatment, s/he might be continued on intermittent tretinoin as a maintenance therapy; if s/he does not, the clinician may prescribe a second-line chemotherapy. Clinicians routinely modify therapy in the face of toxicity, reducing the dose or delaying initiation. Essentially, cancer treatment in practice involves a series of decisions made sequentially over time based on accruing information on the patient. Moreover, in this revolutionary era of advances in biology, the possibility that biomarkers and genetic and genomic information may be used with other baseline or evolving patient characteristics to guide treatment decisions is seemingly within reach.

Despite this, evaluation of cancer treatment is overwhelmingly carried out through standard clinical trials of competing treatment options at a single decision point. Although such studies provide important information on how treatments compare at that decision point, they are not designed to address questions regarding how patients fare over the course of an entire sequence of decisions, nor of how patient information may inform those decisions. Nonetheless, there is great interest in establishing guidance on the optimal treatment strategy involving a series of such decisions; that is, deducing that strategy that would lead to the best outcomes if followed by the population of patients. For example, Grossi et al. (2008) review results of studies of first-, second-, and third-line therapies for advanced non-small cell lung cancer (NSCLC) and attempt to synthesize this evidence to recommend an optimal strategy for using these in practice. A key limitation of this approach is that the apparent "best" treatment at a particular decision point as determined by a study at that point may not be the best when placed in the context of prior and subsequent decision points and patient information. A treatment may have prolonged effects that have implications for the efficacy of future treatments, and therapies that may be beneficial over the course of the disease for one patient may differ from those for another. The authors advocate formal study of "rigorous treatment algorithms" involving the series of three decisions.

There is increasing appreciation, particularly in study of management of chronic behavioral disorders such as depression and drug and alcohol dependence (Murphy et al., 2007ab; Pineau et al., 2007) and of HIV infection (El-Sadr et al., 2006), that considering entire strategies involving a number of key decision points is required to determine how best to treat patients over the course of a disease. Establishing treatment algorithms that dictate at each decision point which treatment among the available options to give based on information on the patient up to that point would provide cancer clinicians with principled, evidence-based guidance for individualizing their treatment decisions to the particular circumstances of the patient over time.

This perspective has led to heightened interest in the formal study of dynamic treatment regimes. A dynamic treatment regime may be viewed as an algorithm for sequential decision-making. A regime involves a sequence of decision points at which decisions on treatment, selecting from several options, would be made. At each point, an associated decision rule that takes as input all information on the patient to that point, such as previous treatment history and past and current values of biomarkers and other measures that may reflect disease progression, and outputs the next step of treatment. Clearly, with several treatment options at each decision point and many possible ways of distilling the accruing information to define decision rules, there are numerous possible regimes, and an obvious objective is to determine the regime that leads the best outcomes overall; i.e., if followed by all patients in the population, would result in the most beneficial mean outcome.

A very simple, generic example illustrates. Suppose that there are two first-line chemotherapeutic options, C_1 and C_2, say, for patients at a particular stage of a certain cancer. Among patients who respond to this induction therapy, where "response" may be defined on a cancer-specific basis, it would be standard to begin

a maintenance or intensification treatment; suppose there two such options, M_1 and M_2. For subjects who do not respond, routine practice would be to prescribe a salvage therapy; assume two options S_1 and S_2. In this scenario, there are eight possible dynamic treatment regimes taking the form "Give first-line chemotherapy C_i followed by maintenance therapy M_j if the patient responds; otherwise, if s/he does not, give salvage therapy S_k," where $i, j, k = 1, 2$ in all 8 possible different combinations. In this simple setting, for any given regime, the first decision point occurs at the time of treatment initiation, and the decision rule does not take into account any baseline patient-specific information. The second decision point occurs at the time response is ascertained; here, the decision rule takes the single variable, response (yes or no), as input and assigns the subject to maintenance or salvage therapy accordingly. Among this set of 8 possible regimes, there is a "best" regime defined by a particular combination (C_i, M_j, S_k) in the sense that, if all patients followed it, the greatest mean outcome would be achieved. The decision rules here are primitive; the first uses no information on the patient, while the second uses only information on response to specify treatment. Clearly, developing regimes that employ more sophisticated decision rules synthesizing all information available on the patient would address the goal of individualizing the entire strategy to the patient given the information and options available.

The implications for revolutionizing the treatment of cancer patients are enormous. Major challenges in cancer research are the bottleneck between the laboratory research that suggests new treatments and clinical practice and the best ways to use existing treatments. Taking the view that treatment is a sequential decision-making process and thinking in the context of individualized regimes has the potential to effect a paradigm shift in the way these treatments are evaluated. Among newly conceived candidate treatments, very few make it to human clinical trials, and only 10% of these demonstrate enough efficacy to be approved for marketing (see Food and Drug Administration, 2004; Högberg, 2005). It may well be that, although such a treatment fails in the primary analysis of a traditional, single-decision point clinical trial because its benefit to patients with certain characteristics is "averaged out" with outcomes for patients for whom it is not efficacious, subgroup analysis may reflect this benefit. Were such treatments included as options at an appropriate decision point for such patients in a subsequent study to evaluate and develop dynamic treatment regimes, there is the possibility that, when used with existing treatments as part of an entire strategy, they could emerge as important components of an overall regime. Moreover, such studies might reveal where in the decision-making process and for whom existing treatments are especially advantageous, taking into account their prior and future effects.

With many decision points, treatment options, and high-dimensional patient information, the number of possible regimes is quite large. Given data on treatment decisions over time, intervening information that may have been used to make those decisions, and outcomes for a sample of patients, a new statistical methodology is required to traverse the myriad possibilities and construct optimal regimes. Furthermore, a framework for conception and design of clinical trials to collect such information for this purpose and for evaluating the benefit of specific, pre-conceived regimes is needed. Our four specific aims represent key methodological advances toward these objectives. Taken together, their fundamental premise is that viewing ongoing treatment of cancer patients through the lens of dynamic treatment regimes offers an exciting opportunity for a transformative model for cancer research. While standard, single-decision trials attempt to correct for individual differences and prior history in assessing treatments, development of dynamic treatment regimes leverages patient differences to inform the entire decision-making process, with great implications for improving the treatment of cancer patients.

3.2 Aim 1: Learning Methods for Optimal Dynamic Treatment Regimes

Existing statistical methods do not have the needed flexibility to be used to develop optimal dynamic treatment regimes based on the type of data described in the previous section. Reinforcement learning methods from computer science (e.g., Sutton and Barto, 1998) offer an appropriate framework for this purpose. Reinforcement learning is a powerful artificial intelligence technique that has recently been used to teach an autonomous controller to fly a helicopter upside down in a sustained hover; see Figure 1, demonstrating the potential of reinforcement learning for solving problems that are complex and counter-intuitive (Ng et al., 2006). As posed in computer science, reinforcement learning involves trying a sequence of actions, recording both the long and short term consequences of those actions, estimating the relationship between actions and consequences, and then selecting the "policy," i.e., the sequence of decision rules dictating actions, that results in the most desirable outcomes. The connection to dynamic treatment regimes is evident: "actions" are treatments, "consequences" are evolving measures of health status, including outcomes; and "policies" are dynamic treatment regimes.

Susan Murphy, Chair of the External Advisory Committee for the Program Project, has been instrumental

in bringing reinforcement learning methods to the attention of the statistical community and establishing the connection between these methods and discovery of optimal dynamic treatment regimes (Murphy, 2003; see also Robins, 2004 and Moodie, Richardson, and Stephens, 2007). She has also promoted application of these methods to developing regimes for treatment of behavioral disorders, as noted above. Here, time frames over which response to treatments may be evaluated are quite short, giving patients the opportunity to try many different treatments while experiencing the same severity or stage of the disorder. Things are different for life-threatening diseases such as cancer, where the disease may progress over a single course of treatment. Hence, adaptation of reinforcement learning methods to the particular challenges of cancer treatment is critical.

The work of Dr. Murphy and colleagues represents only initial steps toward harnessing the power of reinforcement learning for developing optimal regimes. There has been only limited assessment of how learning methods should be used in practice or of the relative advantages of different methods, and none in the setting of cancer research. In this aim, we will carry out a comprehensive study of these promising techniques in this context.

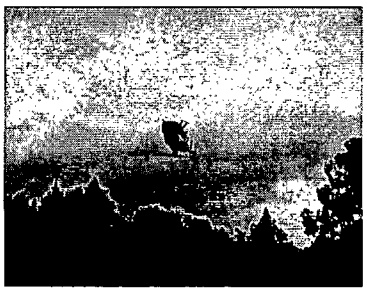


Figure 1: Helicopter in autonomous sustained hover. Figure 3 from Ng et al. (2006)

Different learning methods have been proposed; chief among these are Q-learning (Watkins, 1989; Watkins and Dayan 1992) and A-learning (Murphy, 2003; Blatt et al., 2004); formal descriptions are given in Section **5.2**. Little is known about their relative merits (Almirall et al., 2005). These methods involve developing statistical models for patient outcomes and other quantities as a function of past history, but how best to develop these models, particularly in the face of high-dimensional histories, is not known. A number of methods are available for model selection, including traditional statistical methods such as forward selection; shrinkage methods familiar in the statistical literature, such as the Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996), the adaptive LASSO (Zou, 2006; Wang and Leng, 2007; Zhang and Lu, 2007), the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001), and the False Selection Rate (FSR) methods developed by members of our team (Wu, Boos, and Stefanski, 2007; Boos, Stefanski and Wu, 2008, Crews, Boos, and Stefanski, 2008); and methods developed in the machine learning literature, such as support vector regression (Vapnik et al., 1997) and extremely randomized trees (Ernst, Geurts, and Wehenkel, 2005; Geurts, Ernst, and Wehenkel, 2006). A study of the performance of these methods in the context of Q-learning and A-learning would be valuable. During the project period, we will carry out a systematic study of all of these issues.

Finally, the nature of the data must be considered. In some settings, e.g., prospective cohort studies, data that are observational in nature may be available that record measures of disease progression, outcomes, and treatments. Data from a single-decision clinical trial may include follow-up information on disease course and subsequent treatments received. The extent to which learning methods can be used with these data must be explored. A much more fruitful mechanism for obtaining the needed data is to conduct clinical trials with the specific goal of discovering optimal dynamic treatment regimes. In Aim 4, we study sequentially randomized trials for this purpose, and propose a promising new type of such trial, the clinical reinforcement trial, that is specifically designed to yield the information necessary to develop optimal regimes of interest in cancer. Our comprehensive study of learning methods will include evaluation of the performance of the methods with both types of data and will be a critical preliminary step toward the development of these new trials in Aim 4.

3.3 Aim 2: Identifying Optimal Dynamic Treatment Regimes From a Restricted, Feasible Set

As we have noted, the number of possible dynamic treatment regimes can be very large. More precisely, with many decision points, an array of treatment options at each, and, most importantly, the high dimensionality of the patient information to be used in constructing rules, there is effectively an infinite number of regimes depending on high-dimensional information. The reinforcement learning methods discussed in Aim 1 focus exclusively on finding the optimal regime that distills the full complement of patient information into decision rules. The resulting regime may thus involve rules that are difficult to interpret in practice, as they may incorporate complex, non-intuitive combinations of patient information, and may require complicated manipulation of that information to obtain the next step of treatment. This may make the optimal regime less appealing to some practitioners. Even if this is not the case, because the goal of learning methods is only to identify the optimal regime, they do not provide a framework for exploring the relationship of mean outcome to the factors involved in the decision rules.

For example, if the rule at a particular decision point output the dose of chemotherapy to administer, clinicians may be interested in understanding the extent to which mean outcome changes as a function of that dose. If mean outcome is not sensitive in a clinically meaningful way to the dose given at that point for a range of feasible doses, giving a lower dose on the basis of cost or toxicity may be preferred.

These observations inspire us to develop an alternative framework for developing dynamic treatment regimes that attempts to circumvent the issues associated with high-dimensional information and to accommodate the situation where clinicians are interested in developing regimes that depend on a key subset of patient information on grounds of cost and clinical practice. We will develop methods for estimating population mean outcome for regimes that are restricted by such considerations. The ability to estimate mean outcome for a slate of competing regimes addresses the second issue above and allows identification of the optimal regime within the restricted set. An important part of this effort will be to compare regimes so identified to the optimal regime derived via learning methods under realistic conditions to assess the extent to which such a restriction is useful.

3.4 Aim 3: Inferential Methods for Dynamic Treatment Regimes

The question of how to conduct statistical inference in the context of reinforcement learning is crucial for ensuring that optimal dynamic treatment regimes are identified with acceptable precision from observed data. Existing approaches for accomplishing this for Q-learning are of two general types in computer science. The first line of attack is to develop rules for establishing broad bounds on "generalization error," which assesses the accuracy of an estimator for the difference between the optimal regime (or "policy") and any given regime (see Bartlett and Tewari, 2007; Murphy, 2005). Note that generalization error in this context is somewhat different from the use of the term in the machine learning literature (e.g., Laber and Murphy, 2008). Unfortunately, generalization error bounds tend to be much too conservative to be of practical use in clinical trial design.

The second approach is to consider inference for the parameters in the models that must be developed in implementation of learning methods. This approach is easier to use in our setting to obtain interpretable clinical quantities, and, moreover, practical generalization error bounds can often be derived from parametric inference. Hence, we will restrict our attention to the parametric inference approach. The available work in this area is very recent, and is challenged by the non-regularity of the associated estimators (Chakraborty, Strecher, and Murphy, 2008). The issue is that the optimal regime is usually obtained by maximizing over an estimated regression function, and these maxima end up being non-differentiable functions of estimated parameters. To circumvent this problem, Chakraborty et al. (2008) use smooth approximations to these functions based on a technique called soft-thresholding. While this appears to facilitate construction of confidence intervals, the approach also leads to a decrease in the accuracy of estimation because the models are intentionally misspecified to avoid discontinuities. Thus, there is still a great need for development of methods for inference that avoid soft-thresholding, provided that the technical challenges can be overcome.

The problem is closely related to the classic statistical problem of estimating parameters that are constrained by a boundary. The difficulty is that the limiting distribution is different when the parameter is on the boundary versus when it is off, and, it is not known in advance whether or not the parameter is on the boundary. One approach is to carry out a test of the being on the boundary and then conduct inference based its conclusion. This process generates the so-called "post-model-section estimators," which have a long history (e.g., Bancroft, 1944; Bancroft and Han, 1977; Sen, 1986). Leeb and Pötscher (2006) show that this approach can lead to guaranteed inconsistency, depending on the goals of inference and the underlying model. The issue is that the resulting estimators are non-regular, and non-regular estimators can perform arbitrarily poorly in certain settings (Leeb and Pötscher, 2006; Kosorok, 2008, section 18.1). On the other hand, non-regular estimators can also have good properties in certain settings, such as shrinkage estimation of high-dimensional covariance matrices (for a review, see Schäfer and Strimmer, 2005; see also Bickel and Levina, 2008).

To summarize, the problem of finding valid methods of inference for reinforcement learning is important but unresolved. There are a number of new ideas (e.g., empirical process methods) and old ideas (such as post-model-selection estimators), which, when combined, hold promise for addressing this problem. During the project period, we will develop inferential methods based on this strategy.

3.5 Aim 4: Design of Sequentially Randomized Trials for Dynamic Treatment Regimes

Our first three aims will lead to significant advances in techniques for discovering and making inference on dynamic treatment regimes from data. As we have noted, these methods may be applied to observational

data from existing sources; however, because these data were not collected with this purpose in mind, they are likely to lack key information that would support the development of the most effective regimes possible. A more fruitful strategy is to carry out trials that are explicitly designed not only to evaluate a finite, pre-conceived set of relatively simple regimes but also to collect information that can form the basis for the development of more sophisticated regimes via learning methods and the methods in Aim 2. Such designs, which involve sequential randomization of subjects to treatment options at each of a number of pre-determined decision points, have been advocated by Lavori and Dawson (2004), Murphy (2005), Thall et al. (2007), and many others, by whom they have been referred to as "Sequential Multiple Assignment Randomized Trials" (SMART). In behavioral research, prominent such studies have included Sequenced Treatment Alternatives to Relieve Depression (STAR*D) (Rush et al., 2004) and Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) (Lieberman et al., 2005).

For example, in the simple setting of the two-decision point regimes in Section **3.1** defined as "Give first-line chemotherapy C_i followed by maintenance therapy M_j if the patient responds; otherwise, if s/he does not, give salvage therapy S_k," which we write in shorthand as $C_iM_jS_k$, where $i, j, k = 1, 2$ in all 8 possible different combinations, such a trial to evaluate these 8 possible, simple regimes would naturally involve two randomizations. The first would be to one of the two first-line chemotherapy options, C_1 or C_2, at baseline. The second would correspond to the decision point defined by response; at the time a patient responds, if s/he does, s/he would be randomized to one of the two maintenance options M_1 or M_2; patients who do not respond by some maximum time would be randomized to one of the two salvage options S_1 or S_2. Figure 2 shows the trial design schematically, where "•" *represents randomization.* Under this randomization scheme, there will be subjects whose "realized treatment experience" is consistent with at least one of the 8 regimes of interest; e.g., a subject who is randomized initially to C_1, responds, and is then randomized to M_1 has an experience that could have resulted from having followed *either* of the regimes $C_1M_1S_1$ or $C_1M_1S_2$. Thus, subjects in such a trial can provide information on more than one regime, which can be exploited in their design; see Section **5.5**. Of course, it should be clear that, as the number of decision points and/or options at each increase, the numbers of subjects whose realized experience is consistent with a given regime could become quite small, making the design of these studies challenging. In our first sub-aim, we propose a framework for the design of sequentially randomized trials for the purpose of identifying the "best" regime among those represented. In particular, we will develop criteria and approaches for sample size calculation that will allow trialists to determine the numbers of decision points and treatment options at each that may be incorporated while achieving acceptable precision.

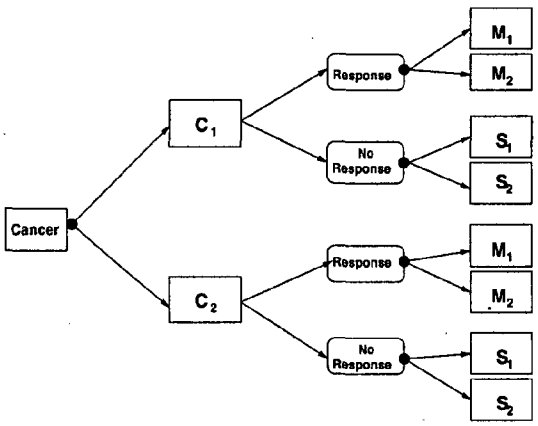


Figure 2: Schema for a sequentially randomized trial with 2 decision points.

It is worth noting that there have been numerous sequentially randomized cancer trials, many of the first of which were designed by Stephen George, one of the PD/PIs for the Program Project. For example, two key, recent Cancer and Leukemia Group B (CALGB) studies that used sequential randomization are CALGB 9710, a phase III trial of concurrent tretinoin and chemotherapy with or without arsenic trioxide as initial consolidation therapy followed by maintenance therapy with intermittent tretinoin versus intermittent tretinoin plus mercaptopurine and methotrexate for patients with untreated acute promyelocytic leukemia; and CALGB 19808, a phase III trial of induction chemotherapy with or without MDR-modulation with PSC-833 followed by cytogenetic risk-adapted intensification therapy followed by immunotherapy with rIL-2 vs. observation in previously untreated patients with AML < 60 years old. Both of these studies had essentially the same design: An initial randomization to one of the two options for induction therapy followed by a subsequent randomization to one of the two maintenance therapy options for those patients responding to the induction therapy. Patients not responding had only a single option, care as prescribed by their primary physicians, and hence were not randomized. It should be clear that in each of these trials, there are four embedded regimes of the form, using the above generic notation, "Give induction therapy C_i followed by maintenance therapy M_j if the patient responds; otherwise, if s/he does not, send to primary physician," for the 4 combinations of $i, j = 1, 2$. Although such trials have been

commonplace for quite some time, the analysis of these trials has focused on separate analyses comparing the induction treatments without regard to subsequent treatment and the maintenance therapies among all responders on either induction treatment and not on the evaluating the embedded regimes. Methods for doing so are available; see Lunceford, Davidian, and Tsiatis (2002), Wahed and Tsiatis (2004, 2006), and Murphy (2005). In any event, that these trials have been executed successfully demonstrates that sequentially randomized trials for the purposes we advocate are logistically feasible in cancer research.

In simple sequentially randomized studies like these, ideally, in the intervening periods between decision points, all patient information that might be important for determining the treatment option to be given at the next decision point would be recorded. The resulting data set would be a rich resource for application of learning methods for identifying the optimal regime. It is in this spirit that we propose our second sub-aim, in which we advocate trials that are conceived with this objective, integrating the goal of using learning methods into the design, which we term clinical reinforcement trials. Such trials involve a series of decision points, which are either specific time points measured from trial onset or decision points in the treatment process such as starting times of each new line of cancer treatment. For each decision time, a set of possible treatments to be randomized is identified. The treatment options can be a continuum, e.g., doses or initiation times (see below), or a finite set, and can be assigned according to pre-specified decision rules that involve accruing patient information. A utility function is identified that can be assessed at each time point and contains an appropriately weighted combination of outcomes available at each interval between decision times and at the end of the final treatment interval. Such designs extend SMART designs in allowing a continuum of treatments. We will develop these trials first for NSCLC and then for other settings such as breast cancer, lung cancer, and ovarian cancer. Our goal is to conceive and, in collaboration with Dr. Socinski at the University of North Carolina (UNC-CH), lay the groundwork for carrying out such a trial in NSCLC, the background for which we now describe.

For NSCLC patients who present with a good performance status and stage IIIB/IV disease, platinum-based chemotherapy is the primary treatment, which has been shown to offer a modest survival advantage over best supportive care alone in single-decision trials. Approximately 40-50 percent of patients in recent first-line trials received second-line treatment. Some patients who maintain a good performance status and tolerate therapy without significant toxicities will receive third-line therapy (Stinchcombe and Socinski, 2008). First-line treatment primarily consists of platinum-based doublets that include cisplatin, gemcitabine, pemetrexed, paclitaxel, carboplatin, or vinorelbine. Numerous studies have compared these various platinum doublets, and the great majority of these trials have concluded that all such regimens appear to be comparable in their clinical efficacy. In addition to platinum-based doublets, some phase III studies have examined the efficacy of various targeted therapies, with both mixed (Sandler et al., 2006) and positive (Pirker et al., 2008) results. The strategies of first-line treatment are essentially based on these four targeted combination therapies, where the choice depends on a number of factors, including histology type, toxicity profile, smoking history, VEGF level, EGFR expression, and race (Scagliotti et al., 2008; Sandler et al., 2006; Pirker et al., 2008). Similar to the first-line therapy, the three approved second-line agents (docetaxel, pemetrexed, and erlotinib) appear to have similar response and efficacy for overall survival outcomes but very different toxicity profiles. Choice of agent also depends on many factors, including prior chemotherapy history, risk for neutropenia, EGFR expression, and patient preference (Shepherd et al., 2000; Hanna et al., 2004; Fidias et al., 2007; Ciuleanu et al., 2008).

In addition to the complexity of the problem of selecting compounds for first- and second-line treatments based on prognostic factors, another primary challenge is to determine the optimal starting time for the second-line therapy, either immediately or delayed after induction therapy, yielding the highest overall survival probability. Although Fidias et al. (2007) demonstrate that immediate transition to second-line therapy using docetaxel is better than waiting for relapse, whether these findings are specific to docetaxel or whether some brief delay for patient recovery would be beneficial for overall survival remains unknown, and this has not been studied in the context of an overall strategy. A main focus of this sub-aim will be to develop a trial design for NSCLC based on these and other considerations to that will facilitate development of dynamic treatment regimes to assign first- and second-line treatment and timing of second-line treatment based on individual-specific prognostic factors to optimize outcomes for NSCLC patients.

4 PRELIMINARY STUDIES

4.1 Investigators

The research associated with each of the four aims will be conducted by a highly qualified group of investigators from North Carolina State University and UNC-CH. Anastasios A. Tsiatis (NCSU), a Co-Director for the overall Program Project, will serve as the Project Leader. Dr. Tsiatis is an internationally recognized expert on semiparametric theory (Tsiatis, 2006), causal inference and dynamic treatment regimes (e.g., Lunceford et al., 2002; Wahed and Tsiatis, 2004, 2006; Tsiatis, 2006, chapter 13; Johnson and Tsiatis, 2004, 2005; Zhang et al., 2009), and clinical trials (e.g., Tsiatis et al., 2008; Zhang, Tsiatis, and Davidian, 2008), making him well-suited to lead the project. Marie Davidian (NCSU) and Michael R. Kosorok (UNC-CH) will serve as Co-Leaders of the project, both of whom are PD/PIs for the overall Program Project. In addition to the joint research with Dr. Tsiatis on causal inference and dynamic treatment regimes reviewed above, which cements her expertise in these areas, Dr. Davidian has additional work in causal inference (e.g., Lunceford and Davidian, 2004) that provides her with excellent background for collaborating on this project. In addition, Dr. Davidian is an investigator on a large project funded by the National Institute of Allergy and Infectious Disease (NIAID) using mechanistic mathematical models of HIV infection dynamics combined with statistical models to design dynamic HIV treatment regimes using mathematical control theory and clinical trials to evaluate them (e.g., Rosenberg, Davidian, and Banks, 2007). This work does not overlap with the approaches taken in this project, but offers a complementary perspective on approaches to developing dynamic treatment regimes. Dr. Kosorok has expertise in areas including clinical trials, high-dimensional data, and, notably, empirical processes and semiparametric inference (Kosorok, 2008), and applications of empirical processes to statistical learning in biomedicine (Zhao, Kosorok, and Zeng, 2008). Dr. Kosorok will apply his expertise to the reinforcement learning approach to design and analysis of clinical trials to develop dynamic treatment regimes. He will also exploit use empirical process methods for developing valid methods of inference for dynamic treatment regimes.

Howard Bondell (NCSU) will lead Aim 1, working with Dennis Boos (NCSU), Dr. Davidian, Len Stefanski (NCSU), Dr. Tsiatis, and Helen Zhang (NCSU). Dr. Bondell is an expert in model selection (e.g., Bondell and Reich, 2008,2009; Bondell and Li, 2009), including nonparametric model selection (Storlie et al., 2008), which will play a prominent role in this part of the project. Drs. Boos and Stefanski are also both experts in model selection and are the originators of the FSR approach (Luo, Stefanski, and Boos, 2006; Wu et al., 2007; Boos et al., 2008, Crews et al., 2008) that we will consider. Dr. Zhang is a recognized authority in model selection and especially nonparametric model selection, machine learning, and data mining (e.g., Zhang et al., 2004; Zhang, 2006; Lin and Zhang, 2007; Zhang and Lu, 2007; Liu et al., 2007; Zou and Zhang, 2009), and will contribute her expertise in these areas and in particular support vector machines.

Dr. Tsiatis will lead the efforts on Aim 2, working with Drs. Davidian, Kosorok, and Zhang, who will contribute expertise on causal inference, dynamic treatment regimes, empirical processes and semiparametrics, and model selection. Dr. Kosorok will lead Aim 3, working with Dr. Tsiatis and Donglin Zeng (UNC). Drs. Kosorok and Tsiatis will draw on their aforementioned expertise in empirical processes and semiparametrics. Dr. Zeng, who has expertise is in semiparametric inference and high-dimensional data, (e.g., Johnson, Lin, and Zeng, 2008; Zhao et al., 2008), will provide additional qualifications for addressing this aim. Aim 4 will be led by Dr. Kosorok, working with Drs. Davidian, Tsiatis, and Zeng, all of whom will contribute their expertise in causal inference, clinical trials, dynamic treatment regimes, and empirical processes and semiparametrics.

Mark Socinski, MD, Professor of Medicine and a member of the Lineberger Comprehensive Cancer Center (LCCC) at UNC, will collaborate closely with the team, providing a critical subject-matter perspective. Dr. Socinski is a medical oncologist with a wealth of experience in cancer treatment and cancer clinical trials, especially in NSCLC (Socinski et al., 2008; Stinchcombe and Socinski, 2008). He has also served as Chair of the Data Safety and Monitoring Subcommittee of the Protocol Review Committee for the LCCC. Dr. Socinski will advise the investigators on the design of simulation studies so that the scenarios evaluated are realistic reflections of what would be expected or would be feasible in cancer clinical trials. He will work closely with Drs. Kosorok and Zeng on the design of a clinical reinforcement trial in Stage IIIB/IV NSCLC; see Section **5.5**.

4.2 Preliminary Studies

Because our four aims together constitute an integrated plan for study of dynamic treatment regimes in cancer, rather than recount preliminary work of the investigators on a by-aim basis, we provide an overall review.

Drs. Tsiatis and Davidian have considerable experience with the study of dynamic treatment regimes. In Lunceford et al. (2002), Drs. Tsiatis and Davidian were among the first to recognize that the sequentially randomized trials conducted by CALGB could be exploited to evaluate the dynamic treatment regimes embedded within them; they proposed methods for analysis, inspired by CALGB protocol 8923, a two-decision-point study similar to the CALGB trials in Section **3.5** in older patients with de novo acute myeloid leukemia. These methods were extended to yield more efficient inferences in Wahed and Tsiatis (2004, 2006). Dr. Tsiatis was instrumental in recognizing that key post hoc clinical trial questions could be cast as dynamic treatment regime problems. In Johnson and Tsiatis (2004, 2005), inspired by a cardiovascular disease (CVD) trial, he developed methodology for determining the optimal duration of therapy from a trial in which duration was left to the discretion of clinicians by considering different duration regimes. In Zhang et al. (2008), Drs. Tsiatis and Davidian, motivated by another CVD trial, posed the problem of subjects who discontinue or switch assigned therapy in terms of dynamic treatment regimes, which permitted development of valid methods for comparing the treatments in the trial in the hypothetical case that all subjects had followed their assigned therapies.

Dr. Tsiatis and Davidian have also been involved in spearheading numerous activities focusing on dynamic treatment regimes. They served as investigator and co-PI, respectively, in a successful 2005 R21 grant application by Dr. Murphy to establish the multidisciplinary "Methodological Challenges in Developing Adaptive Treatment Strategies Network (MCATS)," comprising researchers from diverse areas, including statistics, computer science, engineering, behavioral science, and HIV research. Drs. Tsiatis and Davidian helped organize and were active participants in all of the "brainstorming" workshops conducted by the Network, which led to a Network white paper (Murphy et al, 2007b) and a special issue of the journal *Drug and Alcohol Dependence* devoted to articles on dynamic treatment regimes (e.g., Murphy et al., 2007a; Rosenberg et al., 2007). Dr. Tsiatis was invited in April 2008 by Dr. Alan Krensky, Director of the NIH Office of Portfolio Analysis and Strategic Initiatives (OPASI) to speak to the NIH Institute and Center Directors and other scientists at NIH in the prestigious OPASI Rounds Lecture Series on incorporating study of dynamic treatment regimes into the clinical trials paradigm. His lecture, "Novel Study Designs for Treatment Strategies that Reflect Actual Clinical Practice," was attended by several Directors as well as numerous biostatisticians and clinicians from across the NIH. Drs. Tsiatis and Davidian have also, separately and together, given numerous invited presentations on dynamic treatment regimes at international and national conferences and to audiences of biostatisticians and clinicians at other institutions, including a joint presentation at the Duke Clinical Research Institute in August 2007.

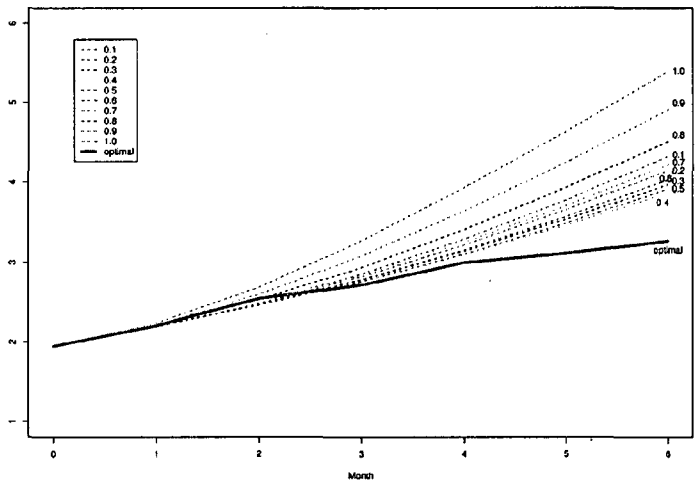


Figure 3: Disease severity (lower is better) as a function of time (in months) for the optimal treatment based on reinforcement learning (solid) versus fixed dose options (dotted).

Drs. Davidian and Tsiatis were Co-Program Leaders, with Dr. Murphy, of the Statistical and Applied Mathematical Science Institute (SAMSI) June 2007 Summer Program on Challenges in Dynamic Treatment Regimes and Multistage Decision-Making. The two-week program, which they organized, brought together many of the world's preeminent authorities from the statistical, computer science, mathematical, and engineering research communities on dynamic treatment regimes and reinforcement learning for an intensive two weeks of activities, including a formal program of tutorials and research presentations followed by a week in which participants formed Working Groups to initiate cross-disciplinary collaboration and formulation of a future research agenda. Drs. Kosorok and Zeng were active participants in the program.

Drs. Kosorok and Zeng also have significant experience with dynamic treatment regimes and reinforcement learning. In October 2007, they initiated a research working group, the "Reinforcement Learning Group," at UNC that meets weekly to work on reinforcement learning, dynamic treatment regimes, statistical learning, and related high-dimensional

problems in biostatistics. The impetus came in part from the SAMSI program. Participants in the Reinforcement Learning Group include Drs. Kosorok, Fine (investigator on Project 2), and Zeng from the UNC Department of Biostatistics, Dr. Liu (investigator on Project 4) from the UNC Department of Statistics and Operations Research, and Dr. W. Wang from the UNC Department of Computer Science, as well as 6–8 students and post-doctoral fellows. Some of the research resulting from this collaboration inspires key components of Aims 3 and 4.

In particular, the Group recently carried out a preliminary assessment of the use of reinforcement learning methods in cancer clinical trials through a simulation study to discover optimal dynamic treatment regimes involving timing and dose for a generic cancer (Zhao et al., 2008). They developed a mechanistic disease model based on a simple difference equation that balances a simulated chemotherapy agent's efficacy and toxicity, from which "virtual patients" could be generated and used this to simulate a virtual, 1000 patient clinical reinforcement trial in which patients were treated with 6 months of chemotherapy. At the beginning of each month, patient quality of life and tumor size were both measured, and a random dose of "chemotherapy" in the interval $[0, 1]$ was assigned subject to constraints, such as requiring the dose to be in $[0.5, 1]$ for the first months, meant to simulate standard-of-care restrictions. Reinforcement learning methods were then used on the virtual data from the trial to find the optimal regime involving timing and dosing based on biomarkers available from the patient. Figure 3 presents the mean disease severity of 200 new simulated patients following the optimal regime so determined (dark solid line) compared to that if the same 200 patients were to follow each of 10 fixed dose regimens (dotted lines) ranging over a grid from highest to lowest. Clearly, the optimal regime is superior by 6 months after initiation of treatment, even though it is not optimal at 2 months. This demonstrates explicitly how considering entire regimes can lead to the proper trade-off between short and long term effects of treatment while adapting to the individual patient. These preliminary results are encouraging; however, considerable work along the lines of the four aims proposed here is required before these methods can be applied in practice.

5 RESEARCH DESIGN AND METHODS

5.1 Notation and Assumptions

Before we describe the research to be carried out for each aim, we establish notation and make explicit assumptions that will be adopted throughout and provide a formal, conceptual framework for identifying and evaluating dynamic treatment regimes on the basis of data from both observational sources and sequentially randomized trials. Assume that there are $T + 1$ time-ordered decision points at which treatment decisions will be made, indexed by $0, 1, \ldots, T$; this indexing convention uses "0" for the first for consistency with popular notation for indexing "baseline" in a clinical trial. These may correspond to fixed times or to events in the disease process that dictate such decisions. In the reinforcement learning literature, the available data are realizations of the time-ordered random variables

$$(S_0, A_0, R_0, \ldots, S_T, A_T, R_T, S_{T+1}), \tag{5.1}$$

where, in the case relevant to our context, S_j denotes the environmental "state", A_j denotes the "action," and R_j denotes the "reward" at the jth decision time, $j = 0, \ldots, T$. For brevity, write $\overline{S}_j = (S_0, \ldots, S_j)$ and $\overline{A}_j = (A_0, \ldots, A_j)$ to denote the histories of the states and observed actions, respectively, up to the jth decision, and let $\overline{S} = \overline{S}_T$ and $\overline{A} = \overline{A}_T$ represent the histories up to the final decision. It is assumed that the reward R_j is a function of state resulting from the history $\overline{A}_j$ of treatment up to the jth time and the ensuing state $\overline{S}_{j+1}$, i.e., $R_j = r_j(\overline{S}_{j+1}, \overline{A}_j)$. Given data on (5.1), the goal of reinforcement learning is to find the sequence of actions as a function of past states and actions, i.e., "policies" $\pi_j(\overline{s}_j, \overline{a}_{j-1}) = a_j$, that will lead to the maximum expected discounted return, where the return is $\sum_{j=0}^{T} \gamma^j R_j$, where $\gamma \leq 1$ is some discount factor that may be interpreted as a control to balance immediate and future rewards; $\gamma = 1$ corresponds to maximizing rewards over the long run. In Section **5.2**, we describe how learning methods address this objective.

In the context of a cancer patient, state S_j corresponds to a vector of patient information (covariates) recorded between the $(j-1)$th and jth decision times, where S_0 is the initial information available; e.g., a vector of baseline covariates in a clinical trial. The action A_j is the treatment given at the jth decision point, where A_j takes on possible values in a set $\mathcal{A}_j$ containing the treatment options available at that point. The policies $\pi(\cdot)$ define what we refer to as g, the dynamic treatment regimes, as we will demonstrate explicitly in Section **5.3**. The reward R_j is some intermediate patient outcome measure ascertained between decisions j and $j + 1$ that contributes to the overall outcome of interest. Because the rewards are determined entirely by the past treatment-covariate history, without loss of generality, we may take the rewards $R_0, \ldots, R_{T-1}$ to be

identically equal to zero; take the last reward to be equal to the outcome of interest (for example, the primary endpoint in a clinical trial) and write $Y = R_T$; and take $\gamma = 1$, in which case, the policy that gives the maximum expected discounted return is exactly the same as the dynamic treatment regime that maximizes the expected outcome. In this case, the data may be viewed as realizations of $(S_0, A_0, \ldots, S_T, A_T, Y)$. In the sequel, we consider both this setting and (5.1); for the following discussion, we adopt the situation of a single outcome Y.

In a sequentially randomized clinical trial, the observed treatments $A_j \in \mathcal{A}_j$, $j = 0, \ldots, T$ received are randomly assigned according to an experimental design with probabilities that may be functions of a patient's past covariate-treatment history $(\overline{S}_j, \overline{A}_{j-1})$. For definiteness, consider the example in Section **3.5**, where $T = 1$. Here, at the first decision point $j = 0$, $\mathcal{A}_0$ consists of the two first-line chemotherapeutic options; i.e., $\mathcal{A}_0 = \{\mathrm{C}_1, \mathrm{C}_2\}$. The next decision point, $j = 1$, corresponds to observation of whether or not the patient has responded; thus, included in the intervening information S_1 is a variable $R = 1$ if a response occurs and $R = 0$ if not. Let $\mathcal{M} = \{\mathrm{M}_1, \mathrm{M}_2\}$ be the set of the two maintenance options, and $\mathcal{S} = \{\mathrm{S}_1, \mathrm{S}_2\}$ that of the two salvage options. Then the set of possible treatment options at $j = 1$ is $\mathcal{A}_1 = (\mathcal{M} \cup \mathcal{S})$; however, patients who respond $(R = 1)$ may only be randomized to one of the options in $\mathcal{M}$, and those who do not $(R = 0)$ only to one of the options in $\mathcal{S}$. Thus, the randomization probabilities are determined by a patient's past history; here, the variable R in S_1. In contrast, the treatments $(A_0, \ldots, A_T)$ that are recorded for a patient in an observational study are not given according to a designed experiment and randomization, but rather are assigned according to a decision process carried out by clinicians (and patients) over time. In either case, the data that are available for analysis are independent and identically distributed (iid)

$$(\overline{S}_{T,i}, \overline{A}_{T,i}, Y_i), \quad i = 1, \ldots, n, \tag{5.2}$$

representing observations from n randomly selected patients from the population of interest.

Recall that a dynamic treatment regime is an algorithm consisting of a collection of decision rules dictating how to treat a patient over time as a function of his/her past history. Formally, a dynamic treatment regime is a function $g = (g_0, \ldots, g_T)$, where, for every $\overline{s}_j$, a realization of $\overline{S}_j$, $g_j(\overline{s}_j) = a_j \in \Phi_j(\overline{s}_j, \overline{a}_{j-1}) \subset \mathcal{A}_j$, and $\Phi_j(\overline{s}_j, \overline{a}_{j-1})$ is the set of treatments that may be given to a patient at j, which may be a function of his/her past covariate-treatment history. Denoting by $\mathcal{G}$ the set of all such dynamic treatment regimes to be considered, the goals in which we are interested are (i) to estimate the mean of Y if the entire population were to follow a particular regime $g \in \mathcal{G}$, and (ii) to identify the optimal regime $g^{\mathrm{opt}} \in \mathcal{G}$ such that the corresponding mean of Y is greatest among all $g \in \mathcal{G}$ (assuming larger values of Y are preferred). We develop methods to achieve these goals in Section **5.2–5.4**.

In order to proceed, we must define precisely the parameter of interest, i.e., "mean outcome if the entire population were to follow regime g." We accomplish this by appealing to the notion of potential outcomes (Neyman, 1923; Rubin, 1974; Robins, 1986) in the causal inference literature. Define the set of potential outcomes by

$$W = [\{S_0, S_1^*(a_0), \ldots, S_j^*(\overline{a}_{j-1}), \ldots, S_T^*(\overline{a}_{T-1}), Y^*(\overline{a})\} \text{ for all } \overline{a} \in \overline{A}], \tag{5.3}$$

where $S_j^*(\overline{a}_{j-1})$ denotes the value of the covariates S_j that would be observed for an arbitrary patient in the population between the $(j-1)$th and jth decision point in the hypothetical situation that the patient were to receive the prior series of treatments $\overline{a}_{j-1}$. Similarly, $Y^*(\overline{a})$ denotes the potential value of the clinical outcome in the hypothetical situation that the patient were to receive the sequence of treatments $\overline{a}$. The $S_j^*(\overline{a}_{j-1})$ and $Y^*(\overline{a})$ are referred to as potential outcomes for obvious reasons. We do not define a potential covariate corresponding to S_0 because covariates prior to the first decision point are assumed not affected by subsequent treatments. We may then define the potential covariates and outcome for patients in the population were they to follow regime g by $\{S_0, S_1^*(g), \ldots, S_T^*(g), Y^*(g)\}$ for $g \in \mathcal{G}$ recursively as

$$S_1^*(g) = S_1^*\{g_0(S_0)\}, \quad S_2^*(g) = S_2^*[g_0(S_0), g_1\{\overline{S}_1^*(g)\}], \quad \cdots, \quad Y^*(g) = Y^*[g_0(S_0), \ldots, g_T\{\overline{S}_T^*(g)\}].$$

The parameter of interest is thus $E\{Y^*(g)\}$, which, for goal (i) above, we wish to estimate for every $g \in \mathcal{G}$ using the observed data (5.2).

To do this, certain assumptions are necessary to identify the distribution of $\{\overline{S}^*(g), Y^*(g)\}$ from the distribution of the observed data $(\overline{S}, \overline{A}, Y)$. The first assumption is the so-called stable unit treatment value assumption (SUTVA, Rubin, 1978),

$$S_j = S_j^*(\overline{A}_{j-1}),\ j = 1, \ldots, T, \text{ and } Y = Y^*(\overline{A}); \tag{5.4}$$

(5.4) states that the observed covariates and outcome are those that would be seen under the treatments actually received. The second assumption is

$$p_{A_j|\overline{S}_j,\overline{A}_{j-1}}(a_j|\overline{s}_j,\overline{a}_{j-1}) > 0 \text{ for all } a_j \in \Phi_j(\overline{s}_j,\overline{a}_{j-1}), \tag{5.5}$$

whenever $p_{\overline{S}_j,\overline{A}_{j-1}}(\overline{s}_j,\overline{a}_{j-1}) > 0$, where $p_X(x)$ denotes the density or conditional density of variables X at the realization x. That is, for any treatment-covariate history that has positive probability of being realized, there must be a positive probability that a patient will receive a treatment from the set $\Phi_j(\overline{s}_j,\overline{a}_{j-1})$. Finally, the most critical assumption is that of no unmeasured confounders; namely,

$$A_j \perp\!\!\!\perp W|\overline{S}_j,\overline{A}_{j-1}, \tag{5.6}$$

where "$\perp\!\!\!\perp$" denotes independence or conditional independence. (5.6) states that, conditional on the past covariate-treatment history through the jth decision point, the probability of treatment assignment at that decision point is independent of all the potential outcomes. For a sequentially randomized trial, (5.6) is guaranteed, as patients are randomized according to probabilities at the jth decision point that are functions only of $(\overline{S}_j,\overline{A}_{j-1})$ and hence are independent of W. However, in an observational study, (5.6) may or may not hold. It would necessarily hold if clinicians make treatment decisions at the jth decision point based only on the information $(\overline{S}_j,\overline{A}_{j-1})$ available up to that point; this premise is certainly reasonable, as treatment decisions cannot be based on potential outcomes that are unavailable. The critical issue is whether or not decisions are based on additional information not captured in $(\overline{S}_j,\overline{A}_{j-1})$ and where this information is correlated with the potential outcomes W; if such data are not captured in the database, then (5.6) may be violated.

Henceforth, we assume that (5.4)-(5.6) hold, recognizing that they, and (5.6) in particular, must be critically assessed with observational data. Under (5.4)–(5.6), Robins (1994) and others have shown that the distribution of the potential outcomes $Y^*(g)$ for a dynamic treatment regime $g \in \mathcal{G}$ may be identified from the distribution of the observed data. There are several approaches for estimating $E\{Y^*(g)\}$, including direct regression modeling such as Robins' g-computation algorithm and g-estimation (Robins, 1994) and inverse probability weighted methods (van der Laan, Murphy, and Robins, 2001). The need for these key assumptions is not stated explicitly in the reinforcement learning literature; however, they are required for learning methods to yield valid conclusions. These assumptions underlie our proposed methods for obtaining the optimal, or at least a "good," dynamic treatment regime from either a sequentially randomized trial or observational data.

5.2 Aim 1: Learning methods for Optimal Dynamic Treatment Regimes

We do not give a comprehensive description of reinforcement learning methods but instead restrict to the special case of their formulation relevant to our setting. A detailed account of the history of reinforcement learning is given in Sutton and Barto (1998); see also Kaelbling et al. (1996). We focus on two particular learning methods, which we now describe; other methods are described, for example, by Rummery (1994).

Q-Learning. In the general case of Q-learning, the data are summarized as realizations of (5.1). Q-learning is a recursive algorithm that can be used under assumptions (5.4)–(5.6) and the setting in Section **5.1** to find the optimal dynamic treatment regime; i.e., g^{opt}, where $E\{Y^*(g^{\text{opt}})\} \geq E\{Y^*(g)\}$ for all $g \in \mathcal{G}$. The Q-learning algorithm in the general case defines sequences of so-called Q-functions, value functions, and policy functions, for $j = T, T-1, \ldots, 0$, given by

$$\begin{aligned} Q_j(\overline{S}_j,\overline{A}_j) &= E\{R_j + V_{j+1}(\overline{S}_j,\overline{A}_{j-1})|\overline{S}_j,\overline{A}_j\}, \qquad (5.7)\\ V_j(\overline{S}_j,\overline{A}_{j-1}) &= \max_{a_j} Q_j(\overline{S}_j,\overline{A}_{j-1},a_j)\\ \pi_j(\overline{s}_j,\overline{a}_{j-1}) &= \arg\max_{a_j} Q_j(\overline{s}_j,\overline{a}_{j-1},a_j), \qquad (5.8) \end{aligned}$$

respectively, where we take $V_{T+1}(\overline{S}_{T+1},\overline{A}_T) = 0$. Desirable properties of Q-learning algorithms have been demonstrated by Jaakkola et al. (1994) and Tsitsiklis (1994). To describe the algorithm here, we consider the case in Section **5.1** where the rewards $R_0,\ldots,R_{T-1}$ are zero and the outcome $Y = R_T$. Write $\widetilde{V}_j = V_j(\overline{S}_j,\overline{A}_{j-1})$ for brevity, and, for notational convenience, let $\widetilde{V}_{T+1} = Y$. The following approach will lead to the optimal dynamic treatment regime; namely

$$g_0^{\text{opt}}(s_0) = \pi_0(s_0), \quad g_1^{\text{opt}}(\overline{s}_1) = \pi_1\{\overline{s}_1,\pi_0(s_0)\}, \quad g_2^{\text{opt}}(\overline{s}_2) = \pi_2[\overline{s}_2,\pi_0(s_0),\pi_1\{\overline{s}_1,\pi_0(s_0)\}],\cdots,$$

and the expected outcome for the optimal dynamic treatment regime will be $E(\widetilde{V}_0) = E\{Y^*(g^{\text{opt}})\}$.

The key to Q-learning is to develop regression models for the Q functions given in (5.7). One might specify the Q-functions in terms of a finite number of parameters; write $Q_j(\overline{S}_j, \overline{A}_j; \psi_j)$, $j = T, T-1, \ldots, 0$. These could be linear or non-linear models including the variables $\overline{S}_j$ and $\overline{A}_j$ as main effects as well as interactions of these; we discuss this shortly. In a "backward" iterative fashion, the estimators $\widehat{\psi}_j$ can be obtained for $j = T, T-1, \ldots, 0$ by solving the estimating equations (optimal generalized estimating equations)

$$\sum_{i=1}^{n} \frac{\partial Q_j(\overline{S}_{ji}, \overline{A}_{ji}, \psi_j)}{\partial \psi_j} \Sigma_j^{-1}(\overline{S}_{ji}, \overline{A}_{ji})\{\widetilde{V}_{(j+1)i} - Q_j(\overline{S}_{ji}, \overline{A}_{ji}, \psi_j)\} = 0, \quad \widetilde{V}_j = \max_{a_j} Q_j(\overline{S}_j, \overline{A}_{j-1}, a_j; \widehat{\psi}_j),$$

where $\Sigma_j(\overline{S}_j, \overline{A}_j) = \text{var}(\widetilde{V}_{j+1} | \overline{S}_j, \overline{A}_j)$ would be represented using posited models. Then define $\pi_j(\overline{s}_j, \overline{a}_{j-1}) = \arg\max_{a_j} Q_j(\overline{s}_j, \overline{a}_{j-1}, a_j; \widehat{\psi}_j)$.

Although the Q-learning procedure seems straightforward, implementation in practice poses challenges. Validity of the method requires correct specification of the models for the Q-functions. This may be especially difficult in high-dimensional situations. Accordingly, the use of model selection methods to determine the Q-function is required; we defer discussion until later in this section.

A-Learning. Advantage learning, or A-learning, is a semiparametric version of Q-learning where fewer assumptions on the Q-functions are necessary in order to derive optimal dynamic treatment regimes. Consequently, the appeal of these methods is that they may yield results that are more robust to model misspecification than the Q-learning methods described above. To illustrate how A-learning works, we consider the case where all treatment decisions are binary; i.e., $A_j = (0, 1)$, $j = 0, \ldots, T$, although the development is easily generalized.

To derive the optimal action at the jth decision point, $\pi_j(\overline{s}_j, \overline{a}_{j-1})$, defined by (5.8), note that it suffices to know only the treatment contrast $C_j(\overline{s}_j, \overline{a}_{j-1}) = Q_j(\overline{s}_j, \overline{a}_{j-1}, 1) - Q_j(\overline{s}_j, \overline{a}_{j-1}, 0)$. Clearly,

$$\pi_j(\overline{s}_j, \overline{a}_{j-1}) = I\{C_j(\overline{s}_j, \overline{a}_{j-1}) > 0\},$$

where $I(\cdot)$ is the indicator function. Note that any function of $(\overline{s}_j, \overline{a}_j)$, $Q_j(\overline{s}_j, \overline{a}_j)$, say, can be written $h_j(\overline{s}_j, \overline{a}_{j-1}) + a_j C_j(\overline{s}_j, \overline{a}_{j-1})$, for arbitrary functions $h_j(\overline{s}_j, \overline{a}_{j-1})$ and $C_j(\overline{s}_j, \overline{a}_{j-1})$, where $h_j(\overline{s}_j, \overline{a}_{j-1}) = Q_j(\overline{s}_j, \overline{a}_{j-1}, 0)$. This suggests that a more robust procedure may be to develop parametric models for the Q-contrast function C_j and estimate the parameters from the observed data. This is the premise of A-learning.

To illustrate how A-learning works, consider the first step in Q-learning, which involves a model for $E(Y|\overline{S}, \overline{A})$ in terms of parameters ψ_T. Instead, consider the semiparametric model for $E(Y|\overline{S}, \overline{A})$ given by

$$h_T(\overline{S}_T, \overline{A}_{T-1}) + A_T C_T(\overline{S}_T, \overline{A}_{T-1}; \xi_T),$$

where $h_T(\overline{S}_T, \overline{A}_{T-1})$ is an arbitrary function of $\overline{S}_T$ and $\overline{A}_{T-1}$, and only the Q-contrast function is modeled using a finite number of parameters ξ_T. If, in addition, we know or can estimate the propensity score

$$E(A_T|\overline{S}_T, \overline{A}_{T-1}) = P(A_T = 1|\overline{S}_T, \overline{A}_{T-1}), \tag{5.9}$$

then semiparametric theory shows that all consistent, asymptotically normal estimators for ξ_T must be solutions to estimating equations of the form

$$\sum_{i=1}^{n} \phi_T(\overline{S}_{Ti}, \overline{A}_{(T-1)i})\{A_{Ti} - E(A_T|\overline{S}_{Ti}, \overline{A}_{(T-1)i})\}\{Y_i - A_{Ti} C_T(\overline{S}_{Ti}, \overline{A}_{(T-1)i}; \xi_T) - \theta_T(\overline{S}_{Ti}, \overline{A}_{(T-1)i})\} = 0 \tag{5.10}$$

for arbitrary $\phi_T(\overline{S}_T, \overline{A}_{T-1})$ and $\theta_T(\overline{S}_T, \overline{A}_{T-1})$. These sets of equations are what Robins (1994) refers to as g-estimation used in estimating "blip-functions" in structural nested mean models. The optimal estimating equation uses $\phi_T(\overline{S}_T, \overline{A}_{T-1}) = \partial/\partial\xi_T\{C_T(\overline{S}_T, \overline{A}_{T-1}; \xi_T)\}\{\text{var}(Y|\overline{S}_T, \overline{A}_{T-1})\}^{-1}$ and $\theta_T(\overline{S}_T, \overline{A}_{T-1}) = h_T(\overline{S}_T, \overline{A}_{T-1})$. As $h_T(\overline{S}_T, \overline{A}_{T-1})$ and $\text{var}(Y|\overline{S}_T, \overline{A}_{T-1})$ are unknown, one could model these in an attempt to achieve optimal efficiency; even if the chosen models are incorrect, the resulting estimator for ξ_T would still be consistent and asymptotically normal for ξ_T, assuming that the model for $C_T(\overline{S}_{Ti}, \overline{A}_{(T-1)i})$ is correctly specified. Implementation requires that propensity score (5.9) is known, as in a sequentially randomized trial, or it must be modeled and estimated if the analysis is based observational data.

Given the estimated Q-contrast function $C_T(\overline{S}_T, \overline{A}_{T-1}; \widehat{\xi}_T)$, say, then an obvious estimator for the value function $\widetilde{V}_T = \max_{a_T} Q_T(\overline{S}_T, \overline{A}_{T-1}, a_T)$ is $Y + C_T(\overline{S}_T, \overline{A}_{T-1}; \widehat{\xi}_T)[I\{C_T(\overline{S}_T, \overline{A}_{T-1}; \widehat{\xi}_T) > 0\} - A_T]$. The term $C_T(\overline{S}_T, \overline{A}_{T-1}; \widehat{\xi}_T)[I\{C_T(\overline{S}_T, \overline{A}_{T-1}; \widehat{\xi}_T) > 0\} - A_T]$ is called the advantage function, as it represents the advantage incurred if the patient receives the best treatment at the Tth decision point as compared to the current treatment received.

The A-learning algorithm thus is as follows: Specify a series of Q-contrast function models $C_j(\overline{S}_j, \overline{A}_{j-1}, \xi_j)$, $j = T, T-1, \ldots, 0$, and, in an iterative fashion, estimate the ξ_j, $j = T, T-1, \ldots, 0$, by solving

$$\sum_{i=1}^{n} \frac{\partial C_j(\overline{S}_{ji}, \overline{A}_{(j-1)i}; \xi_j)}{\partial \xi_j} \Sigma_j^{-1}(\overline{S}_{ji}, \overline{A}_{(j-1)i})\{A_{ji} - E(A_j|\overline{S}_{ji}, \overline{A}_{(j-1)i})\}$$
$$\times\{\widetilde{V}_{(j+1)i} - A_{ji}C_j(\overline{S}_{ji}, \overline{A}_{(j-1)i}; \xi_j) - h_j(\overline{S}_{ji}, \overline{A}_{(j-1)i})\} = 0,$$

where we adopt the form of the optimal equation; $\Sigma_j(\overline{S}_j, \overline{A}_{j-1})$ is a posited model for $\mathrm{var}(\widetilde{V}_{j+1}|\overline{S}_j, \overline{A}_j)$ that in practice would be fitted and substituted, as would be a model for $h_j(\overline{S}_{j,i}, \overline{A}_{j-1,i})$; the value function $\widetilde{V}_j$ is obtained as $\widetilde{V}_j = \widetilde{V}_{j+1} + C_j(\overline{S}_j, \overline{A}_{j-1}; \widehat{\xi}_j)[I\{C_j(\overline{S}_j, \overline{A}_{j-1}; \widehat{\xi}_j) > 0\} - A_j]$; $\pi_j(\overline{s}_j, \overline{a}_{j-1}) = I\{C_j(\overline{s}_j, \overline{a}_{j-1}; \widehat{\xi}_j) > 0\}$; and $E(\widetilde{V}_0) = E\{Y^*(g^{\mathrm{opt}})\}$.

Comparison of Q- and A-Learning. To our knowledge, except for some of our preliminary studies (see Section **4**), little is known about the performance of either Q- or A-learning in scenarios likely to be encountered in cancer research. Moreover, the relative merits of the techniques in general settings are largely unknown. Clearly, A-learning is more robust than Q-learning, as it makes fewer assumptions on the Q-function; i.e., in Q-learning, the Q-functions must all be correctly specified, while in A-learning only the Q-contrast functions need be. However, this robustness comes at a price of greater variability. Thus, there is a bias-variance trade-off in assessing the performance of these two methods, but insights are difficult to achieve theoretically. Thus, our first objective during the project period will be to carry out studies of these issues through some analytical work and extensive simulation experiments to critically examine the performance of Q- and A-learning in scenarios designed to resemble cancer studies. This will include evaluating relative performance under different sample sizes in order to provide guidelines on the sample size requirements necessary in order to expect reliable inferences on optimal dynamic treatment regimes with these methods.

We will begin by considering the simplest situation of a single-decision study, i.e., $T = 0$. Here, it is possible to compute the relative efficiency of the two methods when the Q-function is correctly specified. For example, assuming

$$E(Y|S_0, A_0) = Q(S_0, A_0, \psi) = h(S_0, \chi) + A_0 C(S_0, \xi)$$

is correctly specified, where $\psi = (\chi, \xi)$, the parameter ξ may be consistently estimated using both Q- and A-learning. Thus, we will be able to assess the relative efficiency of these two methods analytically. In the situation that $h(\cdot)$ is misspecified, Q-learning will no longer result in consistent inferences, while A-learning will; accordingly, we will evaluate the bias that is incurred in estimating ξ using Q-learning under these conditions. We will also derive the asymptotic properties of the Q- and A-learning estimators for $E\{Y^*(g^{\mathrm{opt}})\}$ and use these to assess relative efficiency, using the delta method assuming the parameters are not on the boundary; we will also results of Aim 3 when they are. This exercise, although in a simple case, will give us some basic insight on the relative performance of these two methods and help guide our design of more complex and realistic simulation experiments involving multiple decision points, under which useful analytical evaluations of performance are intractable. We will design these studies based on, for example, studies such as the sequentially randomized trials CALGB 9710 and 19808 discussed in Section **3.5**, data sets compiled by Core B, data from trials in NSCLC mentioned in Section **5.5**, and in consultation with Dr. Socinski. We will also consider how to extend Q- and A-learning to the case of censored survival outcomes.

As noted previously, a major challenge in this context is developing the regression models for the Q- and Q-contrast functions, and, because of the potential that these modeling exercises will involve high-dimensional information, some sort of model selection techniques must be employed to reduce the dimension and identify the relevant information to be included in the models. There are several challenges involved. First, most model selection techniques, such as the LASSO, adaptive LASSO, SCAD, and FSR methods noted in Section **3.2** that focus on minimizing prediction error may not necessarily be the most advantageous approaches for choosing

models that lead to the best treatments given previous information at each decision point. Here, it is critical to identify covariate information that exhibits qualitative interactions with treatment. Moreover, when considering bias-variance trade-offs, it is important to recognize that these propagate as models are recursively built and fitted. Our second objective during the project period will be to carry out a comprehensive study, which will of necessity be empirical, of the performance of a variety of model selection strategies for Q-learning and A-learning in settings reflecting those in cancer research, to provide recommendations on how these critical components of learning methods be carried out in the analysis of data from cancer clinical trials. Again, we will use data such as those described above to inform the design of extensive simulation studies comparing the performance of Q- and A-learning when coupled with various forms of model selection over a broad range of conditions.

We will consider a number of methods in these studies. Traditional selection techniques available in software such as the SAS regression procedures will be studied as "off-the-shelf" methods familiar to practitioners, although we expect these to be grossly inadequate. We will also study the LASSO, adaptive LASSO, and SCAD, which are based on penalizing the model fitting criterion, where the penalty is tuned adaptively. These methods are familiar to biostatisticians and available in R packages and hence, for brevity, we do not review them here. (A brief description of these methods is given in Section 5.1 of Project 2.) We will also study some new or less familiar approaches, which we now describe.

Fast False Selection Rate Method. The recently developed FSR method of Wu et al. (2007) and Boos et al. (2008) is an approach to tuning any existing model selection method so that the proportion of "unimportant" covariates that enter the selected model is on average equal to a pre-specified small value such as 0.05. When used with traditional forward selection to tune the "α to enter," FSR is a type of adaptive False Discovery Rate method (Benjamini and Hochberg, 2000) with intuitive appeal, as it leads to parsimonious models with good prediction error performance. The original version of FSR proposed by Wu et al. (2007) involves adding simulated noise variables to the real predictors in the model, monitoring when they enter a forward selection sequence, and using this information to tune the procedure, and hence can be time consuming to compute. Boos et al. (2008) develop a "fast" approximation to FSR that does not require simulation and is easy to implement using existing software. As we noted above, ability to identify qualitative interactions between treatment and covariates is critical in the context of learning methods; see Section **5.4** for more on this topic. In recent work, Crews et al. (2008) have extended the fast FSR methods to allow separate tuning considerations for main effects and interactions. In learning, it may be beneficial to increase the probability that important qualitative interactions are included in the model; thus, fast FSR, appropriately adapted to the setting of learning methods, has potential to offer a fruitful approach. We will be able to adapt the software implementation of the fast FSR approach developed in Project 2 for this purpose.

Support Vector Regression. This method and the next have their origins in the machine learning literature. Support Vector Regression (SVR) is an extension of Support Vector Machines (SVM) developed by Vapnik (1995). The SVM paradigm was originally designed for the classification problem and provides a compromise between parametric and nonparametric approaches. SVMs are often involved in the solution of learning the relationship between the attributes x and label indices y in a training data set $\{(x_i, y_i) \in X \times \mathcal{Y}\}_{i=1}^n$. In Q-learning, X may be replaced by $\{S, A\}$ representing covariate-treatment information, and we define attributes $x_{ij} \in \overline{S}_j \times \overline{A}_j$, i=1,...,n; in the general case, $\mathcal{Y}$ may be replaced by numerical rewards. In many cases the reward function maps $(\overline{s}_j, \overline{a}_j, s_{j+1})$ to a set that consists of some discrete integer numbers, and, if the size of the set is larger than 2, is a multicategory classification problem (e.g., Crammer and Singer, 2001, 2003; Lee, 2004). However, when the number of the classes is large (more than 4) or in the extreme case where the reward is continuous, and the numerical value is not only a label index but also a continuous outcome, multicategory learning methods may be inadequate. Therefore, support vector regression (SVR), one of the most popular extensions of SVM (Vapnik et al., 1997), is more suitable.

The ideas underlying SVR are similar but slightly different from SVM. The data are mapped into a feature space by a nonlinear transformation Ψ, which guarantees that any data set becomes arbitrarily separable as the data dimension grows (Cover, 1965). A hyperplane is then fitted to the mapped data. The SVR function is also derived within the reproducing kernel Hilbert space context, but one of the popular loss functions involved in SVR is known as the ϵ-insensitive loss function, defined as $L(f(x_i), y_i) = \{|f(x_i) - y_i| - \epsilon\}_+$, $\epsilon > 0$ (Vapnik, 1995). Other possible loss functions include quadratic loss, Laplace loss, and Huber loss. Given training data

$\{(x_i, y_i) \in X \times Y\}_{i=1}^n$, SVR solves the optimization problem $\min_{w,b,\xi,\xi'} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n(\xi_i + \xi_i')$ subject to

$$(w^T\Psi(x_i) + b) - y_i \leq \epsilon + \xi_i, \quad y_i - (w'\Phi(x_i) + b) \leq \epsilon + \xi_i', \quad \xi_i, \xi_i' \geq 0, \quad i = 1, \ldots, n; \tag{5.11}$$

the tuning parameter C is determined by cross-validation. By minimizing the regularization term $\frac{1}{2}\|w\|^2$ as well as the training error $C\sum_{i=1}^n(\xi_i + \xi_i')$, SVR can avoid both overfitting and underfitting. The slack variables ξ_i and ξ_i' allow for some data points in the feature space to stay outside the confidence band determined by ϵ. In other words, the goal is to find a function that has at most ϵ deviation from the actually obtained targets y_i for all the training data. Errors with deviation larger than ϵ are not accepted. In practice, this is changed to a standard (convex and quadratic) optimization problem using Lagrange multipliers.

Similar to SVM, which calculates a hyperplane, the solution of an SVR function only depends on the support vectors (Cortes and Vapnik, 1995). Usually, support vectors just represent a small fraction of the sample, therefore, the evaluation of the decision function is computationally efficient. This attractive property is especially useful when dealing with data sets with a low ratio of sample size to dimension. To achieve good performance by using SVR, some procedures such as data scaling, kernel and related parameter selection need to be implemented carefully. Compared to least-squares regression where ϵ is always zero, SVR is a more general and flexible approach for regression problems. In very high-dimensional cases, it may be necessary to carry out a dimension reduction, e.g., via sparse principal components techniques, prior to invoking SVR.

Extremely Randomized Trees. Ernst et al. (2005) and Geurts et al. (2006) propose the Extremely Randomized Trees (ERT) method, called the Extra-Trees algorithm, for batch mode reinforcement learning. Unlike classical classification and regression trees such as Kd-tree or pruned CART trees, this nonparametric method builds a model in the form of the average prediction of an ensemble of regression trees, or random forest. Each tree consists of strongly randomizing both attribute and cut-point choice while splitting a tree node. In addition to the number of trees G, this method depends on one parameter, K, the maximum number of cut-direction tests at each node, and $k_{\min}$, the minimum number of elements at each leaf depends on the resulting compromise between computational requirements and prediction accuracy. K determines the strength of the randomization. For $K = 1$, the splits are chosen independently of the output variable. A larger $k_{\min}$ yields smaller trees but higher bias. The ERT algorithm builds G trees using the training data set. To determine a test at a node for each tree, this algorithm randomly selects K attributes with K randomized cut-points. A score is calculated for each test and then the one which has the highest value is kept. The algorithm stops splitting a node when the number of elements in the node is less than $k_{\min}$. The complete ERT algorithm is given in Geurts et al. (2006).

Compared to standard tree-based regression methods, ERT successfully leads to significant improvements in precision, can dramatically decrease variance while at the same time decreasing bias, and is very robust to outliers. ERT has been recently been implemented in a simulation of HIV infection (Ernst et al., 2006) and adaptive treatment of epilepsy (Guez et al., 2008). While this algorithm appears very effective for extracting a well-fitted Q-function from the data set, its drawback is low computational efficiency, especially with increasing sample size of the training data set. We used both ERT and SVR in the work leading to Figure 3 and found the results to be very similar, with ERT was notably more time consuming than SVR. However, the tremendous flexibility the ERT offers may offset the computational cost.

During the project period, we will carry out several steps in our investigation of model selection in the context of reinforcement learning in cancer research. We will begin with the simple case of a single decision point, for which there is only one model selection exercise to conduct. In this situation, this single model selection determines the optimal regime, so this will allow us to evaluate the relative merits of the model selection strategies to address this goal directly. We will then consider situations with more than one decision point; here the model selection methods must both provide good predictions for use in the learning algorithms and ultimately identify the best regime. In both single- and multiple-decision point settings, we will devise simple generative scenarios involving solely discrete S and A, where the optimal regime may be determined (e.g., Chakraborty et al., 2008), and carry out extensive simulation studies using the various selection methods as above. We will also consider more complex, realistic generative scenarios with high-dimensional S where the data are generated according to a sequence of probability distributions as in Robins' (1994) g-computation algorithm. We will then carry out the learning methods incorporating each model selection technique. For the optimal regime estimated by each, $\widehat{g}$, say, we will use the g-computation algorithm based on the true generative model with the decision rules specified by $\widehat{g}$ to determine the mean outcome $E\{Y^*(g)\}$ evaluated at the particular $\widehat{g}$. In simulations, we

will then be able to generate the distribution of such mean outcomes associated with the use of each model selection method, which may be directly compared. We will consider a range of such scenarios based on, e.g., the CALGB studies mentioned previously, and synthesize the results into recommendations for practice.

We close this section by noting that our proposed research, which represents a critical step toward adoption of the study of dynamic treatment regimes in cancer clinical trials and which will involve examining numerous scenarios and intensive, large-scale computations, would be almost impossible to carry out without access to the resources and varied and powerful expertise of the overall Program Project. Not only will the computational resources of Core C be critical to facilitating our efforts, we will be able to exploit progress achieved in Project 4 on using model selection and classification methods to develop single-decision individualized therapies through the joint involvement of Drs. Bondell, Kosorok, Zeng, and Zhang in those efforts and the insights gained on model selection methods in other contexts from Project 2 through the joint involvement of Drs. Bondell, Boos, Davidian, Stefanski, Tsiatis, and Zhang.

5.3 Aim 2: Identifying Optimal Dynamic Treatment Regimes From a Restricted, Feasible Set

To fix ideas, consider a simple example. Suppose interest focuses on a particular induction therapy, a particular maintenance therapy that would be given if a patient responds to induction therapy, and a particular salvage therapy that would be given if s/he does not, all of which may be given in different doses x_I, x_M, and x_S, say, respectively. It should be clear that triplets (x_I, x_M, x_S), where x_I, x_M, and x_S take on different values, correspond to different possible dynamic treatment regimes we can summarize by $g = (x_I, x_M, x_S)$; i.e., regimes of the form "Give induction therapy at dose x_I; if the patient responds, give maintenance therapy at dose x_M; if s/he does not, give salvage therapy at dose x_S." A natural goal is to estimate the mean outcomes that would be achieved if all patients in the population followed each regime g, i.e., estimate $E\{Y^*(g)\} = \mu(x_I, x_M, x_S)$ for each regime g and find the optimal regime of this form that would yield the maximum mean outcome if all patients followed it. Supposing that the range of possible doses is a continuum, we cannot estimate $\mu(x_I, x_M, x_S)$ for any specific value of (x_I, x_M, x_S) without further assumptions; a natural approach is to consider a model for $\mu(x_I, x_M, x_S)$ depending on a finite-dimensional parameter τ. For example, we may consider a quadratic model

$$\mu(x_I, x_M, x_S; \tau) = \tau_0 + \tau_1 x_I + \tau_2 x_M + \tau_3 x_S + \tau_4 x_I^2 + \tau_5 x_M^2 + \tau_6 x_S^2 + \tau_7 x_I x_M + \tau_8 x_I x_S + \tau_9 x_M x_S. \quad (5.12)$$

If this model were an accurate representation of the true mean outcome-dose relationship, then the optimal doses could be found by taking the gradient of (5.12) with respect to x_I, x_M, x_S and setting it equal to zero, yielding

$$\begin{pmatrix} x_I^{\text{opt}} \\ x_M^{\text{opt}} \\ x_S^{\text{opt}} \end{pmatrix} = -\begin{pmatrix} 2\tau_4 & \tau_7 & \tau_8 \\ \tau_7 & 2\tau_5 & \tau_9 \\ \tau_8 & \tau_9 & 2\tau_6 \end{pmatrix}^{-1} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}.$$

Suppose data from a sequentially randomized trial randomizing subjects to several fixed doses (or to a continuum of doses from a continuous distribution) of each agent or an observational database, where the doses given to patients were at clinician discretion, are available of the form $(S_{0,i}, A_{0,i}.S_{1,i}, A_{1,i}, Y_i)$, $i = 1, \ldots, n$. Here, S_0 denotes baseline covariates, $A_0 = X_I$, the dose of induction therapy given; S_1 are additional covariates collected in the intervening period between induction and the time that response is declared to have occurred or not (S_1 includes a response indicator R defined as before); $A_1 = RX_M + (1 - R)X_S$; i.e., the dose of maintenance (salvage) therapy given if response (non-response) was observed; and Y is the outcome. The question of interest is then, given the model (5.12), how to estimate τ from these observed data.

A general approach to estimation of parameters in models with such a restricted set of feasible treatment regimes is required. The method that we propose is based on solving inverse propensity score weighted estimating equations. As motivation, we first review how inverse propensity score weighted estimators are used to obtain an estimator for $E\{Y^*(g)\}$ for a single dynamic treatment regime g. Define the propensity score for a treatment regime g to be

$$P\{A_0 = g_0(S_0), A_1 = g_0(\overline{S}_1), \ldots, A_T = g_T(\overline{S}_T)|W\}, \quad (5.13)$$

where W denotes the set of potential outcomes defined in (5.3). By the law of conditional probabilities, this can be expressed equivalently as $P\{A_0 = g_0(S_0)|W\} \times \ldots \times P\{A_T = g_T(\overline{S}_T)|\overline{A}_{T-1} = \overline{g}_{T-1}(\overline{S}_{T-1}), W\}$. Because of the assumptions of SUTVA (5.4) and no unmeasured confounders (5.6), the jth term in this expression may be written as

$$P\{A_j = g_j(\overline{S}_j)|\overline{A}_{j-1} = \overline{g}_{j-1}(\overline{S}_{j-1}), \overline{S}_j\}. \quad (5.14)$$

If we denote $P(A_j = a_j|\overline{A}_{j-1} = \overline{a}_{j-1}, \overline{S}_j = \overline{s}_j)$ by $f_j(a_j|\overline{a}_{j-1}, \overline{s}_j)$, and assume $f_j(a_j|\overline{a}_{j-1}, \overline{s}_j) > 0$, then the propensity score defined by (5.13) is equal to

$$\prod_{j=0}^{T} f_j\{g_j(\overline{S}_j)|\overline{g}_{j-1}(\overline{S}_{j-1}), \overline{S}_j\}.$$

It may be shown by using an iterated conditional expectation argument that

$$E\left[\frac{\prod_{j=0}^{T} I\{A_j = g_j(\overline{S}_j)\}Y}{\prod_{j=0}^{T} f_j\{g_j(\overline{S}_j)|\overline{g}_{j-1}(\overline{S}_{j-1}), \overline{S}_j\}}\right] = E\{Y^*(g)\}$$

by first conditioning on W and using the SUTVA assumption and (5.14) above.

To describe the proposed methods, we introduce some shorthand notation. Let $D = (\overline{S}, \overline{A})$ denote the observed covariate-treatment history, and let $\eta_g(d)$ be the indicator of whether or not the covariate-treatment combination $d = (\overline{s}, \overline{a})$ is consistent with the treatment regime g (see Section **3.5**); that is, $\eta_g(d) = 1$ if $(\overline{s}, \overline{a}) = \{\overline{s}, \overline{g}(\overline{s})\}$, and 0 otherwise. Letting

$$\zeta(d) = \prod_{j=0}^{T} f_j(a_j|\overline{s}_j, \overline{a}_{j-1}),$$

we can obtain an unbiased estimator for $E\{Y^*(g)\} = \mu(g)$ as $n^{-1}\sum_{i=1}^{n}\{\eta_g(D_i)Y_i\}/\zeta(D_i)$, which may be written equivalently as the solution to the estimating equation

$$\sum_{i=1}^{n} \frac{\eta_g(D_i)\{Y_i - \mu(g)\}}{\zeta(D_i)} = 0. \tag{5.15}$$

As noted in Section **3.5**, for complicated dynamic treatment regimes, the number of individuals whose experience is consistent with the treatment regime g, $\{i : A_{ji} = g_j(\overline{S}_{ji})$ for all $j = 0, \ldots, T\}$, may be so small as to give unstable and unreliable estimates. Of course, for sequentially randomized trials, the probabilities $f_j(a_j|\overline{s}_j, \overline{a}_{j-1})$, $j = 0, \ldots, T$, are known by design, whereas they would need to be modeled and estimated in an analysis based on observational data.

We now return to the our problem. Suppose that we limit the scope of our problem to a feasible subset of dynamic treatment regimes $\mathcal{F} \subset \mathcal{G}$. In the above example, $\mathcal{F}$ consists of the combinations of doses (x_I, x_M, x_S) that would be given for the induction drug, the maintenance drug, and the salvage drug, respectively, where whether or not a maintenance or salvage dose is given depends on the value of R. Thus, in this example, $\mathcal{F}$ restricts attention to regimes where decisions are made only on the basis of R. As with the quadratic model (5.12) in the example, we posit a model for

$$E\{Y^*(g)\} = \mu(g, \tau), \text{ for all } g \in \mathcal{F},$$

as a function of a finite number of parameters τ. The goal is then to estimate τ from the observed data (5.2).

Exploiting the idea of the inverse propensity score weighted estimator solving (5.15), we propose estimating τ by solving the estimating equation

$$\sum_{i=1}^{n} \int_{\mathcal{F}} \frac{\eta_g(D_i)\partial/\partial\tau\{\mu(g,\tau)\}\{Y_i - \mu(g,\tau)\}}{\zeta(D_i)} d\nu(g) = 0, \tag{5.16}$$

where $d\nu(g)$ is some mass function that weights across the different regimes in $\mathcal{F}$. The important thing to note is that a single value of the treatment-covariate history D may be consistent with more than one of the treatment regimes $g \in \mathcal{F}$. That the resulting estimator will be consistent and asymptotically normal will follow because (5.16) is an unbiased estimating equation, a result that can easily be shown because of (5.15). The choice of the weight function $d\nu(g)$ will affect the efficiency of the estimator, and careful study of how to choose $d\nu(g)$ will be required. Once the estimator for τ is obtained, the estimated optimal regime in the restricted set and its mean are determined, and its asymptotic properties may then be deduced via standard methods.

To illustrate, consider the model (5.12). Suppose we have conducted a sequentially randomized trial where patients are randomized with equal probability to doses $x_{Ij},\ j = 1,\ldots,k_1$, for induction therapy, to doses $x_{Mj},\ j = 1,\ldots,k_2$, for maintenance therapy for patients who respond, and to doses $x_{Sj},\ j = 1,\ldots,k_3$, for salvage therapy for patients who do not respond. Considering the treatment regime $g = (x_{Ij_1}, x_{Mj_2}, x_{Sj_3})$, then $\eta_g(D_i) = 1$ if $X_{Ii} = x_{Ij_1}, X_{Mi} = x_{Mj_2}, R_i = 1$ or $X_{Ii} = x_{Ij_1}, X_{Si} = x_{Sj_3}, R_i = 0$ and is equal to 0 otherwise. If we put point mass at the $k_1 \times k_2 \times k_3$ treatment combinations, then

$$\begin{aligned}\zeta(D_i) &= 1/(k_1k_2) \text{ when } R_1 = 1\\ &= 1/(k_1k_3) \text{ when } R_1 = 0.\end{aligned}$$

Consequently, equation (5.16) becomes

$$\sum_{i=1}^{n}\left[k_1k_2R_i\sum_{j=1}^{k_3}\frac{\partial\mu(X_{Ii},X_{Mi},x_{Sj};\tau)}{\partial\tau}\{Y_i-\mu(X_{Ii},X_{Mi},x_{Sj};\tau)\}\right.$$
$$\left.+\ k_1k_3(1-R_i)\sum_{j=1}^{k_2}\frac{\partial\mu(X_{Ii},x_{Mj},X_{Si};\tau)}{\partial\tau}\{Y_i-\mu(X_{Ii},x_{Mj},X_{Si};\tau)\}\right]=0.$$

During the project period, we will carry out a comprehensive study of this approach. We will first derive the asymptotic properties of such estimators. With observational data, there is the additional complication of the need to model and estimate the propensity score. Although the feasible dynamic treatment regimes in $\mathcal{F}$ may be deliberately restricted to involve only some of the covariate data $\overline{S}$ in the decision rules, such data may play an important role in modeling the propensity score with observational data in order that the assumption of no unmeasured confounders be tenable. Therefore, we will derive the asymptotic properties of the proposed estimator when the propensity score is estimated. We will also study the implications of the choice of the weight function $d\nu(g)$ in (5.16) and misspecification of $\mu(g,\tau)$. We will carry out extensive simulation experiments across a range of scenarios designed based on the cancer clinical trials noted previously and data compiled by Core B to evaluate practical performance. An important goal will be to establish guidelines on the sample sizes required to ensure that valid inferences on the dynamic treatment regimes derived from the methods.

As with many inverse probability weighted methods, efficiency gains can be made by using the covariates $\overline{S}$ in augmented inverse propensity score weighted estimators with methods similar to those described in Aims 1 and 2 of Project 2. We will explore this approach during this project period. Because time to event endpoints are often the primary outcome in cancer trials, and because such endpoints are often right censored we will study how the above methods can be adapted for use with censored survival outcomes.

5.4 Aim 3: Inferential Methods for Dynamic Treatment Regimes

Because of its relative simplicity, we focus on Q-learning. We will use empirical process techniques combined with post-model-selection (also called "two-stage") estimation to develop valid methods of inference for parameters in Q-functions used in reinforcement learning methods. We will use a simple, two decision-point Q-learning example to illustrate the proposed approach. The setting is taken from Chakraborty et al. (2008) who use it to describe the soft-thresholding approach to inference mentioned in Section **3.4**.

Suppose we have observed iid data that are realizations of (5.1) for $T = 1$, $(S_{0i}, A_{0i}, R_{0i}, S_{1i}, A_{1i}, R_{1i}, S_{2i})$, $i = 1,\ldots,n$, where R_0, R_1 are the rewards resulting at each decision point. Suppose our goal is to maximize the expected, undiscounted return, where here the return is the total reward $\sum_{j=0}^{1} R_j$. We let the Q-functions for $j = 0, 1$ be represented as

$$Q_j(H_j, A_j; \beta_j, \gamma_j) = \beta_j' H_{j0} + (\gamma_j' H_{j1})A_j, \tag{5.17}$$

where here $\psi + j$ in Section **5.2** is partitioned as $\psi_j = (\beta_j', \gamma_j')'$; we write the history at the jth decision point as $H_j = (\overline{S}_j, \overline{A}_{j-1})$ for brevity; H_{j0} and H_{j1} are subsets of H_j selected for the model at decision point j; $A_j \in \{-1, 1\}$ is a dichotomous treatment choice (we use the set $\{-1,1\}$ instead of $\{0,1\}$ for mathematical convenience); and the parameters of the Q-function at the jth decision point are β_j and γ_j, $j = 0, 1$. The parameters β_j reflect the associate of patient history to outcome, while γ_j reflects the interaction between

patient history and treatment choice. Accordingly, following the discussion at the end of Section **5.2**, the γ_j are of particular interest.

Here, the Q-learning procedure in Section **5.2** involves the following steps (also given in Chakraborty et al. 2008): (i) $(\widehat{\beta}_1, \widehat{\gamma}_1) = \arg\min_{\beta_1,\gamma_1} n^{-1} \sum_{i=1}^n \{R_{1i} - Q_1(H_{1i}, A_{1i}; \beta_1, \gamma_1)\}^2$; (ii) For $i = 1, \ldots, n$, estimate $\widetilde{V}_{0i} = R_{0i} + \max_{a\in\{-1,1\}} Q_1(H_{1i}, a; \widehat{\beta}_1, \widehat{\gamma}_1)$; and (iii) $(\widehat{\beta}_0, \widehat{\psi}_0) = \arg\min_{\beta_0,\gamma_0} n^{-1} \sum_{i=1}^n \{\widetilde{V}_{0i} - Q_0(H_{0i}, A_{0i}; \beta_0, \gamma_0)\}^2$. The form of Q_j in (5.17) yields trivially that $(\widehat{\beta}_1', \widehat{\gamma}_1')' = \mathcal{Z}_1^{-1} \sum_{i=1}^n Z_{1i} R_{1i}$, where $Z_{ji} = (H_{j0}', A_j H_{j1}')'$, $i = 1, \ldots, n$, and $\mathcal{Z}_1 = n^{-1} \sum_{i=1}^n Z_{1i} Z_{1i}'$; that

$$\widetilde{V}_{0i} = R_{0i} + \widehat{\beta}_1' H_{10i} + |\widehat{\gamma}_1' H_{11i}|; \tag{5.18}$$

and that $(\widehat{\beta}_0', \widehat{\gamma}_0')' = \mathcal{Z}_0^{-1} n^{-1} \sum_{i=1}^n Z_{0i} \widetilde{V}_{0i}$, where $\mathcal{Z}_0 = n^{-1} \sum_{i=1}^n Z_{0i} Z_{0i}'$. The difficulty, of course, is the fact that the absolute value function in (5.18) is non-differentiable.

We now sketch a new two-stage approach to finding the limiting distribution and conducting inference without requiring soft-thresholding. We assume that $Y_{0i} = Q_0(H_{0i}, A_{0i}; \beta_{00}, \gamma_{00}) + \epsilon_{0i}$; that $Y_{1i} = Q_1(H_{1i}, A_{1i}; \beta_{10}, \gamma_{10}) + \epsilon_{1i}$; and that the error terms have conditionally mean zero and finite joint covariance given Z_{0i} and Z_{1i}. We also assume that all covariates are bounded and that

$$\begin{pmatrix} I_{00} & I_{01} \\ I_{01}' & I_{11} \end{pmatrix} = E\left\{ \begin{pmatrix} Z_{0i} \\ Z_{1i} \end{pmatrix} \begin{pmatrix} Z_{0i} \\ Z_{1i} \end{pmatrix}' \right\}$$

is positive definite. We also assume that the covariance matrix

$$\begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{pmatrix} = E\left\{ \begin{pmatrix} \epsilon_{0i}^2 & \epsilon_{0i}\epsilon_{1i} \\ \epsilon_{0i}\epsilon_{1i} & \epsilon_{1i}^2 \end{pmatrix} \middle| Z_{0i}, Z_{1i} \right\}$$

is likewise positive definite. These are essentially the same assumptions made in Chakraborty et al. (2008). We further assume that (D) $P(|\gamma_{10}' H_{11i}| = 0) > 0$ holds. Without (D), the discontinuity in (5.18) has no effect asymptotically, and the standard delta-method can yield asymptotic normality of all parameters of interest. Assumption (D) is realistic in practice for both continuous and dichotomous covariates.

Define the function $F_0 : \mathbb{R}^d \mapsto \mathbb{R}^p$, where d is the dimension of γ_1 and p is the dimension of Z_{0i}, as

$$F_0(u) = E\left[Z_{0i} \left(|\gamma_1' H_{11i}| I\{|u' H_{11i}| = 0\} + (\gamma_1' H_{11i}) I\{|u' H_{11i}| > 0\} \right) \right].$$

Let $(G_0, G_1)'$ be a mean-zero, Gaussian random vector with covariance matrix

$$\begin{pmatrix} \sigma_{00} I_{00} & \sigma_{01} I_{01} \\ \sigma_{01} I_{01}' & \sigma_{11} I_{11} \end{pmatrix}.$$

Using careful empirical process techniques (e.g., Kosorok, 2008), we can show under our assumptions that

$$n^{1/2} \begin{pmatrix} \widehat{\beta}_0 - \beta_{00} \\ \widehat{\gamma}_0 - \gamma_{00} \\ \widehat{\beta}_1 - \beta_{10} \\ \widehat{\gamma}_1 - \gamma_{10} \end{pmatrix} \rightsquigarrow \begin{pmatrix} I_{00}^{-1} \left[G_0 + E\{Z_{1i} H_{10i}'\} I_{11.0}^{-1} G_1 + F_0(I_{11.1}^{-1} G_1) \right] \\ I_{11}^{-1} G_1 2 \end{pmatrix}, \tag{5.19}$$

where $\rightsquigarrow$ denotes convergence in distribution; $I_{11.1}^{-1}$ is the lower submatrix of I_{11}^{-1} consisting of the bottom d rows; and $I_{11.0}^{-1}$ is the upper submatrix consisting of the remaining rows of I_{11}^{-1}. This is clearly a complex limiting distribution involving non-differentiable functions. If (D) does not hold, then $F_0(u) = E\{Z_{0i} H_{11i}'\} u$ is a simple linear function, and the limiting distribution on the right side of (5.19) becomes a mean-zero Gaussian process with easily estimated covariance matrix. In general, F_0 does not simplify this nicely, and inference is challenging.

The two-stage approach we propose for inference involves creating an empirical rule for testing whether or not $\gamma_{10}' H_{11i} = 0$ for each observation i and applying a specialized parametric bootstrap. Let r_n be a decreasing sequence of constants such that $r_n \to 0$ and $n^{1/2} r_n \to \infty$, and generate a realization of $(G_0, G_1)'$. For $i = 1, \ldots, n$, let $\delta_{in} = \left(|G_{11}' H_{11i}| > r_n \sqrt{H_{11i}' \widehat{U} H_{11i}} \right)$, where $\widehat{U}$ is an empirical estimator of $\sigma_{11} I_{11.1}^{-1} I_{11} (I_{11.1}^{-1})'$. Now

replace $F_0(u)$ in the right side of (5.19) with $n^{-1}\sum_{i=1}^{n} Z_{0i}\{\delta_{in}H'_{11i}u + (1-\delta_{in})|H'_{11i}u|\}$, and replace I_{00}, $I_{11.0}^{-1}$, $I_{11.1}^{-1}$, I_{11}^{-1} and $E(Z_{0i}H'_{10i})$ with their empirical estimators. It can be shown after careful empirical process analysis that the distribution of this quantity, conditional on the sample data, is asymptotically equivalent to the limiting distribution given on the right side of (5.19) and thus can be used for valid inference without requiring soft-thresholding. This approach is valid even when (D) does not hold. The reason this works in spite of the inherent challenges with two-stage estimation, which in this case happens through the δ_{in} values, is that the region in the space where δ_{in} transitions from 0 to 1 can be made to have a very small probability, because

$$\lim_{\eta \downarrow 0} P\left(0 < \frac{|\gamma'_{10}H_{11i}|}{\|H_{11i}\|} \leq \eta\right) = 0,$$

and that the "size" of the test reflected in the rate r_n goes to zero while remaining asymptotically consistent.

While this is clearly a complicated inference procedure and algorithm, it should be quite fast computationally and practical for implementation. During the project period, we will verify this procedure theoretically and evaluate its performance and operating characteristics in extensive simulation studies under a variety of scenarios designed to reflect the conditions likely to be encountered in cancer research. Two key issues that we will address are how best to select r_n in a possibly data-adaptive way and how to extend this general approach to other, more complicated reinforcement learning settings, for example, to A-learning and with more complicated set-ups (numbers of decision points and treatment options at each). To obtain these results, we will use empirical process methods. Our initial restriction to the two decision-point case is not a serious limitation; as we discuss in Section **5.5**, two decision points may be a feasible setting in the design of sequentially randomized cancer trials, such as the clinical reinforcement trial for NSCLC we propose in Aim 4. While inference on the parameters is of interest, the ultimate focus is on precision of estimation of the expected total reward (mean outcome) achieved by the estimated optimal regime. Because this is a differentiable functional of the parameters and probability distribution of the data, it is possible using empirical process methods to exploit the above results to obtain the limiting distribution of this quantity and a valid bootstrap procedure.

For the development of sample size guidelines for these methods, the distribution given in (5.19) is too complex to yield closed-form sample size formulæ, and it will be necessary to make some simplifying assumptions and approximations. In particular, we can consider replacing $F_0(u)$ with a simple linear function that yields a conservative estimator for the variance of the entire distribution on the right side of (5.19). This would result in a Gaussian random variable for which the derivation of sample size formulæ is relatively straightforward. While inferential methods need to be precise, sample size formulæ, on the other hand, maybe somewhat conservative without losing their practical usefulness.

5.5 Aim 4: Design of Sequentially Randomized Trials for Dynamic Treatment Regimes

We discuss the two sub-aims for this aim in turn.

Design and Sample Size considerations for Sequentially Randomized Trials. As we noted in Section **3.5**, sample size determination is a key challenge in the design of sequentially randomized trials, owing in part to the fact that, as the numbers of decision points and treatment options at each increase, the numbers of subjects with experience consistent with each regime decreases. Thus, there is a need for methods to allow evaluation of the precision that can be expected as a function of the numbers of decision points where randomization will occur and the options at each. We propose a framework for systematic determination of the total sample size for these trials on the basis of criteria relevant to the comparison of the regimes embedded in such a trial.

To fix ideas, we consider the setting in Section **5.1** where there is a primary outcome Y of interest to be ascertained on each subject. The goal is to design a sequentially randomized trial to identify which among a set of K dynamic treatment regimes, where K is a function of the number of decision points and treatment options at each, would yield the greatest mean outcome (assuming larger values of Y are preferred). Let μ_k, $k = 1, \ldots, K$, denote the expectation of the potential outcome corresponding to the kth regime; that is, the mean outcome that would be seen if all patients in the population followed regime k. Appealing to ideas advocated in the classical literature on ranking and selection (Bechhofer, 1954; Bechhofer, Kiefer, and Sobel, 1968) and also discussed by Oetting, Levy, and Murphy (2007), a sensible criterion for evaluation of dynamic treatment regimes is that the trial allow determination of the best dynamic treatment regime; i.e., $\arg\max(\mu_1, \ldots, \mu_K)$, such that, if indeed that treatment regime is better than the remaining $K-1$ by some tolerable limit δ, say, this will be detected with sufficiently high probability. Without loss of generality, assuming that treatment regime 1 is

the best and is better than all the rest by at least δ; i.e., $\{\mu_1 - \max_{j=2,\ldots,K}(\mu_j)\} \geq \delta$, then we want a design that would guarantee that

$$P\{\widehat{\mu}_1 \geq \max_{j=2,\ldots,K}(\widehat{\mu}_j)\} \geq 1-\beta \tag{5.20}$$

whenever $\{\mu_1 - \max_{j=2,\ldots,K}(\mu_j)\} \geq \delta$, where $\widehat{\mu}_j$ is an estimator for the mean response of dynamic treatment regime j and β is a small probability chosen by the investigator.

A variety of estimators for μ_j the may be used with data from sequentially randomized trials have been proposed (e.g., Lunceford et al., 2002; Murphy, 2005; Wahed and Tsiatis, 2004, 2006), which generally have the property that $(\widehat{\mu}_1,\ldots,\widehat{\mu}_K)'$ is asymptotically normal with mean $(\mu_1,\ldots,\mu_K)'$ and $(K\times K)$ covariance matrix Λ that can be estimated from the data. If we can deduce the form of Λ for different designs and sample sizes, then we can evaluate probabilities in (5.20) and use them to assess the relative feasibility of the design.

For illustration, consider the sequentially randomized design depicted in Figure 2, which embeds the $K=8$ simple dynamic treatment regimes we denoted as (C_i, M_j, S_k) for $i,j,k = 1,2$, discussed in Sections **3.1**, **3.5**, and **5.1**. We may summarize all of the possible realized treatment experiences that could be had by patients following any of these regimes as in Table 1, where "R" and "NR" denote "response observed" or "no response observed," so that, for example, experience 1 corresponds to being assigned induction chemotherapy C_1, responding to it, and then being assigned to maintenance therapy M_1. It is critical to recognize that the experiences in the table are *not* themselves dynamic treatment regimes but rather are results of *following* dynamic treatment regimes.

1.	C_1	R	M_1
2.	C_1	R	M_2
3.	C_1	NR	S_1
4.	C_1	NR	S_2
5.	C_2	R	M_1
6.	C_2	R	M_2
7.	C_2	NR	S_1
8.	C_2	NR	S_2

Table 1: Realized experiences from the sequentially randomized trial.

In the actual trial, there will be n_ℓ subjects whose realized experience under sequential randomization corresponds to experience ℓ in the table, which we denote by the event $E=\ell$. Let $\overline{Y}_\ell$ be the sample average outcome for these subjects. If we define $\theta_\ell = E(Y|E=\ell)$ and $p_s = P(\text{response}|C_s)$, $s=1,2$, then it is easy to show that the mean outcomes for the $K=8$ dynamic treatment regimes embedded in the trial may be expressed as

$$\begin{array}{llll}\mu_1=\theta_1p_1+\theta_3(1-p_1), & \mu_2=\theta_1p_1+\theta_4(1-p_1), & \mu_3=\theta_2p_1+\theta_3(1-p_1), & \mu_4=\theta_2p_1+\theta_4(1-p_1),\\ \mu_5=\theta_5p_2+\theta_7(1-p_2), & \mu_6=\theta_5p_2+\theta_8(1-p_2), & \mu_7=\theta_6p_2+\theta_7(1-p_2), & \mu_8=\theta_6p_2+\theta_8(1-p_2),\end{array}$$

and these means may be estimated by substituting $\overline{Y}_\ell$ for θ_ℓ, $(n_1+n_2)/(n_1+n_2+n_3+n_4)$ for p_1, and $(n_5+n_6)/(n_5+n_5+n_7+n_8)$ for p_2 in these formulæ. Denote the resulting estimators by $\widehat{\mu}_j$, $j=1,\ldots,8$.

Under some simplifying assumptions; namely, assuming constant variance across ℓ, i.e.,

$$\sigma^2 = \text{var}(Y|E=\ell),\ \ell=1,\ldots,8, \quad \text{and} \quad \sigma^2 >> |\theta_\ell - \theta_{\ell'}| \text{ for all } \ell\neq\ell', \tag{5.21}$$

which may be reasonable approximations for design purposes, it is straightforward to show that $(\widehat{\mu}_1,\ldots,\widehat{\mu}_8)$ will be asymptotically normal with mean $(\mu_1,\ldots,\mu_8)'$ and covariance matrix

$$\Lambda = \frac{4\sigma^2}{n}\left(\begin{array}{cccc|cccc} 1 & p_1 & 1-p_1 & 0 & 0 & 0 & 0 & 0\\ p_1 & 1 & 0 & 1-p_1 & 0 & 0 & 0 & 0\\ 1-p_1 & 0 & 1 & p_1 & 0 & 0 & 0 & 0\\ 0 & 1-p_1 & p_1 & 1 & 0 & 0 & 0 & 0\\ \hline 0 & 0 & 0 & 0 & 1 & p_2 & 1-p_2 & 0\\ 0 & 0 & 0 & 0 & p_2 & 1 & 0 & 1-p_2\\ 0 & 0 & 0 & 0 & 1-p_2 & 0 & 1 & p_2\\ 0 & 0 & 0 & 0 & 0 & 1-p_2 & p_2 & 1 \end{array}\right).$$

For simplicity, assume that $p_1 = p_2$.

Because of the symmetry of Λ, to compute the lower bound for

$$P\{\widehat{\mu}_1 \geq \max_{j=2,\ldots,8}(\widehat{\mu}_j)\} \tag{5.22}$$

whenever $\{\mu_1 - \max_{j=2,\ldots,8}(\mu_j)\} \geq \delta$, for a given sample size n and probability of response p_1, it suffices to compute (5.22) assuming that $(\widehat{\mu}_1, \ldots, \widehat{\mu}_8)$ is multivariate normal with mean $(\delta, 0, \ldots, 0)'$ and covariance matrix Λ above. This probability can be computed easily by simulation. Note that this configuration is least favorable among all such configurations with $\{\mu_1 - \max_{j=2,\ldots,8}(\mu_j)\} \geq \delta$ in the sense of yielding the largest sample size and hence is conservative.

To illustrate, we present results from such simulations. Specifically, in Table 2, we show the sample size (up to a multiplicative constant σ^2/δ^2) for different values for the probability in equation (5.22) and for different p_1.

Hence, if we design a sequentially randomized trial as in Figure 2 and expect a response rate of $p_1 = 0.7$, then to ensure that we identify the best dynamic treatment regime among those represented in the trial with probability 0.90 if that treatment regime has a mean outcome at least δ units greater than the rest, then, according to Table 2, we need a sample size of at least $27.7 \times (\sigma^2/\delta^2)$.

	Probability		
p_1	0.80	0.90	0.95
0.9	16.4	27.0	37.1
0.8	16.8	27.4	37.5
0.7	17.1	27.7	37.8
0.6	17.4	27.9	38.1
0.5	17.6	28.1	38.4

Table 2: Sample size up to σ^2/δ^2.

Note that the sample size is relatively insensitive to the choice of the probability of response p_1, suggesting that the simplifying assumption that $p_1 = p_2$ is a reasonable approximation. More intriguing is the comparison to standard sample size calculations. If instead we consider a simple, two-arm randomized trial and use the standard sample size calculation, choosing sample size to ensure that a level-α test (two-sided) would detect a treatment difference in mean outcome of δ units with power at least $1 - \beta$, then we would take $n = 4(z_{\alpha/2} + z_\beta)^2(\sigma^2/\delta^2)$, where z_β is the $(1-\beta)$th quantile of a standard normal distribution. For the routine choices $\alpha = 0.05$ and power equal to $1 - \beta = 0.90$, the sample size would equal $4(1.96 + 1.28)^2(\sigma^2/\delta^2) = 42.0 \times (\sigma^2/\delta^2)$. Thus, we see that, with the same sample size used for a standard, two-sample comparison, we could have designed a sequentially randomized trial that evaluates $K = 8$ dynamic treatment regimes and has a greater than 95% chance of identifying the best regime if that regime has a mean outcome at least δ units greater than the rest.

Note that this suggests that one could design a sequentially randomized trial having a specified probability of detecting the best regime that would also have high power for carrying out a standard primary analysis comparing the two induction treatments to which subjects are randomized at the first decision point. Ordinarily, in standard trials with a single decision-point randomization, what happens to subjects after they receive their assigned treatments is left to the discretion of clinicians and subjects, and the analysis proceeds according to the intention-to-treat principle. From the perspective of such a primary analysis, one may view sequentially randomized trials as single decision-point trials where aspects of what happens to the subjects after randomization are "systematized." This feature does *not* compromise the validity of an intention-to-treat analysis of the induction treatments; indeed, this analysis is routine in sequentially randomized cancer trials. The advantage is that designing the trial to have sufficient power for this analysis *and* sufficient precision to identify the best regime among those represented with high probability places us in a position to learn much more with a similar commitment of resources. If extensive patient information is also collected throughout the trial, the resulting data will furthermore be a rich resource for identifying the optimal regime using learning methods.

The calculations above are for a very specific design and involve some simplifying assumptions. During the project period, we will study the impact of the assumptions in (5.21), especially the second, on the sample size calculations, and we will generalize the sample size methods to apply to sequentially randomized studies with multiple decision points, multiple treatment options, and more complex decision rules. The expressions and computations involved will of necessity be much more complicated than those presented here. By so doing, we will be able to provide comprehensive guidance on the level of complexity of the regimes that may be incorporated in a trial while preserving feasible statistical accuracy. Our ultimate objective is to develop software to assist cancer trialists in the design of sequentially randomized trials; see Section **5.6**.

Clinical Reinforcement Trials for Cancer Treatment. The foregoing sub-aim focuses on design where the objective is to evaluate the specific regimes embedded in the trial. As we have emphasized, a key objective of this project is to advance the premise of carrying out sequentially randomized trials with the primary objective of obtaining a rich data resource for developing optimal dynamic treatment regimes using reinforcement learning methods. As proof of principle of this exciting idea of clinical reinforcement trials, in this sub-aim, we will carry out a detailed study of this approach in the context of Stage IIIB/IV NSCLC. Notably, the implications of the first

sub-aim suggest that these trials can also facilitate standard analyses.

Most of our research will be through simulation of virtual patients and virtual clinical reinforcement trials in order to assess sample size and performance where the goal is development of optimal regimes. Our first step will be to develop a realistic mechanistic difference-equation framework for modeling the progression of NSCLC and how it is affected by treatment from which virtual patients may be simulated. We have already had success with using a simple difference-equation approach in the preliminary work leading to Figure 3. In the proposed efforts, we will develop a more comprehensive, realistic such framework by using existing clinical trial data on NSCLC, such as LCCC protocols 9719, "A phase III, randomized trial comparing a defined duration versus continuous administration of combination chemotherapy in advance non-small cell carcinoma of the lung," and 2003, "A phase II, randomized trial comparing weekly administration of taxol with carboplatin to an every-three-week regimen for the treatment of advanced stage IIIB/IV non-small cell lung cancer," in order to better tune the models to match observed patient behavior.

The three aspects of clinical reinforcement trials, numbers of decision points, numbers of treatment options, and outcome measure (utility function), will be specialized to NSCLC. We will have at most two decision times, one possibly at the beginning of the first line of treatment, and the second at the end of first-line treatment. A third line is currently only available for certain patients, and there is only one FDA approved third-line treatment (so only one viable treatment option at a third decision point); accordingly, we will not incorporate a third decision point. We will strongly emphasize the second line of treatment given at the second decision point; as noted in Section **3.5**, timing of initiation of second-line therapy appears to be the most pressing issue clinically. We will include a finite set of possible treatments (agents) for the first decision time (beginning of first line therapy), which is restricted somewhat by established clinical practice. For the second decision time, we will have a finite set of agents and a possible continuum of start times for second line of therapy, where, again, we will incorporate constraints on treatment choices as dictated by clinical practice. For example, we will use a limited choice of targeted start times that fall within the usual 6 week median window before relapse and initiation of second-line therapy at the earlier of the targeted start time and relapse. We will use overall survival as our primary outcome (utility function), as this outcome is arguably the most crucial. We will also compare and contrast overall survival with a utility function that assesses quality adjusted survival.

For the analysis component, we will use Q-learning methodology combined with the model selection methods shown to work best in Aim 1 to estimate the optimal regime as a function of patient variables and biomarkers at each decision time. We have already demonstrated in our preliminary work in a simple setting, as shown in Figure 3, that Q-learning can lead to optimal regimes that lead to improved patient outcomes over fixed treatment, and we expect to demonstrate similar gains in this more realistic setting.

Exploiting our simulation framework, we will study systematically the sample sizes required to achieve precise inferences on the optimal dynamic treatment regime from such trials and evaluate the performance of the inferential methods in Aim 3 in this particular context. Our preliminary such simulation studies seem to suggest that sample sizes required to make reliable inferences on dynamic treatment regimes are similar to those required for typical phase III trials, similar to the findings for evaluating a fixed set of regimes in the previous sub-aim. However, considerable, systematic study is needed to develop reliable sample size guidelines.

Based on the results of these simulation-based studies and in close collaboration with Dr. Socinski, we will design a Stage IIIB/IV NSCLC clinical reinforcement trial. This will require identifying and refining all of the needed aspects described above while ensuring that the resulting regimes to which subjects would be randomized are consistent with clinical practice. We expect that arriving at a final design that can be realistically implemented will take several iterations before a suitable and efficient design is achieved.

This general approach is possible with other cancers, such as breast and colon cancers. We will undertake a study of other cancer treatment questions and develop simulation-based frameworks that can be used to develop clinical reinforcement trials. This study will enable us to use the differences among settings for various cancers to formulate general guidelines for developing clinical reinforcement trials in cancer research.

5.6 Software Implementation and Dissemination

All public-use software developed in this project will be made available on dedicated pages on the Program Project website, including downloadable code and instructions, documentation, and examples.

Aim 1: As a first step toward our comprehensive study of learning methods, we will develop efficient, robust programs for implementation of Q- and A-learning, where the models for the Q-functions and Q-contrast func-

tions, respectively, at each decision point are prespecified. We will make heavy use of Core C resources for developing and testing these programs. These programs will facilitate the simulations and data analyses required to carry out our first objective of understanding the relative merits of the two learning approaches. For model selection methods for which reliable, suitable implementations are not readily available, such as the Fast FSR, SVR, and ERT methods, we will develop efficient implementations, exploiting the resources of Core C. These programs will be integrated with the Q- and A-learning routines to facilitate our large-scale study of model selection methods in this context. Ultimately, based on the results, we will work with Core C to embed versions of these programs implementing the best methods into software, documentation, and examples for public dissemination. Because of the complexity involved, we envision that this software will require some intervention by the user to tailor it for a specific use; thus, the source code will be transparent and heavily documented.

Aim 2: We will work with Core C to develop software implementations of the methods for public dissemination.

Aim 3: We will work with Core C to incorporate the inferential methods for optimal dynamic treatment regimes in the public-use Q-learning software developed in Aim 1.

Aim 4: A critical objective of this aim is to develop guidelines and software for practical design of sequentially randomized trials. We plan to work closely with Core C programmers to develop a general public-use, software package suitable for use by analysts involved in protocol development based on the research in sub-aim 1; this will included numerous illustrative examples. For that in sub-aim 2, working with Core C, we will develop simulation-based software for public dissemination that will enable practitioners to carry out virtual clinical reinforcement trials and sample size determinations for such trials. The software will be written so as to be transparent to researchers who wish to modify it for their own needs.

5.7 Timetable

For all Aims, manuscripts will be prepared and submitted as results worthy of publication become available.

Aim 1: The first 2 years will be devoted to development of the Q- and A-learning programs and their use in our initial investigation comparing the two methods. In year 3, we will expand these studies to assessing the role of model selection and develop programs implementing these methods if none are available, which will be used in the continuing effort in Year 4 to carry out a comprehensive study of their performance. Year 5 will be devoted to developing the programs into software for public use along with accompanying documentation.

Aim 2: The first year will be spent fully developing the methods and deriving their theoretical properties. Year 2 will be devoted to developing efficient programs in collaboration with Core C and on research, including review of studies compiled by Core B, to devise realistic scenarios where a restricted set of feasible regimes would be of interest to inform our extensive simulation studies. Work will begin on translating the programs into public use software, continuing in Year 3. Year 3 will also be devoted to extending the methods to incorporate "augmentation terms" to gain efficiency, as discussed in Aim 1 of Project 2, and to censored survival outcomes and efficient implementation of these methods. In Year 4, theoretical study and extensive simulations will be carried out. Year 5 will be devoted to completion of public-use software implementations of all methods.

Aim 3: The first two years will be devoted to theoretical work and, in collaboration with Core C, development of efficient implementations of the inferential methods that may be integrated with the Q-learning programs in Aim 1. Year 3 will be devoted to extensive simulations of the methods' performance. Years 4 and 5 will involve development of sample size guidelines and incorporation of the methods in the public-use software in Aim 1.

Aim 4: For sub-aim 1, in Year 1, we will carry out the necessary analytical work and develop efficient programs that allow for varying assumptions, numbers of decision points, and numbers of treatment options for computation of sample size. Years 2 and 3 will be devoted to analytical and empirical study of sample size of general trial designs using these programs. In Year 4, we will use the results to develop guidelines for practitioners and worked examples be included in the documentation for the public-use software to be developed in Year 5. For sub-aim 2, we will spend Years 1 and 2 developing the NSCLC simulation framework, working closely with Core C programmers, and using it to generate virtual trial data that we will use with Q-learning to develop clinical reinforcement trial designs. We will use the evolving results to develop the draft NSCLC trial protocol by the end of Year 3. In Years 3 and 4, we will begin to study other cancers and develop simulation frameworks that can be used to develop clinical reinforcement trials in these settings. Year 5 will be devoted to development of the public-use simulation-based software.

6 INCLUSION ENROLLMENT REPORT

N/A

7 BIBLIOGRAPHY AND REFERENCES CITED

Almirall, D., Gunter, L. L., and Murphy, S. A. (2005). Efficient A-learning for dynamic treatment regimes: A handout. Available at— `http://www.stat.lsa.umich.edu/ samurphy/nida/seminars.html`.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* **10**, 1100–1120.

Bancroft, T. A. and Han, C. P. (1977). Inference based on conditional specification: a note and bibliography. *International Statistical Review* **45**, 117–127.

Bartlett, P. L. and Tewari, A. (2007). Sample complexity of policy search with known dynamics. In: *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, BC, December 4–7, 2006*. MIT Press, Cambridge, MA.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* **25**, 16-3-9.

Bechhofer, R. E., Kiefer, J., and Sobel, M. (1968). *Sequential Identification and Ranking Procedures*. Chicago: The Univ. of Chicago Press.

Benjamani, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**, 60–83.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics* **36**, 199–227.

Blatt, D., Murphy, S. A., and Zhu, J. (2004). A-learning for approximate planning. *Unpublished Manuscript*.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and clustering of predictors with OSCAR. *Biometrics* **64**, 115–123.

Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, in press.

Bondell, H. D. and Li, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *Journal of the Royal Statistical Society, Series B*, in press.

Boos, D. D., Stefanski, L. A., and Wu, Y. (2008). Fast FSR variable selection with applications to clinical trials. *Biometrics*, `doi:10.1111/j.1541-0420.2008.01127.x`. NIHMSID:NIHMS76165.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Boca Raton, Florida: Chapman and Hall/CRC Press.

Chakraborty, B., Strecher, V., and Murphy, S. A. (2008). Bias correction and confidence intervals for fitted Q-iteration. Unpublished paper presented at the NIPS-08 Workshop on Model Uncertainty and Risk in Reinforcement Learning, Whistler, BC, December 13, 2008.

Ciuleanu, T. E., Brodowicz, T., Belani, C. P., Kim, J., Krzakowski, M., Laack, E., Wu, Y., Peterson, P., Adachi, S., and Zielinski, C. C. (2008). Maintenance pemetrexed plus best supportive care (BSC) versus placebo plus BSC: A phase III study. *ASCO*, May 20 suppl, abstract 8011.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20**, 273–297.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers* **14**, 326–334.

Crammer, K. and Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* **2**, 265–292.

Crammer, K. and Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research* **3**, 951–991.

Crews, H. B., Boos, D. D., and Stefanski, L. A. (2008). Fast false selection rate methods for analyzing response surface experimental designs. Institute of Statistics Mimeo Series # 2619, Raleigh.

El-Sadr, W. M., Lundgren, J. D., Neaton, J. D., Gordin, F., Abrams, D., Babiker, A., Clumeck, N., Cohen, C. J., Cohn, D., Cooper, D., Emery, S., Fatkenheuer, G., Gazzard, B., Grund, B., Hoy, J., Klingman, K., Losso, M., Ramos, J. M., Markowitz, N., Neuhaus, J., Phillips, A., and Rappoport, C. (2006). CD4+ count-guided interruption of antiretroviral treatment: The Strategies for Management of Antiretroviral Therapy (SMART) Study Group. *New England Journal of Medicine*, **355**, 2283–2296.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch model reinforcement learning. *Journal of Machine Learning Research* **6**, 503–556.

Ernst, D., Stan, G.-B., Gonçalves, J., and Wehenkel, L. (2006). Clinical data based optimal STI strategies for HIV: a reinforcement learning approach. *Proceedings of the 45th IEEE Conference on Decision and Control.*

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle property. *Journal of the American Statistical Association* **96**, 1348–1360.

Fidias, P., Dakhil, S., Lyss, A., Loesch, D., Waterhouse, D., Cunneen, J., Chen, R., Treat, J., Obasaju, C., and Schiller, J. (2007). Phase III study of immediate versus delayed docetaxel after induction therapy with gemcitabine plus carboplatin in advanced non-small-cell lung cancer: Updated report with survival. *ASCO* **25**, June 20 suppl, LBA7516.

Food and Drug Administration (2004). Innovations or stagnation: Challenge and opportunity on the critical path to new medical products. White paper.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning* **11**, 3–42.

Grossi, F., Gridelli, C., Aita, M. and De Marinis, F. (2008). Identifying an optimum treatment strategy for patients with advanced non-small cell lung cancer. *Critical Reviews in Onocology/Hematology* **67**, 16–26.

Hanna, N., Shepherd, F. A., Fossella, F.V., Pereira, J. R., de Marinis, F., von Pawel, J., Gatzemeier, U., Chang, T., Tsao, Y., Pless, M., Muller, T., Lim, H. L., Desch, C., Szondy, K., Gervais, R., Shaharyar, Manegold, C., Paul, S., Paoletti, P., Einhorn, L., and Bunn, P. A. (2004). Randomized phase III trial of pemetrexed versus docetaxel in patients with non-small-cell lung cancer previously treated with chemotherapy. *Journal of Clinical Oncology* **22**, 1589–1597.

Högberg, T. (2005). Widening bottlenecks in drug discovery: Glimpses from Drug Discovery Technology Europe. *Drug Discovery Today* **10**, 820–822.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation* **6**, 1185–1201.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 272–280.

Johnson, B. A. and Tsiatis, A. A. (2004). Estimating mean response as a function of treatment duration in an observational study, where duration may be informatively censored. *Biometrics* **60**, 315–323.

Johnson, B. A. and Tsiatis, A. A. (2005). Semiparametric inference in observational duration-response studies, with duration possibly right-censored. *Biometrika* **92**, 605–618.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* New York: Springer.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: a survey. *Journal of Artificial Intelligence Research* **4**, 237–285.

Laber, E. B. and Murphy, S. A. (2008). Small sample inference for generalization error in classification using the CUD bound. *Proceedings of the 2008 Uncertainty in Artificial Intelligence Conference.* AUAI Press, 357–365.

Lavori, P. and Dawson, R. (2004). Dynamic treatment regimes: practical design considerations. *Clinical Trials* **1**, 9–20.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* **99**, 67–81.

Leeb, H., and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* **34**, 2554–2591.

Lieberman, J., Stroup. T. S., McEvoy, J., Swartz, M., Rosenheck, R., Perkins, D. O., Keefe, R. S. E., Davis, S. M., Davis, C. E., Lebowitz, B. D., Severe, J., Hsiao, J. K. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine*, **353**, 1209–1223.

Lin, Y. and Zhang, H. H. (2006). Component Selection and smoothing in smoothing spline analysis of variance models - COSSO. *Annals of Statistics* **34**, 2272–2297.

Liu, Y., Zhang, H. H., Park, C., and Zhu, J. (2008). Support vector machines with adaptive Lq penalty. *Computational Statistics and Data Analysis* **51**, 6380–6394.

Lunceford, J., Davidian, M., and Tsiatis, A. A. (2002) Estimation of the survival distribution of treatment regimes in two-stage randomization designs in clinical trials. *Biometrics* **58**, 48–57.

Luo, X., Stefanski, L. A., and Boos, D. D. (2006). Tuning variable selection procedures. *Technometrics* **48**, 168–175.

Moodie, E. M., Richardson, T. S., Stephens, S. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63**, 447–455.

Murphy, S. A. (2003). Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society, Series B* **58**, 331–366.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* **24**, 1455–1481.

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R., and Ten Have, T. (2007a). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence* **88S**, S24–S30.

Murphy, S. A., Oslin, D. W., Rush, A. J., and Zhu, J. for MCATS (2007b). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology* **32**, 257–262.

Murphy, S. A., van der Laan, M. J., Robins, J. M. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* **96**, 1410–1423.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agricales: Essay de principle. English translation of excerpts by Dabrowska, D. and Speed, T. (1990). *Statistical Science* **5**, 465–480.

Ng, A. Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., Berger, E., and Liang, E.. (2006). Inverted autonomous helicopter flight via reinforcement learning. In *Experimental Robotics IX: The 9th International Symposium on Experimental Robotics*, Eds. H.A. Marcelo Jr., and O. Khatib. New York: Springer, pp. 363–371.

Oetting, A. I., Levy, J. A., and Murphy, S. A. (2007). Evaluation of sample size formulæ for developing adaptive treatment strategies using a SMART design. To appear in *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures* Ed., P.E. Shrout, Arlington VA: American Psychiatric Publishing, Inc. Available at `http://www.stat.lsa.umich.edu/~samurphy/papers/appatechrep_04302007.pdf`.

Pineau, J., Bellmare, M. G., Ruch, A. J., Ghizaru, A., and Murphy, S. A. (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence* **88S**, S52–S60.

Pirker, R., Szczesna. A., von Pawel, J., Krzakowski, M., Ramlau, R., Park, K., Gatzemeier, U., Bajeta, E., Emig, M., and Pereira, J. R. (2008) FLEX: A randomized, multicenter, phase III study of cetuximab in combination with cisplatin/vinorelbine (CV) versus CV alone in the first-line treatment of patients with advanced non-small cell lung cancer (NSCLC). *Journal of Clinical Oncology*, May 20 suppl, abstract 3.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods: Applications to control of the healthy worker survivor effect. *Mathematical Modeling* **7**, 1393–1512.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics – Theory and Methods* **23**, 2379–2412.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, Eds. D.Y. Lin, P.J. Heagerty. New York: Springer, pp. 189–326.

Rosenberg, E. S., Davidian, M., and Banks, H. T. (2007). Using mathematical modeling and control to develop structured treatment interruption strategies for HIV infection. *Drug and Alcohol Dependence* **88S**, S41–S51.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **6**, 34–58.

Rummery, G. A. and Niranjan, M. (1994). On-line Q-learning Using Connectionist Systems. Technical Report, CUED/F-INFENG/TR166, Cambridge University.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nirenberg, A. A., Quitkin, F. M., Kashner, T. M., Kupler, D. J., Rosenbaum, J. F., Alpert, J., Stewart, J., McGrath, P. J., Biggs, M. M., Shores-Wilson, K., Lebowitz, B. D., Ritz, L. and Niederehe, G. for the STAR*D Investigators Group. (2004). Sequenced Treatment Alternatives to Relieve Depression (STAR*D): Rationale and design. *Controlled Clinical Trials* **25**, 119–142.

Sandler, A., Gray, R., Perry, M. C., Brahmer, J., Schiller, J. H., Dowlati, A., Lilenbaum, R., Johnson, and D. H. (2006). Paclitaxel-Carboplatin alone or with bevacizumab for non-small-cell lung cancer. *New England Journal of Medicine* 355, 2542–2550.

Scagliotti, G.V., Parikh, P., von Pawel, J., Biesma, B., Vansteenkiste, J., Manegold, C., Serwatowski, P., Gatzemeier, U., Digumarti, R., Zukin, M., Lee, J. S., Mellemgaard, A., Park, K., Patil, S., Rolski, J., Goksel, T., de Marinis, F., Simms, L., Sugarman, K. P., and Gandara, D. (2008). Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *Journal of Clinical Oncology* **26**, 3543–3551.

Schäfer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**, Article 32.

Sen, P. K. (1986). On the asymptotic distributional risks of shrinkage and preliminary test versions of maximum likelihood estimators. *Sankhya, Series A* **48**, 354–371.

Shepherd, F. A., Dancey, J., Ramlau, R., Mattson, K., Gralla, R., O'Rourke, M., Levitan, N., Gressot, L., Vincent, M., Burkes, R., Coughlin, S., Kim, Y., and Berille, J. (2000). Prospective randomized trial of docetaxel versus best supportive care in patients with non-small-cell lung cancer previously treated with platinum-based chemotherapy. *Journal of Clinical Oncology* **18**, 2095–2103.

Smola, A., and Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing* **14**, 199–222.

Socinski, M. A., Novello, S., Brahmer, J. R., Rosell, R., Sanchez, J. M., Belani, C. P., Govindan, R., Atkins, J. N., Gillenwater, H. H., Palleres, C., Tye, L., Selanu, P., Chao, R. C., and Scagliotti, G. V. (2008). Efficacy and safety of Sunitinib in previously treated, advanced non-small cell lung cancer. *Journal of Clinical Oncology* **26**, 650–656.

Stinchcombe, T. E., and Socinski, M. A. (2008). Considerations for second-line therapy of non-small cell lung cancer. *The Oncologist* **13**(suppl 1), 28–36.

Storlie, C. B., Bondell, H. D., Reich, B. J., and Zhang, H. H. (2008). Surface estimation, variable selection, and the nonparametric oracle property. Under review.

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning* **3**, 9–44.

Sutton, R.S., Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: MIT Press.

Thall, P. F., Millikan, R. E., and Sung, H. G. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine* **19**, 1011–1028.

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. E., and Tannir, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine* **26**, 4687–4702.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning* **16**, 185–202.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27**, 4658–4677. NIHMSID:NIHMS46485.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik, V., Golowich, S., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems* **9**, 281–287.

Wahed A. S. and Tsiatis A. A. (2004). Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics* **60**, 124–133.

Wahed A. S. and Tsiatis A. A. (2006). Semiparametric efficient estimation of survival distribution for two-stage randomization designs with censored data. *Biometrika* **93**, 163–178.

Wang, H. and Leng, C. (2007). Unified LASSO estimation via least squares approximation. *Journal of the American Statistical Association* **102**, 1039–1048.

Watkins, C. J. C. H. (1989). Learning from Delayed Rewards. Ph.D. Thesis. King's College, Cambridge, U.K.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning* **8**, 279–292.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association* **102**, 235–243.

Zhang, H. H. and Lu, W. (2007). Adaptive LASSO for Cox's proportional hazards model. *Biometrika* **93**, 1–13.

Zhang, H. H., Wahba, G., Lin, Y., Volker, M., Ferris, M., Klein, R., and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association* **99**, 659–672.

Zhang, H. H. (2006). Variable selection for support vector machines via smoothing spline ANOVA. *Statistica Sinica* **16**, 659–674.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inference in randomized clinical trials using auxiliary covariates. *Biometrics* **64**, 707–715. PMCID: PMC2574960.

Zhang, M., Tsiatis, A. A., Davidian, M, Pieper, K., and Mahaffey, K. (2009) Inference on treatment effects from a randomized clinical trial in the presence of premature treatment discontinuation: The SYNERGY trial. Unpublished manuscript; submitted to *Biometrics*.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2008). Reinforcement learning design for cancer clinical trials. Submitted.

Zou, H. (2006) The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429

Zou, H., and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, in press.

8 PROTECTION OF HUMAN SUBJECTS

Although the proposed research indirectly involves human subjects through the preparation, in Core B, of de-identified data sets from identifiable patient data sources, the investigators on Project 5 will have access only to the de-identified data. Thus, the investigators on Project 5 will have no access to any identifiable patient information.

9 INCLUSION OF WOMEN AND MINORITIES

The methods we develop will be applicable to studies with both women and minorities and also to studies which examine treatment differences adjusted for gender, ethnicity and race. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, many of the existing data sets to be studied and provided by Core B include women and minorities, although we will not be generating any new data involving human subjects.

10 TARGETED/PLANNED ENROLLMENT TABLE

N/A

11 INCLUSION OF CHILDREN

The methods we develop will be applicable to studies with children and also to studies which examine treatment differences adjusted for age. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, some of the existing data sets to be studied and provided by Core B may include children, although we will not be generating any new data involving human subjects.

12 VERTEBRATE ANIMALS

N/A

13 SELECT AGENT RESEARCH

N/A

14 MULTIPLE PD/PI LEADERSHIP PLAN

N/A

15 CONSORTIUM/CONTRACTUAL ARRANGEMENTS

If the present application is funded, the University of North Carolina at Chapel Hill will execute a subcontract with the consortium institution (North Carolina State University). The inter-institutional agreement will be written consistent with the NIH consortium agreement policy.

16 LETTERS OF SUPPORT - None

17 RESOURCE SHARING PLAN(S)

(a) Data sharing plan: The data-related resources generated by the proposed research consists of new statistical methodology, software packages for implementation of the methodology, and tutorials for the software. The statistical methodology will be shared through peer reviewed publications and national meetings and through other standard means. All accepted publications will be deposited in PubMed Central in accordance with the NIH Public Access Policy. Summaries of the methodology, the software and tutorials will be shared through a public web site managed by Core A, while Core C will assist in preparation of the software and tutorials for dissemination. This project will use de-identified data prepared by Core B to test the methods and to create demonstrations of use of the methods to be included in tutorials. This project will not be involved in sharing of these data; this function will be addressed by Core B.

(b) Sharing model organisms: N/A

(c) GWAS: N/A

CORE A

ADMINISTRATIVE CORE

Core Director: Michael R. Kosorok, PhD

PROJECT SUMMARY (See instructions):

The Administrative Core (Core A) is responsible for organizing the program investigators and staff into an effective and well-coordinated team to develop and implement the statistical methods for cancer clinical trials proposed in the research projects to improve the health of cancer patients. This program is integrated across three institutions whith a lead PD/PI at one institution (UNC-CH) and two additional PD/PIs at the other two institutions (NCSU and Duke). These three PD/PIs form an executive Committee with overall responsibility for the management and administration of the program. Each institution has an additional co-PD/PI to assist the PD/PIs with both the overall and intra-institutional administration of the program project. The Executive Committee, three co-PD/PIs, and individual project leaders form a Steering Committee which provides overall scientific guidance for the program. An External Advisory Committee of experts provides feedback to the Steering Committee on the goals and progress of the program during an annual retreat. Communication and collaboration between project investigators is facilitated with a program project wiki. Communication and dissemination of new results and software are aided with a program project web page. The matrix leadership structure of Core A maximizes the scientific integration of this multi-disciplinary and trans-institutional collaboration.

RELEVANCE (See instructions):

The Administrative Core (Core A) is essential to the success of the proposed project since it coordinates all administration and provides leadership for the five projects, three cores and three institutions involved in this program project. The administrative component is necessary to facilitate the science of this program project and to achieve the overall program aims, to develop new statistical methods that will improve the health of cancer patients.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location			
Organizational Name: The University of North Carolina at Chapel Hill			
DUNS: 608195277			
Street 1: Office of Sponsored Research, CB #1350		Street 2: 104 Airport Dr., Suite 2200	
City: Chapel Hill	County: Orange		State: NC
Province:	Country: USA		Zip/Postal Code: 27599-1350
Project/Performance Site Congressional Districts: NC-004			

Additional Project/Performance Site Location			
Organizational Name: North Carolina State University			
DUNS: 042092122			
Street 1: Research Admin/ SPARCS		Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514	
City: Raleigh	County: Wake		State: NC
Province:	Country: USA		Zip/Postal Code: 27695-7514
Project/Performance Site Congressional Districts: NC-02			

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location		
Organizational Name: Duke University		
DUNS: 044387793		
Street 1: Hock Plaza	Street 2: Box 2716 Med Ct.	
City: Durham	County: Durham	State: NC
Province:	Country: USA	Zip/Postal Code: 27705
Project/Performance Site Congressional Districts: NC-004		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Kosorok, Michael R.	Michael_Kosorok	UNC-CH	Core Director
Cai, Jianwen	Jianwen_Cai	UNC-CH	Core Contributor
Davidian, Marie	Davidian2	NC State University	Core Co-Director
George, Stephen L.	georg001	Duke University	Core Co-Director
Ibrahim, Joseph G.	JOE_IBRAHIM	UNC-CH	Core Co-Director
Jung, Sin-Ho	Jung0005	Duke University	Core Co-Director
Lin, Danyu	DANYU_LIN	UNC-CH	Core Contributor
Tsiatis, Anastasios	butch_tsiatis	NC State University	Core Co-Director

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project
Begg, Colin B.	Memorial Sloan-Kettering Cancer Center	Consultant
Murphy, Susan A.	University of Michigan	Consultant
Owzar, Kouros	Duke University	Consultant
Parmigiani, Giovanni	Johns Hopkins University	Consultant

Human Embryonic Stem Cells X No ☐ Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

METHODS AND SERVICES TO BE PROVIDED

1 SPECIFIC AIMS, BACKGROUND, AND RATIONALE

The Administrative Core (Core A) will be responsible for organizing the program investigators and staff into an effective and well-coordinated team to develop and implement statistical methods for cancer clinical trials that will relieve the bottleneck between biomedical science and clinical practice and improve the health of cancer patients.

1.1 Integration of Research Projects

The main goal of Core A is integration of the five research projects:

Project 1 (Cai): Innovative clinical trial design and analysis

Project 2 (Davidian): Methods for missing and auxiliary data in clinical trials

Project 3 (Ibrahim): Methods for post marketing surveillance and comparative effectiveness research

Project 4 (Lin): Methods for pharmacogenomics and individualized therapy trials

Project 5 (Tsiatis): Methods for discovery and analysis of dynamic treatment regimes

The projects are supported by two additional cores:

Core B (George): Data compilation

Core C (Davidian): Computational resource and dissemination core

Each project has its own research focus and aims. However, in concert, the projects are synergistic in ways that expand the scientific impact of the work. Core A will promote these synergies by facilitating and tracking interaction and integration among projects through the following mechanisms:

- continuing assessment of project and core objectives
- articulating program wide objectives that involve interaction and integration
- organization of monthly program review meetings and annual retreats
- facilitation of weekly and ad-hoc scientific group meetings
- facilitating tracking and communication across projects and cores
- coordination of inter-institutional computing resources.

Overall, this program project will have a non-traditional matrix organization, with projects and aims cutting across institutions as well as statistical areas of expertise cutting across projects. This organization will create a uniquely effective framework for making scientific advances that require multiple areas of expertise and multiple modes of attack. On the other hand, the traditional weekly meeting structure for each project and core is not in general going to be flexible enough to maximize these synergistic opportunities. Our basic approach will be to have regular monthly meetings of a Steering Committee for direction and coordination of the program as a whole in combination with weekly and ad-hoc scientific group meetings dictated by cross-cutting research themes and project-specific objectives. These weekly and ad-hoc scientific group meetings will be organized and directed as needed by project leaders and project investigators. Meetings will be tracked and minutes recorded on the program project wiki, to which all program project investigators and staff will have access.

Before describing integration plan components, we briefly review points of interaction between the projects that inform our integration process, and then we outline existing collaborations already in place as well as new collaborations enabled by the proposed program project. We then describe the key facets of the integration plan: the Executive Committee, the Steering Committee, the weekly and ad-hoc scientific group meetings, the program project web site and wiki, and the Inter-Institutional Computing Committee.

1.1.1 Project Interactions

Project 1, on clinical trial design and analysis, will benefit greatly from developments in Projects 2 and 4, because of both the ubiquitous presence of missing data and the emerging importance of pharmacogenomic information in clinical trials. Project 2, on missing data, will contribute greatly to Project 1, as already mentioned above, as well as to Projects 3, 4, and 5, because missing data always occur in clinical studies. The techniques for finding candidate treatment regimens from existing data developed in Project 3 will inform and contribute to the range of treatments and treatment regimens studied in Projects 4 and 5. The surveillance methodology also developed in Project 3 will feed into the clinical trial design issues examined in Project 1. The candidate individualized therapies and model selection tools developed in Project 4 will be directly applicable to the dynamic treatment regime research of Project 5. The personalized dynamic treatment regimes developed in Project 5 will suggest treatment protocols that can be evaluated in the randomized trials developed in Project 1. In addition to this synergy among the scientific goals, there are a number of cross-cutting themes that transcend multiple projects. Two such themes are statistical learning, which is involved in Projects 3, 4, and 5, and model selection, which plays a role in all of the projects but especially in Projects 2–5. Two other such themes are empirical processes and semiparametric inference which cuts through all 5 projects. There are many additional cross-cutting themes and synergies among the projects, which we do not exhaustively enumerate here but which have played and will continue to play key roles in the proposed research program.

1.1.2 Existing and New Collaborations

The proposed program project includes a number of existing collaborations on topics related to the program that have resulted in either publications or funded grants. Examples of collaborations among University of North Carolina at Chapel Hill (UNC-CH) personnel in biostatistical methodology areas related to the proposed program project include: Drs. Cai and Kosorok on semiparametric methods and time-to-event data (Song et al., 2008a); Drs. Cai and Zeng on joint modeling of longitudinal and time-to-even data (Zeng and Cai, 2005); Drs. Ibrahim and Zeng on missing data (Chen et al., 2006); Drs. Chu, Ibrahim, and Sandler on statistical methodology in cancer (Qu et al., 2008); Drs. Fine and Kosorok on semiparametric methods (Lee et al., 2005) and on microarray methods (Ma et al., 2006); and Drs. Lin and Zeng on statistical genetics (Lin and Zeng, 2006), along with many other joint publications. Examples of collaborations among North Carolina State University (NCSU) personnel in biostatistical methodology areas related to the proposed program project include: Drs. Davidian and Stefanski on longitudinal data (Huang et al., 2009); Drs. Davidian and Tsiatis on joint modeling of longitudinal and time-to-event data (Tsiatis and Davidian, 2004); along with numerous other joint publications. Examples of collaborations among Duke University personnel in biostatistical methodology areas related to the proposed program project include: Drs. George and Jung on clinical trial methods (Jung and George, 2009); Drs. Wang and Pang on cancer biomarkers (Wang et al., 2009); along with many other joint publications.

Examples of collaborations related to the proposed program among personnel spanning two or more of the three institutions include: Drs. Fine at UNC-CH and Tsiatis at NCSU, who collaborated in the areas of semiparametric methods and time-to-event data (Fine and Tsiatis, 2000); Drs. Liu at UNC-CH and Zhang at NCSU, who collaborated on statistical model selection (Zhang et al., 2008); and Drs. Zhou at UNC-CH and Wang at Duke, who collaborated in the area of semiparametric methods (Wang and Zhou, 2006). In addition, Drs. Davidian and Tsiatis at NCSU are adjunct faculty in the Duke Department of Biostatistics and Bioinformatics and have collaborated with faculty there. Moreover, Drs. Kosorok, Cai, Fine, Ibrahim, Wright, and other investigators at UNC-CH teamed up with Drs. Davidian and Tsiatis at NCSU to formulate the Biostatistics Core of the recently funded NIH Clinical and Translational Science Award (1 UL1 RR025747-01) to UNC-CH and partners (including NCSU). Drs. Davidian and Tsiatis at NCSU and Drs. Kosorok, Lin, and Zeng at UNC-CH participated in the Statistical and Applied Mathematical Sciences Institute (SAMSI) summer 2007 research program on "Dynamic Treatment Regimes and Multistage Decision-Making" and initiated collaborations on dynamic treatment regimes and reinforcement learning in clinical trials that formed the basis for Project 5 in the proposed program project. Overall, there exist many strong research collaborations among the program investigators across the three institutions.

Nevertheless, many of the specific collaborations delineated in the proposed projects and cores are new, especially in terms of the breadth and depth of the inter-institutional collaborations. For example, the collaborations between UNC-CH and Duke on Projects 1 and 4 are new. The three-way collaboration between UNC-CH,

NCSU, and Duke on Project 4 is also new. There are very few geographic locations where three strong institutions with significant biostatistical research groups are so close together as occurs with UNC-CH, NCSU, and Duke. The proposed program project will facilitate and leverage this unique and powerful consortium that can clearly exceed the sum of its parts. The proposed integrated projects and multi-institutional synergy will enable dramatic progress in cancer clinical trial methodology that will lead to improved public health.

1.1.3 Executive Committee

The Executive Committee will consist of the three PD/PI's of the program project: Dr. Kosorok at UNC-CH, lead PD/PI and Core A Director; Dr. Davidian at NCSU, PD/PI, Project 2 Leader and Core C Director; and Dr. George at Duke, PD/PI and Core B Director. Dr. Kosorok will chair this committee, which provides overall administrative leadership for the program project. Additional details on the role of the Executive Committee are provided under the Administration section below.

1.1.4 Steering Committee

The Steering Committee will consist of the Executive Committee and also Drs. Ibrahim, Tsiatis, Jung, Cai, and Lin. Dr. Ibrahim at UNC-CH is a co-PD/PI and the Project 3 Leader. Dr. Tsiatis at NCSU is a co-PD/PI and the Project 5 Leader. Dr. Jung at Duke is a co-PD/PI. Drs. Cai and Lin at UNC-CH are the Project Leaders for Projects 1 and 4, respectively. Dr. Kosorok will chair the committee. The role of the Steering Committee will be to provide overall scientific leadership for the program project, continuing assessment of project and core objectives, and articulation of program-wide objectives that involve interaction and integration across projects and cores. The Steering Committee has been meeting in the process of preparing for this grant for at least six months and will continue to meet monthly throughout the program project funding period. Most of these monthly meetings will be via teleconference, but once a quarter the meetings will be face-to-face at the National Institutes of Statistical Sciences (NISS) facilities in the Research Triangle Park, which is centrally located to all three institutions involved on the program project and where we have already met several times. Once a year, one of these face-to-face meetings at NISS will be part of the the program project annual retreat that will be described in greater detail below. Minutes from these meetings will be placed on the program project wiki for tracking and communication to all program project personnel.

1.1.5 Weekly and Ad-hoc Scientific Group Meetings

We will have weekly and ad-hoc scientific group meetings as dictated by cross-cutting research themes and project-specific objectives. These weekly and ad-hoc scientific group meetings will be organized and directed as needed by project leaders and project investigators. Each project and core will meet at least monthly, usually in conjunction with the Steering Committee meetings. Whether or not these project/core meetings are face-to-face or are teleconference or web conference meetings will depend on the needs of the project or core and the issues to be addressed.

There will also be regular meetings based on cross-cutting research themes. For example, there will be weekly face-to-face meetings on the theme of statistical learning (e.g., machine learning and reinforcement learning), which cuts across Projects 3, 4, and 5. These statistical learning meetings have been underway for about a year in the Department of Biostatistics at the University of North Carolina at Chapel Hill (UNC-CH), under the label "Reinforcement Learning Group," and some of the research discussed in those meetings played a role in the development of some of the specific aims in Projects 3, 4, and 5. We will also initiate a meeting series in the Center for Quantitative Sciences in Biomedicine (CQSB) at NCSU on the theme of model selection, which cuts across Projects 2–5. Additional research-thematic and project-focused meetings will be instigated after the program project is underway. The themes and frequency of these meetings will depend on the needs of the overall program, projects, and cores. Our goal is to be flexible and efficient rather than highly formalized with these meetings. We believe that this mix between a formal and ad-hoc structure will maximize our flexibility and adaptability in tackling the complex problems we are addressing in this program project as well as maximize our impact on public health.

1.1.6 Program Project Web Site and Wiki

A dedicated web site for our program project will be developed and housed on a server at UNC-CH. The purpose of this web site is to be a single point of contact for all interested parties. The main page will be accessible to the

general public and will include links to both an external set of pages also available to the general public as well as internal pages available only to program investigators. The external pages will include general information about the program projects, links to published papers and software developed by the program project, along with instructions and tutorials and other items for outreach. The internal pages will include a program project wiki for tracking and communicating among investigators, sharing data, developmental software and other digital information for investigators.

For the preparation of this proposal, we have been using a wiki that is a prototype of the proposed program project wiki. The prototype wiki is currently located on a server in the Department of Statistics at NCSU and is powered by Media-Wiki (http://www.mediwiki.org) shareware. This wiki includes descriptions of all projects and cores as well as minutes from Steering Committee meetings. In the first few weeks of the proposed project, we will move the current wiki to a new server located at UNC-CH dedicated to Core A functions. The transfer of the contents of the wiki is extremely easy at this stage, as no large files are involved, and the number of pages is below 20. The new wiki will include pages for each project and core as well as other resources for the investigators. Each page has the capacity for editing and adding, deleting, and changing additional pages as well as minutes, papers, figures, and short comments used to track progress and communicate between investigators. This is an extremely powerful tool for research coordination that increases research productivity and collaboration between scientists and across the three institutions. We anticipate that the utilization of this wiki will dramatically increase over the first year of the project and continue to grow throughout the five years as progress continues to be made on the proposed research. The wiki will greatly facilitate the matrix and cross-cutting aspects of the program project.

1.1.7 Inter-Institutional Computing Committee

An Inter-Institutional Computing Committee (IICC) will direct computing issues that involve interactions between institutions. This includes addressing processes for transmitting data sets, data resources, software, and other digital resources among the three Cores and among all of the investigators. This committee will be chaired by Dr. Davidian from NCSU and will include as members Dr. Owzar from Duke, who is the co-leader of Core B over computing, and Dr. Kosorok. This committee will meet on an ad-hoc basis as needed but at least four times per year. The IICC has already met several times and has been very effective at addressing the coordination of computing resources between the three institutions. The IICC will orchestrate coordination between the three core servers, with the Core A server located at UNC-CH, the Core B server located at Duke, and the Core C server located at NCSU. The Committee will ensure that the severs can communicate and transfer data among themselves so that all investigators have appropriate access to all core services and that there is duplication of data and software so that, if one institution loses service, there is a backup at at least one other institution. Core A will have no access to identifiable humans subjects data.

1.2 Annual Retreat and Meeting with External Advisory Committee

Core A will organize an annual retreat that includes the Steering Committee and all project investigators and staff as well as a review by the External Advisory Committee (EAC) described below. The meeting will take place at the NISS building located in Research Triangle Park, which is central to the three institutions and near the Radisson Hotel Research Triangle Park. This will be convenient for both the program project investigators and the EAC members, who will be traveling from out of state. A summary of progress to date on the program project will be sent to the EAC members three weeks prior to the meeting. On the day of the meeting, the agenda will consist of an open meeting/retreat for all participants, which will include presentations on scientific progress and accomplishments for all projects and cores as well as a report on overall progress. This will be followed by an executive session for the EAC and a wrap-up meeting with both the EAC and the Steering Committee. Shortly after the meeting, the EAC will provide a written report of the current status of the program project, including commendation of strengths as well as suggested recommendations for improvement. The date of this meeting will be chosen to allow incorporation of results from the meeting and the EAC report into the annual NIH progress report.

1.3 Cost Accounting

The core will maintain cost-accounting of program budgetary resources as well as coordinate submission of annual progress reports across the three institutions. A program administrative manager at UNC-CH, reporting

to Dr. Kosorok, will generate quarterly budget reports for the entire project, detailing expenditures of each project and core by institution. This manager will be assisted by administrative managers at NCSU and Duke, reporting to Drs. Davidian and George, respectively. These managers will also assist Drs. Kosorok, Davidian and George in preparing documentation for the annual progress reports. Submission of annual reports and other official program communications will be coordinated through UNC-CH.

2 ADMINISTRATION

The administrative structure for this program project consists of three interwoven components: overall program administration, intra-institutional administration, and project and core leadership. This is represented in Figure 1. In this figure, boxes delineate administrative units, dotted lines denote a supportive or advisory relationship, solid lines denote a supervisory relationship, and double arrows denote two-way support or joint supervisory roles. Dr. Kosorok at UNC-CH will be the lead PD/PI and provide overall leadership for the program project. Dr. Kosorok's overall leadership responsibilities will be shared by two additional PD/PIs, Dr. Davidian at NCSU and Dr. George at Duke, who, together with Dr. Kosorok will form the Executive Committee. Each member of the Executive Committee will be assisted by an additional co-PD/PI at his/her institution to help with both overall leadership and intra-institutional leadership: Dr. Ibrahim at UNC-CH, Dr. Tsiatis at NCSU, and Dr. Jung at Duke. The PD/PI and co-PD/PI at each institution, along with one or more additional administrative personnel, will constitute an intra-institutional Administrative Office to provide intra-institutional administrative support and leadership. The Executive Committee and the three Administrative Offices will provide administrative support to the Steering Committee–composed of the Executive Committee, co-PD/PIs, and individual project and core leaders (Drs. Kosorok, Davidian, George, Ibrahim, Tsiatis, Jung, Cai and Lin)—which provides overall scientific leadership to the program. The project and core leadership will be as delineated in the project and core descriptions and cuts through and across both the overall program and intra-institutional administration.

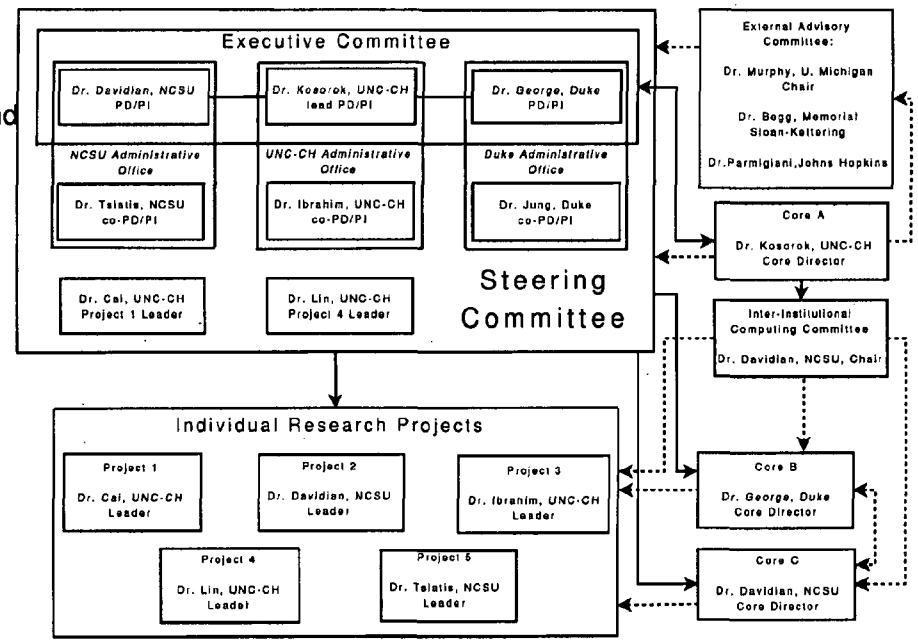


Figure 1: Program project leadership structure.

Dr. Kosorok's duties include managing the integration of the five projects and three core facilities; leading Core A operations; scheduling and conducting regular meetings of the Executive Committee, the UNC-CH Administrative Office, the Steering Committee, and the Annual Retreat and Meeting with the External Advisory Committee. The lead PD/PI will also coordinate yearly summaries of progress and non-competitive renewal materials. Other activities include preparation of annual reports, oversight of budgetary disbursements and financial records, and communication with NIH and other entities that impact the program.

Dr. Kosorok will be assisted in his overall leadership responsibilities and duties by Drs. Davidian and George, who comprise the Executive Committee. Dr. George is also the Core B leader, and, as such, will coordinate trans-institutional data sharing, data compilation and human subjects issues. Dr. George will also have overall responsibility for the Duke budgetary components of the program. Dr. Davidian is also the Core

C leader, Project 2 leader, and chair of the inter-institutional computing committee, which will coordinate trans-institutional computing issues. Drs. Davidian and George will also assist Dr. Kosorok in general with external communications regarding the program, although Dr. Kosorok will be the primary contact person.

The Executive Committee will meet monthly and additionally as needed, mostly via phone conference but also in person at least quarterly. Dr. Kosorok and the Executive Committee will be responsible for the day-to-day management of the overall program. Overall program management activities include organizing and conducting regular meetings of the Executive and Steering Committees, both via phone conference and in person; organizing the annual retreat and meeting with the External Advisory Committee; addressing inter-institutional computing issues; and preparation of annual reports, financial records and non-competitive renewal documents.

Each institutional Administrative Office will also meet monthly and additionally as needed, for day-to-day management of the intra-institutional components of the program at each institution. Intra-institutional management activities include coordination of intra-institutional meetings, computing issues, and budget, and gathering institutional-specific information for the program annual reports.

Dr. Kosorok will be assisted administratively by Ms. Betsy Seagroves, Ms. Evie McKee, and Ms. Tania Osborne. Ms. Seagroves, who serves in the Department of Biostatistics as Administrative Assistant to the Chair (Dr. Kosorok), will assist Dr. Kosorok and the UNC-CH Administrative Office with scheduling related issues. Ms. McKee is the Business and Grants Manager for the Department of Biostatistics and will provide guidance to Dr. Kosorok and Ms. Osborne on budgetary matters. Ms. McKee and Osborne will assist Dr. Kosorok in the day-to-day activities for scheduling, budgetary items and reporting, as well as in handling mail and telephones, maintaining project files, and managing travel of the External Advisory Committee members. Drs. Davidian and George will also be assisted in day-to-day administrative matters by administrative assistants at NCSU and Duke, respectively.

Dr. Kosorok will be responsible for preparation of annual progress reports and dispersion of funds to the subcontracts at NCSU and Duke as well as to the project and core components at UNC-CH. Drs. Davidian and George will be responsible for assisting with the progress reports as well as dispersing funds to the project and core components at NCSU and Duke, respectively. No budgetary changes in subcontracts will be undertaken without approval from the Administrative Office at the affected institution. No budgetary changes affecting projects or cores will be undertaken without consulting with the affected project or core leaders nor without approval of the Executive Committee.

The proposed administrative structure is not a strict hierarchy but is more of a matrix with both intra-institutional leadership under the Administrative Offices and trans-institutional scientific leadership of the overall program as well as of individual projects and cores. While Dr. Kosorok will be the lead PD/PI, the overall leadership and responsibility for the core is shared among all members of the Executive Committee, and no major decisions affecting the project will be made without the consensus of the entire Executive Committee. Moreover, no major changes in scientific focus or budget allocations to projects and cores will be made without input and guidance from the Steering Committee (which includes all project and core leaders). We have been successfully functioning in this manner for about six months now as we have been preparing this application. We do not anticipate there being any conflicts or difficulties that cannot be successfully resolved within this administrative structure. More importantly, the novel matrix administrative structure will facilitate research more effectively than the usual hierarchical approach because of the ability of the matrix approach to foster and coordinate trans-institutional collaboration.

3 EDUCATION

Core A will also organize and facilitate educational opportunities in biostatistical methods research for the student research assistants assigned to the program project. This will include providing opportunities for students to work on site at at least two different institutions on at least two different projects or cores. All three institutions and all projects and cores will participate. We expect that many of the student research activities on the program project will lead to dissertation topics for these students. Core A will also coordinate tracking and reporting of educational accomplishments of students involved on the program project. In addition, students will have opportunities to participate and present at the annual retreat and meeting with the External Advisory Committee, as well as to present project research at conferences.

4 OUTREACH AND DISSEMINATION

All major research results will be published and disseminated through the usual refereed journals and conference proceedings. We also plan to organize a special session at the annual Spring Meeting of Eastern North American Region (ENAR) Meeting of the International Biometric Society (IBS) and/or the Joint Statistical Meetings sponsored by the American Statistical Association, ENAR, and three other major statistical societies. These two meeting venues are the two most well-known professional meetings for biostatistics. We will provide a public web site to provide generally-accessible up-to-date information on our research progress. We will also present research results at conferences of organizations such as the Society for Clinical Trials and the American Society of Clinical Oncology.

Core A will also maintain a liaison with publicity offices in the Department of Biostatistics, the Gillings School of Global Public Health, and the College of Arts and Sciences at UNC-CH; the CQSB, Department of Statistics, and the College of Physical and Mathematical Science at NCSU; and the Department of Biostatistics and Bioinformatics and the School of Medicine at Duke. Important results will be appropriately publicized at the time of publication through news releases to the local media.

4.1 Data and Software Sharing

We also recognize the importance of data and software sharing. Because we will not be generating new clinical data but will only be using existing data with all identifiers removed, there will not be much opportunity to share clinical data outside of the program project investigators. On the other hand, we will be sharing our newly developed software, training and instructional material associated with the new software, and a small number of accompanying demonstration data sets with no identifiers.

Core B will be responsible for gathering clinical data sets without identifiers and formatting them for sharing between the three institutions for research purposes within the program project. Core B will also assist with preparing demonstration data sets for use in demonstrating and training for newly developed software. Core C will be responsible for preparing, streamlining, and testing the newly developed software for dissemination to all interested researchers. Core C will also put together training and demonstration materials for the new software, including in some cases demonstration data sets provided by Core B. The new software and educational materials developed in Core C will be available through the Program Project Web Site. Core A will coordinate the communication and dissemination between and across Cores B and C as well as outreach functions in general.

4.1.1 Program Project Web Site

Current news about the program project and links to articles will also be maintained on the Program Project Web Site which will be maintained as part of the Core A functions. The main page of this web site will be accessible to all who are interested. The public part will consist of information about the program project, news releases, links to research articles, and links to the newly developed software and associated educational materials. All interested researchers will have full access to the software and training materials. There will also be a link to a private, internal part of the web, accessible only to project investigators. The private part of the web will also include the Program Project Wiki described above. The proposed, one-stop Program Project Web Site will allow easy communication both internally among program investigators as well as externally to the public. The web site will be managed by Dr. Kosorok with assistance from the Executive Committee, the UNC-CH Administrative Office, and the Inter-Institutional Computing Committee.

5 FACILITIES

Core A (Kosorok) will be primarily housed in the Department of Biostatistics located in McGavran-Greenberg Hall in the Gillings School of Global Public Health at UNC-CH, with Administrative Offices also located at NCSU and Duke. The UNC-CH Department of Biostatistics, located in the UNC Gillings School of Global Public Health, is one of the largest and highest ranked Biostatistics Departments in the U.S. The Department has over 35 full time faculty members; 130 graduate students pursuing MPH, MS, DrPH and PhD degrees; and 15 undergraduate students. The Department occupies a total of 26,333 square feet; outstanding computer support for all students, faculty and staff; and a state-of-the-art 400 square-foot conference room which seats 20 people and has a drop-down projector, wireless capabilities and conference calling facilities. The Department of Biostatistics at UNC-CH is very supportive of this program project and will do all in its power to ensure its

ongoing success. The Department will contribute $30,000 per year to the program project for all five years of the grant. Dr. Kosorok, as lead PD/PI of the program project, is the Chair of the Department and will ensure that the needed support is provided. UNC-CH is one of the nations foremost research universities, with top rankings in many disciplines. The Gillings School of Global Public Health is the top ranked public school of public health and has seven academic departments, including the Department of Biostatistics, in addition to several centers, programs and institutes. The School also has several high-tech conference rooms, including the Blue Cross and Blue Shield of North Carolina Foundation Auditorium. Both the Gillings School and UNC-CH have pledged their support, including contributing $8,000 and $20,000, respectively, annually to the program project.

The administrative home for the NCSU component of the Project will be the CQSB (Center for Quantitative Sciences in Biomedicine), for which Dr. Davidian, the NCSU PD/PI for the Project, serves as Director and in which several investigators on the Project are members. The Center is jointly supported by the NCSU Colleges of Physical and Mathematical Sciences as well as Agriculture and Life Sciences and shares a state-of-the-art conference facility and two smaller conference rooms each seating 10-15 people with the Center for Research in Scientific Computing (CSRC) on the third floor of Cox Hall. The main conference facility seats 24-30 people in different configurations for conferences, seminars, and instructional events, and has the latest technology, including LCD projection equipment to display presentations on screens at the front and back of the room, Smart Board and Symposium technology, and video-conferencing capabilities. Almost all NCSU Project Investigators have offices in the NCSU Department of Statistics which is one of the largest and oldest departments of statistics or biostatistics in the world, with approximately 70 undergraduate and 170 graduate students, and enjoys excellent office and meeting room facilities and excellent computer support for all of its faculty, students and staff. The Department is very supportive of this program project. With UNC-CH, NCSU is one of the two flagship research institutions of the University of North Carolina system, with major colleges and schools of Agriculture and Life Sciences, Design, Education, Engineering, Humanities and Social Sciences, Management, Natural Resources, Physical and Mathematical Sciences, Textiles, and Veterinary Medicine. The College of Physical and Mathematical Sciences, which houses CQSB and the Department of Statistics, and the College of Agriculture and Life Sciences, which also supports CQSB, are very supportive of the proposed program project, and each will contribute $17,500 to the project in each of the five years of the grant. The University's Vice Chancellor for Research and Graduate Studies will contribute an additional $25,000 in each of the five years.

All of the biostatistical investigators at Duke are faculty members in the Department of Biostatistics and Bioinformatics and the clinical co-investigators at Duke are all members of the DCCC with appointments in the Department of Medicine or the Department of Surgery at Duke. The Biostatistics and Bioinformatics Department currently consists of two divisions: Biostatistics (38 faculty members) and Computational Biology (7 faculty members) and has excellent office and meeting facilities and computational support for all of its faculty, students and staff. The Department of Biostatistics and Bioinformatics at Duke is very supportive of this program project and will contribute $ 15,000.00 per year to the program project for all five years of the grant. Duke University, founded in 1924, is a top ranked private school with many schools and colleges that are highly ranked nationally. The College of Arts and Sciences houses the nationally recognized Department of Statistical Science and the Duke University School of Medicine houses the Biostatistics and Bioinformatics Department. The Duke University School of Medicine is ranked in the top ten with schools twice its age, and is committed to socially relevant education, translational research, compassionate patient care and global healthcare solutions. There is ample meeting space for all program project investigators. The School of Medicine is very support of this program project and will contribute $15,000 per year in addition to the contribution from the Department of Biostatistics and Bioinformatics.

The NISS (National Institute of Statistics Sciences) building is located centrally to the three participating institutions in Research Triangle Park. The building is shared with its sister institute, SAMSI, and has 28,000 square feet of state-of-the-art office and meeting space. The meeting space includes traditional conference rooms, a fully-equipped video conference room, and a lecture and conference room that supports web streaming. Dr. Alan F. Karr, NISS Director, has provided space to us during the past several months and has agreed to continue providing meeting space to our program project without charge.

6 EXPERTISE

Michael R. Kosorok, PhD, Core A Director and Lead PD/PI. Dr. Kosorok is Professor and Chair of Biostatistics and Professor of Statistics and Operations Research at UNC-CH. He is also director of the Biostatistics Core of the UNC-CH CTSA grant. His expertise is in biostatistical methodology for clinical trials, survival analysis, and high dimensional biomedical data, including microarrays, machine learning, and reinforcement learning. He is also an expert in the theoretical foundations underlying biostatistical methodology, especially in the areas of empirical processes and semiparametric inference, and has written a text on the subject (Kosorok, 2008). His contributions have been widely recognized. He is also an associate editor of the *Annals of Statistics* and elected Fellow of both the American Statistical Association and the Institute of Mathematical Statistics. He will direct the entire program project, direct Core A, co-lead Projects 3, 4 and 5, and be a co-investigator on Project 1. He will chair the Executive and Steering Committees, lead the UNC-CH Administrative Office, and be a member of the Inter-Institutional Computing Committee.

Jianwen Cai, PhD, Core A Contributor. Dr. Cai is Professor and Associate Chair of Biostatistics at UNC-CH. She is an internationally recognized expert on clinical trials, survival analysis, and semiparametric methods. She is an elected Fellow of the American Statistical Association. She will lead Project 1, co-lead Core B, and be a co-investigator on Project 2 and a contributor on Core A. She will be a member of the Steering Committee.

Marie Davidian, PhD, Core A Co-Director and PD/PI. Dr. Davidian is William Neal Reynolds Professor in the Department of Statistics and Director of the Center for Quantitative Sciences in Biomedicine at NCSU and Adjunct Professor of Biostatistics and Bioinformatics at Duke. She also serves as Executive Editor of *Biometrics*, regarded by many to be the top journal in the field of biostatistics. She is internationally recognized for her work in longitudinal data, missing data, biomedical modeling, clinical trials, and semiparametric methods. She is an elected Fellow of both the American Statistical Association and the Institute of Mathematical Statistics. She will co-direct the entire program project, lead Project 2, direct Core C, co-lead Project 5, and co-direct Cores A and B. She will be a member of the Executive and Steering Committees, lead the NCSU Administrative Office, and chair the Inter-Institutional Computing Committee.

Stephen L. George, PhD, Core A Co-Director and PD/PI. Dr. George is Professor of Biostatistics and Bioinformatics at Duke and Director of Biostatistics for the Duke Comprehensive Cancer Center and for the Cancer and Leukemia Group B (CALGB). His expertise is in clinical trials, translational science, and prognostic and predictive models. His work is widely recognized. He will co-direct the entire program project, direct Core B, co-lead Project 1, and co-direct Core A. He will be a member of the Executive and Steering Committees and lead the Duke Administrative Office. Dr. George is an elected Fellow of the American Statistical Association.

Joseph G. Ibrahim, PhD, Core A Co-Director and co-PD/PI. Dr. Ibrahim is Alumni Distinguished Professor of Biostatistics at UNC-CH, the Lineberger Comprehensive Cancer Center Director of Biostatistics, and the Director of the UNC Center for Innovative Clinical Trials. His widely recognized expertise is in Bayesian methods, missing data, clinical trials, and cancer genomics. He is an elected Fellow of the American Statistical Association and the Institute of Mathematical Statistics. He will co-direct the entire program project, lead Project 3, co-lead Projects 1 and 2, and co-direct Cores A and B. He will be a member of the Steering Committee and the UNC-CH Administrative Office.

Sin-Ho Jung, PhD, Core A Co-Director and co-PD/PI. Dr. Jung is Professor of Biostatistics and Bioinformatics at Duke. Dr. Jung serves as the Director of the CALGB Biostatistics unit, and as the faculty statistician for the CALGB Lymphoma and Imaging Committees and the Cancer Prevention Subcommittee. He has published design and analysis methods for a wide range of cancer studies and clinical trials. These methods include survival analysis, various types of clustered and longitudinal data analysis, and design and analysis methods for phase II cancer clinical trials. His recent research interests are focused on design and analysis methods for high-throughput projects, including DNA microarrays, SNP and proteomic studies. His work is widely recognized. He will co-direct the entire program project, co-lead Projects 1, 2 and 4, and co-direct Cores A and B. He will be a member of the Steering Committee and the Duke Administrative Office.

Danyu Lin, PhD, Core A Contributor. Dr. Lin is Dennis Gillings Distinguished Professor of Biostatistics at UNC-CH. He is a widely recognized expert in clinical trials, survival analysis, genomics, and semiparametric methods. He is an elected Fellow of the American Statistical Association and the Institute of Mathematical Statistics. He will lead Project 4, co-lead Core C, and be a contributor on Core A. He will be a member of the

Steering Committee.

Kouros Owzar, PhD, Core A Consultant. Dr. Owzar is Associate Professor of Biostatistics and Bioinformatics at Duke. His expertise is in pharmacogenomics, bioinformatics, nonparametrics, and statistical computing. His work is widely recognized. He will co-lead Project 4, co-direct Cores B and C, and be a contributor on Core A. He will be a member of the Inter-Institutional Computing Committee.

Anastasios A. Tsiatis, PhD, Core A Co-Director and co-PD/PI. Dr. Tsiatis is Drexel Professor of Statistics at NCSU and Adjunct Professor of Biostatistics and Bioinformatics at Duke. His expertise is in survival analysis, causal inference, clinical trials, and semiparametric methods. He has written a text on the foundations of semiparametric methods (Tsiatis, 2006). His work is widely recognized. He is an elected Fellow of the American Statistical Association and the Institute of Mathematical Statistics. He will co-direct the entire program project, lead Project 5, co-lead Project 2, co-direct Cores A and B, and be a co-investigator on Project 1. He will be a member of the Steering Committee and the NCSU Administrative Office.

7 DECISION-MAKING

Dr. Kosorok will have authority for all administrative decisions but shares that authority and operates in concert with the other members of the Executive Committee, Drs. Davidian and George. Dr. Kosorok and the Executive Committee will have final authority for all program project decisions. The PD/PIs and co-PD/PIs constituting the Administrative Offices at each institution will have final administrative authority on program project matters at their respective institutions. The Executive Committee (Drs. Kosorok, Davidian and George), co-PD/PIs (Drs. Ibrahim, Tsiatis and Jung), and the other individual project leaders (Drs. Cai and Lin) will serve as the Steering Committee. The Steering Committee will meet monthly as described previously and will have the power to reallocate funds within the program to respond to unanticipated results, challenges and opportunities. Such decisions will be made by consensus and must have the unanimous support of the Executive Committee. Drs. Davidian and George will act for Dr. Kosorok when inter-institutional administrative action is needed in his absence. Drs. Ibrahim, Tsiatis, and Jung will act, respectively, for Drs. Kosorok, Davidian, and George when intra-institutional administrative action is needed within their respective institutions.

8 EXTERNAL ADVISORY COMMITTEE

Deliberations of the Steering Committee will be assisted by an External Advisory Committee consisting of three experts in clinical trial methodology with an international reputation who are external to the three involved institutions. Dr. Susan A. Murphy, H. E. Robbins Professor of Statistics, Professor of Psychiatry, and Research Professor in the Institute for Social Research at the University of Michigan, will serve as Chair of the committee. Dr. Murphy is widely regarded as the world's foremost authority on statistical methodology for clinical trials with dynamic treatment regimes. Drs. Colin B. Begg and Giovanni Parmigiani will serve as members of the committee. Dr. Begg is Attending Biostatistician and Chair of the Department of Epidemiology and Biostatistics at the Memorial Sloan-Kettering Cancer Center in New York and is one of the nation's leading biostatistical methodologists and practitioners in cancer clinical trials. Dr. Parmigiani is Professor of Biostatistics at Johns Hopkins University and is one of the nation's leading biostatistical methodologists and practitioners in cancer genomics. This distinguished group will provide feedback to the steering committee on the progress and goals of the program project on at least an annual basis during the annual retreats.

9 INCLUSION ENROLLMENT REPORT

N/A

10 BIBLIOGRAPHY AND REFERENCES CITED

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2006). Posterior propriety and computation for the Cox regression model with applications to missing covariates. *Biometrika* **93**, 791–807.

Fine, J. P., and Tsiatis, A. A. (2000). Testing for differences in survival with delayed ascertainment. *Biometrics* **56**, 145–153.

Huang, X., Stefanski, L. A., and Davidian, M. (2009). Latent-model robustness in joint models for a primary endpoint and a longitudinal process. *Biometrics*, in press. NIHMSID:NIHMS89685.

Jung, S. H. and George, S. L. (2009). Between-arm comparisons in randomized phase II trials. *Journal of Biopharmaceutical Statistics*, in press.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* New York: Springer.

Lee, B. L., Kosorok, M. R., and Fine, J. P. (2005). The profile sampler. *Journal of the American Statistical Association* **100**, 960–969.

Lin, D. Y., and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* **101**, 89–118.

Ma, S., Kosorok, M. R., Fine J. P. (2006). Additive risk models for survival data with high dimensional covariates. *Biometrics* **62**, 202–210.

Qu, P., Chu, H., Ibrahim, J. G., Peacock, J., Shen, X. J., Tepper, J., Sandler, R. S., and Keku, T. O. (2008). Statistical strategies to improve the efficiency of molecular studies of colorectal cancer prognosis. *British Journal of Cancer* **99**, 2001–2005.

Song, R., Kosorok, M. R., and Cai, J. (2008). Robust covariate-adjusted log-rank statistics and corresponding sample size formula for recurrent events data. *Biometrics* **64**, 741–750.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809-834.

Wang, X., Pang, H., and Schwartz, T. A. (2009). Building and Validating High Throughput Lung Cancer Biomarkers. *Chance*: under revision.

Wang, X., and Zhou, H. (2006). A semiparametric empirical likelihood method for biased sampling schemes in epidemiologic studies with auxiliary covariates. *Biometrics* **62,** 1149-1160.

Zeng, D. and Cai, J. (2005). Asymptotic results for maximum likelihood estimates in joint analysis of repeated measurements and survival time. *The Annals of Statistics* **33,** 2132-2163.

Zhang, H. H., Liu, Y., Wu, Y., and Zhu, J. (2008). Variable selection for multicategory SVM via sup-norm regularization. *Electronic Journal of Statistics* **2**, 149–167.

11 PROTECTION OF HUMAN SUBJECTS

Core A generally will not be involved in the analysis or review of data. All data for the program project comes through Core B (Data Compilations Core). In the event an issue arises where it is necessary for Core A to discuss or review data from Core B, the investigators will have access only to the de-identified data. Thus, the investigators on Core A will have no access to any identifiable patient information.

12 INCLUSION OF WOMEN AND MINORITIES

N/A

13 TARGETED/PLANNED ENROLLMENT TABLE

N/A

14 INCLUSION OF CHILDREN

N/A

15 VERTEBRATE ANIMALS

N/A

16 SELECT AGENT RESEARCH

N/A

17 MULTIPLE PD/PI LEADERSHIP PLAN

N/A

18 CONSORTIUM/CONTRACTUAL ARRANGEMENTS

If the present application is funded, the University of North Carolina at Chapel Hill will execute subcontracts with the consortium institutions (Duke University and North Carolina State University). These inter-institutional agreements will be written consistent with the NIH consortium agreement policy.

19 LETTERS OF SUPPORT

- H. Shelton Earp, MD, Director of the UNC-CH Lineberger Comprehensive Cancer Center, Providing institutional support.
- Barbara K. Rimer, DrPH, Dean of the UNC-CH Gillings School of Global Public Health, Providing monetary and institutional support.
- Tony G. Waldrop, PhD, UNC-CH Vice Chancellor for Research and Economic Development, Providing monetary and institutional support.
- Raymond E. Fornes, PhD, Associate Dean for Research of the NCSU College of Physical and Mathematical Sciences, Providing monetary and institutional support.
- Terri L. Lomax, PhD, NCSU Vide Chancellor for Research and Graduate Studies, Providing monetary and institutional support.
- Steven A. Lommel, PhD, Associate Dean for Research of the NCSU College of Agriculture and Life Sciences, Providing monetary and institutional support.
- Sastry G. Pantula, PhD, Chair of the NCSU Department of Statistics, Providing institutional support.

- Nancy C. Andrews, MD, PhD, Dean of the Duke University School of Medicine, Providing monetary and institutional support.
- Elizabeth R. DeLong, PhD, Chair of the Duke Department of Biostatistics and Bioinformatics, Providing monetary and institutional support.
- H. Kim Lyerly, MD, Director of the Duke Comprehensive Cancer Center, Providing institutional support.
- Richard L. Schilsky, MD, Chair of Cancer and Leukemia Group B, Providing institutional support.
- Alan F. Karr, PhD, Director of the National Institute of Statistical Sciences, Providing meeting space and facilities.
- Colin B. Begg, PhD, Memorial Sloan-Kettering Cancer Center, Member of Program Project External Advisory Committee.
- Susan A. Murphy, PhD, University of Michigan, Chair of Program Project External Advisory Committee.
- Giovanni Parmigiani, PhD, Johns Hopkins University, Member of Program Project External Advisory Committee.

UNC
LINEBERGER COMPREHENSIVE
CANCER CENTER
N.C. CANCER HOSPITAL

December 19, 2008

Michael R. Kosorok, PhD
Lead Principal Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Michael:

The strong partnership between UNC Lineberger Cancer Center and the Department of Biostatistics is decades old but the events of the last five years have taken it to another level. Recruitment has assembled an extraordinary group of statistical methodologic researchers, including Joe Ibrahim, Danyu Lin, Fred Wright, and Jason Fine, who contribute substantially to our cancer research. Under your leadership, an impressive era of scientific productivity is underway that will measurably improve clinical trials methodology.

As Director of the UNC Lineberger Comprehensive Cancer Center, I am writing to express my great enthusiasm and support for your application for an NCI Program Project (P01) entitled "Statistical Methods for Cancer Clinical Trials" at the University of North Carolina at Chapel Hill. This is certainly an important initiative for the UNC Lineberger, the Gillings School of Global Public Health, and for the University of North Carolina at Chapel Hill. With all of the recent advances in biomedicine, there remains a serious bottleneck between laboratory discoveries and their utilization in clinical practice. New clinical trials methodology is needed to keep abreast of and take advantage of molecular genetic discovery. I believe that the innovative program you have outlined will make important breakthroughs in solving this fundamental problem and have broad applicability for breast, colon and lung cancer as well as for other cancers and other diseases. I am very supportive of you and your research group utilizing existing clinical trial data sets housed in the Lineberger Comprehensive Cancer Center.

An important aspect of this program project is the collaboration with North Carolina State University and Duke University. This brings together a diverse group of investigators not only in biostatistics but also in medical oncology, health policy, pharmacogenomics, and computer science.

In summary, your program project application has my highest level of support and commitment. I will do all that I can to help you and your colleagues achieve the goals of this forward-looking P01 and, in the process, to help UNC become a leader in the field of cancer clinical trials.

Sincerely yours,

H. Shelton Earp III, MD
Director and Lineberger Professor
Professor of Medicine and Pharmacology

UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

170 ROSENAU HALL, CB 7400
135 DAUER DRIVE
CHAPEL HILL, NC 27599-7400

T 919.966.3215
F 919.966.7678
brimer@unc.edu

BARBARA K. RIMER, DrPH, MPH
Dean and Alumni Distinguished Professor

December 20, 2008

Michael R. Kosorok, PhD
Chair and Professor, Biostatistics
Lead Principal Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
UNC Gillings School of Global Public Health
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-7420

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Michael:

I write to express enthusiastic support for your National Cancer Institute Program Project (P01) application entitled "Statistical Methods for Cancer Clinical Trials" at the University of North Carolina at Chapel Hill. This is an important initiative for our School, the Lineberger Comprehensive Cancer Center and the University of North Carolina at Chapel Hill.

This is an interdisciplinary effort that brings together researchers from biostatistics, health policy, pharmacogenomics, medicine and computer science. The innovative program you have outlined in this proposal will significantly increase translation from laboratory discoveries to clinical practice which could lead to important improvements in public health – particularly in cancer research. I am especially enthusiastic about the cross-campus collaborations between Duke University, UNC and NC State University. The results of all of us working together could lead to great advances that ultimately can benefit patients.

We will contribute $8,000.00 per year toward this grant for each of the five years of the award. I wish we could do more, but we face additional rounds of budget cuts.

The program project "Statistical Methods in Cancer Clinical Trials" has my highest level of support and commitment. I pledge to do what I can to help achieve the goals you have set.

Warm regards,

Barbara K. Rimer

Barbara K. Rimer

BKR/smb

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

OFFICE OF THE VICE CHANCELLOR FOR
RESEARCH AND ECONOMIC DEVELOPMENT

312 SOUTH BUILDING
CAMPUS BOX 4000
CHAPEL HILL, NC 27599-4000

T 919-962-1319
F 919-962-1476

December 19, 2008

Michael R. Kosorok, PhD
Lead Principal Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Michael:

As Vice Chancellor for Research and Economic Development, I am writing to express very enthusiastic institutional support for your application for a National Cancer Institute Program Project (P01) entitled "Statistical Methods for Cancer Clinical Trials" at the University of North Carolina at Chapel Hill. This is certainly an important initiative for the Gillings School of Global Public Health, the Lineberger Comprehensive Cancer Center, and for the University of North Carolina at Chapel Hill. As you know, my own research background is in physiology, so I am attuned to the exciting developments in basic biomedical research knowledge and the enormous potential that exists to translate this knowledge into improvements in public health. Even with all of the recent advances in biomedicine, there remains a serious bottleneck between laboratory discoveries and their utilization in clinical practice. I believe that the innovative program you have outlined in this proposal will significantly relieve this bottleneck and lead to important improvements in public health, especially in cancer. I am also pleased that your collaborators include a diverse range of disciplines and departments across the university, not only in the Gillings School of Global Public Health but also researchers from the School of Medicine, the School of Pharmacy, and the College of Arts and Sciences.

Another important facet of this project is the collaboration with North Carolina State University and Duke University that will be both leveraged and facilitated by your program. We are very supportive of inter-university cooperation of this kind and recognize that this combination of institutions offers a uniquely powerful resource for making advances in clinical trial methods research that will have a high public health impact.

Because of the importance and value of this project for the university, we will contribute $20,000.00 per year towards this grant for each of the five years of the award.

In summary, the program project "Statistical Methods in Cancer Clinical Trials" has my highest level of support and commitment. I pledge to do whatever I can to see that we achieve the goals you have laid out and, in the process, become a leader in the field of cancer clinical trials.

Sincerely,

Tony G. Waldrop, PhD
Vice Chancellor for Research and Economic Development

North Carolina State University is a land grant university and a constituent institution of The University of North Carolina

College of Physical and Mathematical Sciences

NC STATE UNIVERSITY

Office of the Associate Dean for Research
Campus Box 8209 / 300 Cox Hall
Raleigh, NC 27695-8209
919.515.7865 (phone)
919.515.7668 (fax)

January 9, 2009

Marie Davidian, PhD
Program Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
Center for Quantitative Sciences in Biomedicine and
Department of Statistics
North Carolina State University
Raleigh, NC 27695

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Marie:

I am writing to offer my enthusiastic endorsement and commitment to your application for a National Cancer Institute Program Project (P01) Award, entitled "Statistical Methods for Cancer Clinical Trials," which will be a joint venture between North Carolina State University, Duke University, and the University of North Carolina at Chapel Hill. This important project is consistent with the College's emphasis on health-related research and in particular with the mission of the Center for Quantitative Sciences in Biomedicine (CQSB), which our College strongly supports and which will be the administrative home for the project.

I am well aware of the critical role the quantitative sciences and statistical science in particular play in the development of new methodology for the conception, design, and analysis of clinical trials, and I am excited at the prospect that the innovative program of research proposed in this project will lead to new advances that can speed discoveries in the laboratory to clinical practice in the treatment of cancer. I am also pleased that the project involves a significant collaboration leveraging the complementary expertise at our institution, Duke University, and the University of North Carolina at Chapel Hill, which together comprise an unparalleled resource for this sort of effort.

The College is pleased to contribute $17,500 per year for each of the five years of the award to the CQSB in support of the activities of this project. The project has my highest level of support. I look forward to hearing of the progress you make on your ambitious research program, and I and the College are happy to assist you I any way we can to ensure that in the goals of the project are achieved.

Sincerely,

Raymond E. Fornes, PhD
Associate Dean for Research, College of Physical and Mathematical Sciences

North Carolina State University is a land grant university and a constituent institution of The University of North Carolina

Office of Research and Graduate Studies

NC STATE UNIVERSITY

Office of the Vice Chancellor
Campus Box 7003
103 Holladay Hall
Raleigh, NC 27695-7003

919.515.2117
919.515.7521 (fax)

January 12, 2009

Marie Davidian, PhD
Program Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
Center for Quantitative Sciences in Biomedicine and
Department of Statistics
North Carolina State University
Raleigh, NC 27695

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Marie:

As Vice Chancellor for Research and Graduate Studies at North Carolina State University, I am pleased to offer my enthusiastic support for your inter-university grant application for a National Cancer Institute Program Project (P01), entitled "Statistical Methods for Cancer Clinical Trials." This exciting initiative, which will translate advances in basic biomedical science to clinical practice and develop new ways to conceive and evaluate treatment strategies for cancer, is consistent with the University's emphasis on health-related research. It is also an ideal endeavor in which to exploit the strengths of the Center for Quantitative Sciences in Biomedicine (CQSB), which the University strongly supports and which will serve as the administrative home for the project.

I am especially pleased that this project offers yet another opportunity for trans-institutional collaboration with Duke and the University of North Carolina at Chapel Hill. The resources at our three institutions for carrying out such a transformative project are unique and abundant, and integrating them in the way that you propose is certain to result in advances in clinical trial methods research that will have high visibility and impact.

To recognize the value of this project to the University's research mission and to assist you in achieving your ambitious research objectives, we will contribute $25,000 per year in each of the five years of the award to the CQSB in support of the activities of this project. I look forward to assisting you in any way I can to ensure the success of the project.

Sincerely,

Terri L. Lomax

Terri L. Lomax, PhD
Vice Chancellor for Research and Graduate Studies

TLL/mh

North Carolina State University is a land-grant university and a constituent institution of the University of North Carolina

College of Agriculture and Life Sciences
North Carolina Agricultural Research Service

NC STATE UNIVERSITY

Campus Box 7643
Raleigh, NC 27695-7643

919.515.2717
919.515.7745 (fax)
ag_research@ncsu.edu

January 9, 2009

Marie Davidian, PhD
Program Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
Center for Quantitative Sciences in Biomedicine and
Department of Statistics
North Carolina State University
Raleigh, NC 27695

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Marie:

I am pleased to offer my unqualified support of your grant application for a National Cancer Institute Program Project (P01) entitled "Statistical Methods for Cancer Clinical Trials." This exciting project fits well with our College's focus on the life sciences and health as well as with that of the Center for Quantitative Sciences in Biomedicine (CQSB), to which our College is strongly committed. I am especially enthusiastic about the opportunity this project represents for trans-institutional collaboration, which will draw on the complementary strengths of our institution, Duke, and the University of North Carolina at Chapel Hill, and I am pleased that the CQSB is a partner in this important effort.

In support of this transformative project, the College is pleased to commit $17,500 per year for each of the five years of the project to the CQSB in support of the activities of this project. The College and I pledge to assist you in any way possible to advance the goals of the project and contribute new innovations to cancer clinical trials methodology.

Sincerely,

Steven A. Lommel, PhD
Associate Dean for Research, College of Agriculture and Life Sciences

Cc: Ray Fornes
Mike Cross
Joy Martin
Gail Hill

North Carolina State University is a land-grant university and a constituent institution of The University of North Carolina

Sastry G. Pantula
Head, Department of Statistics
Director, Institute of Statistics
Campus Box 8203 / Patterson Hall
Raleigh, NC 27695-8203

College of Physical and Mathematical Sciences
College of Agriculture and Life Sciences

919.515.1949
919.515.7591 (fax)
pantula@stat.ncsu.edu
www.stat.ncsu.edu

NC STATE UNIVERSITY

January 14,2009

Marie Davidian, PhD
Program Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trial" Program Project
Center for Quantitative Sciences in Biomedicine and
Department of Statistics, North Carolina State University
Raleigh, NC 27695

Re: PAR-09-025, National Cancer Institute Program Project (P01) Applications

Dear Marie:

I am delighted to lend my enthusiastic support to your application for a National Cancer Institute Program Project (P01) Award, entitled "Statistical Methods for Cancer Clinical Trials," which will be a joint venture between North Carolina State University, Duke University, and the University of North Carolina at Chapel Hill and which will involve a number of faculty from the Department of Statistics. This project fits well with the Department's many activities related to the health sciences, including our popular graduate and research programs in biostatistics, bioinformatics, and biomathematics; our established relationship with Duke Clinical Research Institute through our joint training and graduate internship programs; and our recently minted relationship with the Department of Biostatistics at the University of North Carolina at Chapel Hill (UNC-CH) through your membership in the Biostatistics Core of the Translational and Clinical Sciences (TraCS) Center. I am especially excited about the opportunity the project represents for expanded and deeper collaboration among faculty in our Department, the Department of Biostatistics and Bioinformatics at Duke, and the Department of Biostatistics at UNC-CH. The project will be an important resource for all of our faculty, students, and postdocs, who will be exposed to the cutting-edge research on methodology for cancer clinical trials that you propose.

I am pleased that the both Colleges of Physical and Mathematical Sciences and Agriculture and Life Sciences, in which our Department jointly resides, have committed generous support to the project. In addition to the funds they have provided, I am pleased to commit space in the new Mathematics and Statistics Building, to which the Department will move in May 2009, to house the equipment that will host the project software repository as well as office and meeting space for project activities as needed.

The Department is eager to contribute to the success of this high-profile and important project in any way we can. Please do not hesitate to contact me if you need further resources. Also, as the President-Elect of the American Statistical Association, I am thrilled to see this proposal and its benefits to our profession and for human health.

Sincerely,

Sastry G. Pantula

Sastry G. Pantula

Duke University
School of Medicine

Nancy C. Andrews, M.D., Ph.D.
Dean, Duke University School of Medicine
Vice Chancellor for Academic Affairs

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As Dean of the Duke University School of Medicine and Vice Chancellor of Academic Affairs, I am writing to express my strong support for the Duke participation in the multi-institutional Program Project (P01) application to the National Cancer Institute entitled "Statistical Methods for Cancer Clinical Trials". The overall scientific goal of this project, to develop highly innovative methods for cancer clinical trials, is especially important in speeding the introduction of effective new therapies into practice, and is in line with the strategic research plans for the School of Medicine. The involvement of two other major universities in our region, the University of North Carolina and North Carolina State University, provides an outstanding opportunity for collaborative research.

Because of the importance and value of this program, The School of Medicine will contribute $15,000 per year towards this grant for each of the five years of the award.

In summary, the program project "Statistical Methods in Cancer Clinical Trials" has my enthusiastic support and commitment. I pledge to do whatever I can to see that we achieve the goals you have laid out.

Sincerely,

Nancy Andrews, M.D., Ph.D.
Dean, School of Medicine

BOX DUMC 2927
Durham, NC 27710
TEL 919.684.2455
FAX 919.684.0208
URL medschool.duke.edu
dukemedicine.org

DUKE UNIVERSITY MEDICAL CENTER
Department of Biostatistics and Bioinformatics

Telephone: (919) 684-9447
Facsimile (919) 681-7918

January 13, 2009

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As chair of the Department of Biostatistics and Bioinformatics, I am writing to express my enthusiastic support for the P01 application entitled "Statistical Methods for Cancer Clinical Trials", in which Duke will participate jointly with the University of North Carolina – Chapel Hill and North Carolina State University. The overall goal of the research, to transform the current paradigm for drug discovery and translation to clinic, resulting in improved survival and quality of life for cancer patients, is extremely important in itself. And the opportunity for our faculty to engage in high level collaborative research in statistical methodology is consistent with the strategic plans of our department.

Because of the importance of the research, I am willing to commit $15,000 per year of the grant for use in offsetting the costs of research. In addition, I will help in whatever other ways are needed to help this program succeed.

Sincerely,

Elizabeth R. DeLong, Ph.D.
Professor and Chair

Mail: DUMC 2721 • DURHAM, NORTH CAROLINA 27710
Courier: 11th Floor Hock Plaza • 2424 Erwin Road • Durham, NC 27705

DukeMedicine

Duke Comprehensive Cancer Center
A National Cancer Institute-designated Comprehensive Cancer Center

H. Kim Lyerly, MD
George Barth Geller Professor of Research in Cancer
Director
Duke Comprehensive Cancer Center

January 13, 2009

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As Director of the Duke Comprehensive Cancer Center, I am writing to express my strong support for your Program Project (P01) application entitled "Statistical Methods for Cancer Clinical Trials". The overall scientific goal of this project, to develop highly innovative methods for cancer clinical trials, is highly relevant to the strategic plans of the DCCC. Efficient statistical methods are extremely important in accelerating the development of anti-cancer therapy and in translating results into clinical practice. Developments from your proposed research program can be quickly implemented in cancer research projects here because of your role as the director of the biostatistics unit in the DCCC.

I am enthusiastic about this program and I pledge to help in whatever I can to see that you achieve the goals you have laid out.

Sincerely,

H. Kim Lyerly, M.D.
George Barth Geller Professor of Research in Cancer
Director, Duke Comprehensive Cancer Center

BOX DUMC 2714, Durham, NC 27710
LOC 2424 Erwin Road
Hock Plaza, Suite 601
Durham, NC 27705
TEL 919.684.5613
FAX 919.684.5653
EMAIL lyerl001@mc.duke.edu
URL www.cancer.duke.edu
dukemedicine.org

Cancer and Leukemia Group B
CENTRAL OFFICE OF THE CHAIRMAN

230 W. Monroe Street, Suite 2050
Chicago, IL 60606-4703

TEL (773) 702-9171
FAX (312) 345-0117

www.calgb.org

Richard L. Schilsky, M.D.
Chairman

January 14, 2009

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As chair of the Cancer and Leukemia Group B (CALGB), I am writing to express my enthusiastic support for your P01 application entitled "Statistical Methods for Cancer Clinical Trials". Indeed, the CALGB will be a major partner in this research through the participation of several clinical co-investigators participating from Duke and UNC, through the involvement of statisticians from the CALGB Statistical Center, which you direct as Group Statistician, and through the sharing of data from selected CALGB studies to illustrate the methods that are developed. The overall goal of the research, to transform the current paradigm for drug discovery and translation to practice, resulting in improved survival and quality of life for cancer patients, is a shared goal of the CALGB. For all of these reasons, it is anticipated that the results from this program can and will be implemented directly and immediately into the design and analysis of CALGB studies, to the benefit of all.

In summary, I enthusiastically support this program and look forward to our partnership in achieving its aims.

Sincerely,

Richard L. Schilsky, M.D.
Chair, Cancer and Leukemia Group B
Professor of Medicine
University of Chicago

NISS

National Institute of Statistical Sciences
PO Box 14006, Research Triangle Park, NC 27709–4006
Tel: 919.685.9300 FAX: 919.685.9310
www.niss.org

Alan F. Karr, Director
karr@niss.org

January 5, 2009

Dr. Michael Kosorok
Dr. Marie Davidian
Dr. Stephen George
Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg, CB 7420
Chapel Hill, North Carolina 27599–7420

Dear Michael, Marie, and Steve,

I am delighted to hear that you and your colleagues at Duke University, North Carolina State University, and the University of North Carolina at Chapel Hill are collaborating on a application for a joint Program Project grant from the National Cancer Institute on "Statistical Methods for Cancer Clinical Trials."

The Research Triangle is a natural setting for trans-institutional research projects such as the one you are proposing, and, as you know, the National Institute of Statistical Sciences (NISS) and the Statistical and Applied Mathematical Sciences Institute (SAMSI) have a long history of catalyzing and facilitating such interactions. NISS would be pleased to support this important initiative by making our centrally-located facilities available to you and other project personnel for meetings during the project period, as we have already done during the months leading up to the submission of your application. These facilities include "traditional" conference rooms, a fully-equipped video conference room and a lecture room and conference that support web streaming of events.

I wish you success with this proposal, and NISS looks forward to hosting activities associated with the project.

Sincerely,

Sincerely,

Colin B. Begg, PhD
Eugene W. Kettering Chair
Department of Epidemiology & Biostatistics

December 23, 2008

Michael R. Kosorok, PhD
Marie Davidian, PhD
Stephen L. George, PhD
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg Hall
CB 7420
Chapel Hill, NC 27599-7420

Re: Statistical Methods in Cancer Clinical Trials

Dear Michael, Marie and Steve,

I am writing to confirm my willingness to serve on the External Advisory Committee for your joint P01 Program Project application entitled "Statistical Methods in Cancer Clinical Trials". Although the key ingredients of clinical trial methodology have been established for many decades, the new drug development paradigm of trying to create new agents that are specifically targeted to the characteristics of relatively small subgroups of patients, with the ultimate goal of "personalized medicine", promises to change the landscape for designing and analyzing clinical trials. At this juncture we certainly need fresh, innovative approaches to maximize the efficiency of the drug development and testing strategies in the context of this paradigm. Statistical methods must play a central role in this effort. The group of investigators you have put together to tackle these difficult issues is impressive, drawing on the considerable strengths of your three institutions. Your team encompasses several prominent experts in both statistical theory and the application of clinical trials, and so you are in a great position to enhance our knowledge in this important area. I am very happy to serve on your External Advisory Committee, and generally to help in any way I can.

With best wishes,

Colin

Colin Begg, PhD
Attending Biostatistician
Chair, Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center
307 East 63rd Street, 3rd Floor, New York, New York 10065
Telephone 646.735.8108 • FAX 646.735.0009
E-mail: beggc@mskcc.org

NCI-designated Comprehensive Cancer Center

The University of Michigan
Department of Statistics

12/17/2008

Michael R. Kosorok, PhD, Marie Davidian, PhD, Stephen L. George, PhD
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg Hall
CB 7420
Chapel Hill, NC 27599-7420

Re: Statistical Methods in Cancer Clinical Trials

Dear Michael, Marie and Steve,

I am very happy to serve as chair of the External Advisory Committee for your joint Program Project "Statistical Methods in Cancer Clinical Trials" that you are submitting to the National Cancer Institute. This is an exiting project that will bring together the combined strengths of Duke University, North Carolina State University and the University of North Carolina at Chapel Hill.

I am keenly aware of the importance of clinical trials in the discovery of new treatments, and, as you know, have worked for many years in my own research on creating new clinical trial methods, especially in the areas of dynamic treatment regimes and reinforcement learning. I believe that the application of these new areas to cancer, as well as many of the other novel approaches proposed in your application, will likely have a large impact on public health.

As a member of the External Advisory Committee, I am looking forward to following your research progress and providing feedback on at least an annual basis. I wish you success in your application.

Sincerely,

Susan Murphy

H. E. Robbins Professor of Statistics
Research Professor, Institute for Social Research
Professor of Psychiatry

Address	*Telephone*	*Fax*	*email/URL*
Department of Statistics	734-647-3684	734-763-4676	samurphy@umich.edu
444D West Hall			http://www.stat.lsa.umich.edu/~samurphy
The University of Michigan			
Ann Arbor, MI 48109-1107			

THE SIDNEY KIMMEL COMPREHENSIVE CANCER CENTER AT JOHNS HOPKINS

CLINICAL TRIALS AND BIOMETRY

Oncology Biostatistics
550 North Broadway, Suite 1103
Baltimore, Maryland 21205-2013
Office (410) 955-4884 Fax (410) 955-0859
http://www.cancerbiostats.onc.jhmi.edu

Program Leader
Steven Goodman, MD, PhD
Professor, Acting Director
(410) 955-4596

Faculty
Giovanni Parmigiani, PhD
Professor, Director Bioinformatics Core
(410) 614-3426

Michael Ochs, PhD
Associate Professor
(410)955-8830

Peng Huang, PhD
Visiting Associate Professor
(410)502-0944

Jeanne Kowalski, PhD
Assistant Professor
(410) 955-4286

Leslie Cope, PhD
Assistant Professor
(410) 502-0945

Sarah Wheelan, MD, PhD
Assistant Professor
(410) 955-8841

Xiaobu Ye, MD, MS
Research Associate
(410) 614-6261

Senior Biostatisticians
Marianna Zahurak, MS
(410) 955-4219

Biostatisticians
Amanda Blackford, ScM
(410) 614-0361

Hua-Ling Tsai, MS
(410) 502-6529

Zhe Zhang, MS
(410) 502-0946

Administrative Staff
Helen Cromwell
Administrative Manager
(410) 955-4885

Alisa Moore
Administrative Coordinator
(410) 614-3432

December 18, 2008

Michael R. Kosorok, PhD
Marie Davidian, PhD
Stephen L. George, PhD
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg Hall
CB 7420
Chapel Hill, NC 27599-7420

Re: Statistical Methods in Cancer Clinical Trials

Dear Michael, Marie and Steve,

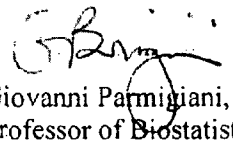
I am pleased to serve as a member of the External Advisory Committee for your joint Program Project "Statistical Methods in Cancer Clinical Trials" that you are submitting to the National Cancer Institute.

Many of your research aims involve genomics and Bayesian methods, areas of research in which I have been very active. Thus I am keenly aware of the potential these methods have in improving design and analysis of cancer clinical trials. I believe your proposed research in cancer clinical trials is fundamental and will likely have a large public health impact. This is an exciting project that will bring together the combined strengths of Duke University, North Carolina State University and the University of North Carolina at Chapel Hill.

NCI CCC

As a member of the External Advisory Committee, I am looking forward to following your research progress and providing feedback on at least an annual basis. I wish you success on your application.

Best wishes,



Giovanni Parmigiani, PhD
Professor of Biostatistics
Johns Hopkins University

20 RESOURCE SHARING PLAN(S)

The services of the Core A will be utilized by all of the projects in this program. Our estimate of the percentage utilization of the core by the five project are as follows:

Project 1 - 20%
2 - 20%
3 - 20%
4 - 20%
5 - 20%

The following is the external component of our resource sharing plan:

(a) Data Sharing Plan: The data-related resources generated by the program project consist of new statistical methodology, software packages for implementation of the methodology, and tutorials for the software. The statistical methodology will be shared through peer reviewed publications and national meetings and through other standard means. All accepted publications will be deposited in PubMed Central in accordance with the NIH Public Access Policy. Summaries of the methodology, the software and tutorials will be shared through a public web site managed by Core A, while Core C (Computational Resource and Dissemination Core) will prepare the software and tutorials for dissemination. Core C will only use de-identified data prepared by Core B (Data Compilation Core). Core A will not be involved in sharing of the data prepared in Core B; this function will be addressed by Core B.

(b) Sharing model organisms: N/A

(c) GWAS: N/A

CORE B

DATA COMPILATION CORE

Core Director: Stephen L. George, PhD

PROJECT SUMMARY (See instructions):

The Data Compilation Core (Core B) will develop and maintain a central resource of analysis-ready, annotated and documented data sets from clinical trials and related studies to be utilized by the investigators of the program. These data sets will be used to evaluate the methods developed in this program as well as to demonstrate the software developed in the Computational Resource Core (Core C). The primary source of the data will be the clinical trials and related studies of the Cancer and Leukemia Group B (CALGB), one of the major NCI-sponsored cancer cooperative groups. In addition, data from cancer research studies conducted at two large NCI-designated Comprehensive Cancer Centers, the Lineberger Comprehensive Cancer Center at UNC and the Duke Comprehensive Cancer Center, will also be utilized. This is a major advantage for the program in that the data sets provided can be exceptionally well annotated and documented, with the direct involvement of clinical and statistical scientists who were involved in the primary design and analysis of the studies.

RELEVANCE (See instructions):

A major disadvantage of using public data sets is that the investigator is often unable to understand the clinical and molecular data as the data are provided without appropriate documentation. Indeed, it is not possible to carry out a thorough statistical analysis of data from clinical trials without taking into account and understanding the design of the study, the specifics of the data collection process, the history of the study and the medical issues. This core will address these issues by providing analysis-ready data sets with extensive annotation and documentation.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location			
Organizational Name: The University of North Carolina at Chapel Hill			
DUNS: 608195277			
Street 1: Office of Sponsored Research, CB #1350	Street 2: 104 Airport Dr., Suite 2200		
City: Chapel Hill	County: Orange	State: NC	
Province:	Country: USA	Zip/Postal Code: 27599-1350	
Project/Performance Site Congressional Districts: NC04			

Additional Project/Performance Site Location			
Organizational Name: North Carolina State University			
DUNS: 042092122			
Street 1: Research Admin/ SPARCS	Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514		
City: Raleigh	County: Wake	State: NC	
Province:	Country: USA	Zip/Postal Code: 27695-7514	
Project/Performance Site Congressional Districts: NC-02			

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location		
Organizational Name: Duke University		
DUNS: 044387793		
Street 1: Hock Plaza	Street 2: Box 2716 Med Ct.	
City: Durham	County: Durham	State: NC
Province:	Country: USA	Zip/Postal Code: 27705
Project/Performance Site Congressional Districts: NC-004		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

Additional Project/Performance Site Location		
Organizational Name:		
DUNS:		
Street 1:	Street 2:	
City:	County:	State:
Province:	Country:	Zip/Postal Code:
Project/Performance Site Congressional Districts:		

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
George, Stephen L.	georg001	Duke University	Core Director
Chu, Haitao	Hchu11	UNC-CH	Core Co-Director
Davidian, Marie	Davidian2	NC State University	Core Co-Director
Ibrahim, Joseph G.	JOE_IBRAHIM	UNC-CH	Core Co-Director
Owzar, Kouros	KOWZAR	Duke University	Core Co-Director
Tsiatis, Anastasios	butch_tsiatis	NC State University	Core Co-Director

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project

Human Embryonic Stem Cells ☒ **No** ☐ **Yes**

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

METHODS AND SERVICES TO BE PROVIDED

1. SPECIFIC AIMS

The Data Compilation Core (Core B) will develop and maintain a central resource of analysis-ready, well annotated, and documented data sets from clinical trials and related studies to be utilized by the investigators of the program. Its mission will be to address the following specific aims in support of the program projects.

- **Identification and acquisition of data sets.** The primary data sources for this program will be the studies conducted by the Cancer and Leukemia Group B (CALGB) supplemented by studies conducted at the comprehensive cancer centers at Duke and UNC. As the CALGB Statistical Center is located at Duke (Dr. George is the Group Statistician) and the heads of the biostatistics units of the two cancer centers (Drs. George and Ibrahim) are participating in this program, we are especially well situated to identify appropriate data sets and to obtain them for the program researchers.
- **De-identification and anonymizing the data sets.** Data from clinical trials are subject to patient confidentiality regulations. Furthermore, because the data we will use are owned by external groups, it is necessary to get permission for use of the data for the purpose of research by project investigators. The owner of the data may also restrict the permission to a certain set of analyses. The core will follow all federal and local regulations and obtain the necessary clearances before processing any data. The core will also work on de-identifying and, when needed, anonymizing the data. For each data set, the core will develop a detailed data access plan which will outline the access levels for project investigators and will outline the specific analyses for which a data set may be used.
- **Annotation and documentation of data sets.** The core will carefully annotate and document the data sets so that project investigators can conduct more appropriate analyses and draw more informed conclusions. The core will also facilitate queries regarding the data sets on behalf of the project investigators.
- **Formatting of data sets.** One of the challenging aspects of analyzing data using statistical software is that the data are often not formatted appropriately. The core will provide data sets that are analysis-ready in the sense that they are merged and formatted so that they could be easily imported into statistical software.
- **Managing Information Technology (IT) resources.** The data sets have to be managed and made accessible using appropriate information technology systems. The core will acquire and maintain the appropriate information technology infrastructure to maintain and distribute the data set in an efficient, robust and secure manner.

2. BACKGROUND AND SIGNIFICANCE

This broad overall objective of this program is to develop and implement new methodology for the design and analysis of cancer clinical trials. In order to evaluate and illustrate these new methods, it is important to have access to analysis-ready, well annotated and documented data from actual clinical trials. The Data Compilation Core, henceforth referenced as Core B for brevity, will provide such data to the investigators in this program.

The Core B leadership is comprised of leaders in the field of clinical biostatistics and bioinformatics. Core B will be under the directorship of Dr. Stephen George. In addition to directing Core B and serving as a PD/PI for this program, Dr. George is the Group Statistician for the CALGB and directs the biostatistics units in the Duke Comprehensive Cancer Center (DCCC) and the Duke Translational Medicine Institute (DTMI). Because of these roles, he is well positioned to carry out the aims of this core in an efficient and effective manner.

In the day-to-day activities of the core, Dr. George will be assisted by the core co-Director Dr. Kouros Owzar. Dr. Owzar is an Assistant Professor in the Department of Biostatistics and Bioinformatics at the Duke University Medical Center. His research interests are in the areas of pharmacogenomics, survival analysis, statistical computing and, statistical dependence, specifically copulas. Dr. Owzar serves as the director of the

CALGB Bioinformatics Unit and the Director of the Radiation Countermeasures Center of Research Excellence Biostatistics and Computational Biology Core. Dr. Owzar serves as the chair of the Department of Biostatistics and Bioinformatics computing committee. Other co-Directors of Core B are Drs. Chu and Ibrahim at the University of North Carolina at Chapel Hill (UNC-CH) and Drs. Davidian and Tsiatis at North Carolina State University (NCSU).

We plan to partner with the CALGB Information Systems Unit to address the information technology needs for the core. Kimberly Johnson, Director of the CALGB Information Systems Unit, will participate in Core B to ensure the successful operation of the core. Ms. Johnson has over 25 years of information systems experience in research computing and clinical trials informatics. She is responsible for the overall direction of CALGB information systems and represents the CALGB to external entities by participating in several National Cancer Institute (NCI) and Cooperative Group information technology, most notably caBIG, initiatives. In the day-to-day information technology related activities, Ms. Johnson will be assisted by Amish Shah, who serves as the Deputy Director of the CALGB Information Systems Unit. More specifically, Mr. Shah will oversee the mentoring, training and supervision of the core Systems Administrator.

3. METHODS AND SERVICES TO BE PROVIDED

3.1 Data Services

3.1.1 Data Acquisition and Identification

Core B will facilitate the acquisition and identification of data sets for the project. The CALGB and Duke and UNC-CH cancer centers are expected to be the major contributors of data for the project. Data from the clinical trials discussed in the individual project narratives will be among the many clinical trial data sets compiled by the core for use by the investigators. As Dr. George is the Group Statistician for the CALGB Statistical Center and the Executive Director of the DCCC biostatistics unit; and Dr. Ibrahim is the Director of the UNC-CH Lineberger Comprehensive Cancer Center (LCCC) biostatistics unit, the process of identifying data will be greatly facilitated. Once the data have been identified and permission is granted to retrieve the data for the project, the process of acquiring the data will be facilitated, as Drs. George and Ibrahim know the processes for each group. Core B investigators are highly experienced in dealing with clinical, demographics and molecular data from clinical studies. This will be a major advantage in determining if a particular data set is appropriate for a given project and if so which data points should be requested.

We will now describe several specific examples of CALGB trials that play an important role in the proposed program project research. Pharmacogenomics, a topic studied in Project 4 (Methods for Pharmacogenomics and Individualized Therapy Trials), features extensively in CALGB trials such as protocols 80303 (pancreatic cancer), 40101 (breast cancer), and 90401 (prostate cancer). Similarly, collection of both time to death or relapse and longitudinal biomarkers such as prostate specific antigen (PSA) and measures such as quality of life (QOL), a topic studied in Projects 1 (Innovative Clinical Trial Design and Analysis) and 2 (Methods for Missing and Auxiliary Data in Clinical Trials), features extensively in CALGB trials such as protocol 90401 and 49907 (breast cancer). In addition, there is growing recognition that treatment of cancer is an ongoing process involving a sequence of treatment decisions, and the study of the entire sequence of treatment decisions requires a fundamentally new approach to design and analysis. This is the them of Project 5 (Methods for Discovery and Analysis of Dynamic Treatment Regimes), and the design of CALGB studies such as protocol 19808 in acute myelogenous leukemia, which involves a series of randomizations at treatment decision points, offers a promising starting point in this area.

Although clinical trials are the gold standard for evaluating the safety and efficacy of cancer therapies, less than 2% of patients with incident cancers enroll on NCI sponsored clinical trials. Furthermore, the fraction of trial enrollees is lower in racial/ethnic minority groups as well as older patients (Murthy, Krumholz, and Gross 2004). Those limitations have created gaps for the generalization of clinical trial findings to general populations. On the other side, the observational studies such as the Surveillance Epidemiology and End Results (SEER) and Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) datasets, discussed below, are more representative for the general population. By combining data from both clinical trials such as ACCENT (see below) and observational studies such as SEER and CanCORS, we will develop novel meta-analysis

methods as proposed in Project 3 (Methods for Post Marketing Surveillance and Comparative Effectiveness Research) to generalize clinical trial findings to general populations.

As examples of the types of data sets to be made available to investigators in this program, we describe briefly some data sets that will be particularly useful for the research of Project 3 involving Bayesian parametric and semiparametric models for meta-analysis of continuous or discrete data, longitudinal data, and time to event data.

SOS/ACORN. The SOS/ACORN data are a unique dataset containing extensive patient-reported outcomes, administrative claims data, and electronic medical records data. The ACORN data warehouse currently imports data from 11 community oncology sites in 10 states. Data are comprised of (1) Electronic Medical Record, (2) Claims/Billing Systems, and (3) Patient Care and Education System (PACE). The PACE system is an e/Tablet based education and assessment system that administers the Patient Care Monitor (PCM), a multi-dimensional, point of care symptom based patient reported outcome (PRO) measure. PCM 2.0 (previously known as the Cancer Care Monitor) is an e/Tablet based assessment of clinically relevant PROs administered to patients as part of routine care. Patients at the community oncology sites complete the PCM at regular medical visits prior to seeing health care professionals and before completing any visit-related laboratory work such as blood draws and scans. Patients answer questions on e/Tablets, and their responses are wirelessly relayed to their health care professionals. Patients provided written informed consent for their de-identified data to be used for research.

CanCORS. The NCI-funded Cancer Care Outcomes and Research Consortium (CanCORS) is comprised of 7 research groups with integrated health systems, VA medical centers, and population-based registries covering nearly 30 million people. The majority of enrollment in CanCORS occurred in 2003 through 2005, and as it pertains to this program, included approximately 4,000 adults over the age of 21 with adenocarcinoma of the colon or rectum, including an oversampling of African Americans and other minorities. Patient surveys, physician surveys, and extensive medical records abstraction comprise the data representing the care and outcomes of this population from 3 months prior to diagnosis through 15 months post diagnosis. Data elements reflect decision making about major treatments, patient-reported symptoms and outcomes, assessments of patient care, and financial burdens related to treatment. Many items and scales were obtained from published surveys on patient-reported outcomes, symptoms, and quality of life. The CanCORS Consortium is one of the most comprehensive observational studies ever undertaken to understand the experiences, treatments, and outcomes of patients with lung cancer or colorectal cancer in the United States.

SEER-Medicare and SEER. The data come from the Surveillance, Epidemiology and End Results (SEER) program of cancer registries linked to administrative and claims data from Medicare, the primary health insurer for 97% of the US population age 65 and older. The SEER Program is a surveillance system of population-based cancer registries spanning 14 states that systematically collects information on all newly diagnosed cancer cases that occur among the 26% of all persons residing in SEER areas. With a program standard of 98% complete case ascertainment within their regions, the registries and cancer cases within them are very closely representative of the US population. Medicare claims files include extensive diagnostic and treatment data for Medicare beneficiaries, including inpatient services for 97% of Americans age 65 and older, and outpatient and physician services for 93%. The SEER-Medicare data reflect the linkage between these registries and Medicare insurance claims, providing detailed information about elderly persons with cancer and offering a large population-based cohort that is one of few that can be used to longitudinally examine individuals from diagnosis through treatment and follow-up. As described by the Institute of Medicine, the SEER-Medicare data comprise a valuable and powerful resource for population-based examinations of cancer incidence, treatment patters, and care quality and outcomes.

PS-2/ACCENT. Recently, Drs. Goldberg and colleagues compiled individual patient data for 6,286 patients enrolled among 9 phase III randomized trials of chemotherapy for metastatic colorectal cancer. These data were complied in order to address whether patients who present for first-line colorectal cancer therapy with impaired function, as measured by a performance status of 2 (thus, PS 2), benefit from chemotherapy to the same extent as those with only minimal or no decline in function at the time of initiating therapy. As most trials enroll fewer than 10% PS 2 patients, and functional assessments are lacking from most population databases, a pooled analysis was thought to be the only way to answer this question. The analysis confirmed PS 2 patients benefit just as much as PS 0-1 patients do from chemotherapy in terms of progression-free survival, overall

survival, and response rate despite the markedly worse prognosis of PS 2 patients. Not surprisingly, however, toxicity was greater in PS 2 patients, including greater incidence of nausea and vomiting, and a 12% higher 60 day mortality. The database used to perform this analysis of outcomes by performance status is managed at the Mayo Clinic. Because clinical trial selection criteria limit inclusion to the healthiest of patients, the outcomes of patients treated outside the realm of protocol specified treatment is likely worse. Thus, this is a unique dataset, as it has over 500 PS 2 patients. This unique dataset will provide a minimum basic standard against which effectiveness of chemotherapy may be compared.

The ACCENT database is comprised of individual patient data from phase III clinical trials of adjuvant colon cancer therapy. Originally compiled to confirm the benefit of adjuvant FU, this database has now expanded to include data from 20,898 patients treated on trials between 1977 and 1999. This database has been used to demonstrate the excellent predictive ability of 3 year disease-free survival for 5 year overall survival in adjuvant colon cancer trials, leading to acceptance of 3 year DFS as an adequate endpoint for approval of adjuvant therapies by the FDA. Among other things, this database has also been used to explore racial differences in adjuvant therapy benefit for patients treated on clinical trials. Just as the PS 2 database represents a large number of patients with metastatic colorectal cancer, the ACCENT database serves the same purpose as a benchmark for our investigations of stage III colon cancer.

DEcIDE Network. The DEcIDE (Developing Evidence to Inform Decisions about Effectiveness) Network is a new network of research centers that AHRQ created in 2005 to generate new knowledge. The DEcIDE Network conducts accelerated practical studies about the outcomes, comparative clinical effectiveness, safety, and appropriateness of health care items and services. The network is comprised of 13 research-based health organizations with access to electronic health information databases and the capacity to conduct rapid turnaround research (http://effectivehealthcare.ahrq.gov/). Initial research focuses on the outcomes of prescription drug use and other interventions for which randomized controlled trials would not be feasible or timely, or would raise ethical concerns that are difficult to address. Other DEcIDE Network projects may focus on electronic registries, methods for analyzing health databases, and prospective observational or interventional studies. The DEcIDE research network of the Effective Health Care (EHC) program was developed to provide a variety of services and products to support the development of new scientific knowledge through studies on the outcomes of health care items and services.

3.1.2 Data Protection

Given that the data sets to be compiled by the core are obtained from clinical trials and other studies with individual patient records, adherence to patient data confidentiality and security is paramount. Furthermore, because the data are owned by external entities, such as the CALGB, it is important that the analyses are carried out within the framework of the agreements with the owner. To this end, the core will produce a data access plan for each data set. This plan will describe which individuals will have access to the data as well as any restrictions with respect to the usage of the data. We will always provide formal acknowledgment of the source of the data, in some cases, co-authorship for the clinical investigators who collected the data may be appropriate in publications by the project investigators using the data.

Drs. George and Owzar have extensive experience in data sharing issues both from regulatory and ownership points of view through their work with the CALGB and the DCCC. This experience will greatly facilitate the implementation of our data sharing policies and activities and interactions with regulatory bodies such as IRBs and the owners of the data. In fact, many of the existing policies and procedures used by the CALGB and the DCCC will be adopted by the core. These have been carefully developed over many years, in compliance with the changing landscape of data sharing in clinical research.

3.1.3 Data Documentation and Annotation

To conduct appropriate analyses and to draw informed conclusions from the results, one needs to understand the data from a scientific and clinical point of view. No level of technical rigor substitute for an understanding of the scientific question addressed by the study and the data collection methods.

To this end, the core will provide data annotation and documentation services. Specifically, each data point provided will be given a meaningful variable name along with a detailed description.

3.1.4 Data Formatting

For data to be useful, they must be properly formatted so that they can be readily imported into statistical software. In many cases, this requires merging multiple data sets together, often with very complicated queries. The data may need to be provided in long format, one row per observation, or wide format, one row per case. For single nucleotide polymorphism (SNP) data, this may require extracting genotype calls from millions of rows and creating a standardized file for import into analysis software. These large standardized files then need to be linked to the phenotypic data.

The core will work with individual project investigators to provide data in a format best suited for their projects. The core investigators have extensive experience in using the R statistical environment and SAS for data management. They also have extensive experience using Python and C/C++ languages for more complicated situations.

The core will employ a data programmer, who will report to Dr. Owzar and assume primary responsibility for formatting the data. The code written for formatting the data will be kept under source code management. For complicated situations, unit testing will be employed to protect against unintended consequences due to code modification.

3.2 Information Technology Infrastructure

The core will acquire, manage, and maintain its own dedicated hardware resources. Rather than building a completely new costly information technology infrastructure, for cost effectiveness the core will partner with the CALGB Information Systems (IS) unit at Duke University. This partnership will enable the core to house its server in a professional grade server room. Rather than purchasing a complete backup system, the core will purchase extensions to the existing CALGB backup system to support the archiving needs of the core. The cost effectiveness is not limited to the cost of hardware acquisition and housing. The CALGB IS unit also has extensive experience in how to retrieve, store, archive and manage clinical and correlative science data using state of the art methods within a secure environment. The core will be able to harness this valuable experience. Specifically, the core will collaborate with Ms. Johnson and Mr. Shah to ensure appropriate support in this partnership.

To protect sensitive confidential information, applications incorporate sound security practices and comply with guidelines for Health Insurance Portability and Accountability Act (HIPAA). Web applications safeguard protected health information (PHI) by requiring secure logins, limiting access to authorized users, and incorporating timeouts for inactivity within an open application. Applications will encrypt all data transmissions to ensure security and confidentiality of data as it is entered and viewed. Users are provided accounts and roles that determine their access to data or functions.

The core will employ best practices for IT. For example, the core will employ a source code management system to keep track of the development of its code produced for data formatting and the software for its hardware systems. As another example, the core will employ unit testing to ensure that modifications to its code do not have unintended consequences such as regressions. The CALGB IS Unit has a comprehensive set of policies and procedures in place to ensure adherence to these practices. The core will adopt these policies and procedures.

3.2.1 Data Server

The core will acquire and maintain a file-server to host and disseminate the data sets. This system will be referred to as the project data server. Since the core will make both raw and pre-processed genomic data available to the project investigators, a file server with large disk space will be acquired for this purpose. To provide some level of redundancy against hard-drive failure, the file server will have hardware RAID with battery backup. To reduce bottlenecks induced when processing or accessing large volumes of data, the server will be equipped with multi-core processors, large amount of dynamic memory and a high-end multi-port network card.

The file server will be housed in the state-of-the-art CALGB server room, which includes raised flooring, automated fire detection and suppression systems, an uninterruptible power supply, key card security systems, and a separate cooling system. The server will be connected to the Duke University high-speed network link to allow high bandwidth access from inside and outside of Duke.

The data sets will be posted on the data server using a web-based content management system (CMS). This will allow project investigators to download the data using a secure interface. For each data set, a dedicated page will be created within the content management system. Access and permissions on this page will be customized according to the data access plan. Along with the data sets, the data access plan and data documentation material will be posted on the page. Possible choices for the CMS used by Core B are plone, alfresco, MediaWiki or xwiki. All of these systems are under active development. The final choice will be made at the time when the grant is funded after a careful comparative review of the available systems to determine which system is best suited for the needs of the core. The CALGB IS unit has extensive experience in installing, deploying, and customizing these types of systems. The CALGB protocol for file and web server management and security will be followed.

As an example of the experience of the Core B investigators, Dr. Owzar and his colleagues from the CALGB Bioinformatics and IS units are currently developing a content management system using the Python web framework called Django, a modern object-oriented infrastructure designed for building web applications using the Model-View-Controller paradigm. The concept of a web framework is that of a modular system, where different components can be replaced, as long as the whole conforms to the underlying Application Programming Interface (API). This has the advantage of scalability over more traditional systems. In particular, Django has a transparent database API and built-in support for file upload. For performance reasons, Django stores files on the file system rather than in the database. However, a reference to the location of the file on the file system, along with file metadata, is stored in a database. We believe this is a good choice, particularly in the context of large files like those found in bioinformatics. The alternative, storing files in the database, is problematic for large files because of database size limitations. This feature will enable seamless integration with other systems, as files can be extracted using database queries or copied directly off the file-system. Another important feature of this type of setup is the ability to integrate with other applications, including those developed through the caBIG project, via the API. To assess the viability of the proposed web application, Dr. Owzar and his colleagues have developed a prototype that implements secure authentication, three upload modes (single-file, multiple-file in zipped archive and multiple-file using a point-and-click graphical interface), and remote data capture through database queries. The results of the prototype indicate the proposed approach is feasible and further development is warranted. Given that the core expects to host a considerable amount of large genomics data, the core will evaluate the use of this system as a supplement to its CMS.

3.2.2 Backup

Core B is tasked with the stewardship of the data. Although the RAID array of the file server will provide some level of redundancy, it will not be relied on as a backup system to protect against catastrophic hardware failure, catastrophic software corruption, and unrecoverable user errors. The core will employ a tape-based system for archiving the data sets. The tape media will be stored remotely. The CALGB protocol for backups will be followed.

3.2.3 Secure E-mail System

The core will provide a secure e-mail system for transmission of sensitive and protected data. A typical example for which such a system is needed is if one has to query patient data for a given sample ID. The commercial enterprise system used by the CALGB IS Unit will be employed for this purpose.

3.2.4 Issue Tracking System

The core will implement an electronic issue tracking system to allow project investigators to report issues and request support. This will allow us to track the progress of the support requests and allow the leadership to

assess the responsiveness of the core staff to issues. We will employ the CALGB JIRA issue tracking system for this purpose.

3.2.5 Mailing List

The core will maintain a mailing list to transmit communications efficiently to the project investigators and personnel. The mailing membership list will be actively maintained to ensure that new members are added in an expeditious manner and that individuals no longer involved with the projects are removed. The CALGB mailing list system will be employed.

4. INCLUSION ENROLLMENT REPORT

N/A

5. BIBLIOGRAPHY AND REFERENCES CITED

N/A

6. PROTECTION OF HUMAN SUBJECTS

Core B will receive and manage information from human subjects. It will, however, not directly solicit information from patients. Data received, maintained and analyzed by the core will fall under HIPAA regulations. Thus, all procedures for handling data will be created to adhere to these regulations. The core will manage IRB approvals required for receiving and managing the data. Patient confidentiality and data security for electronic data are of paramount importance. Data security is ensured through physical security and backup of data, separating patient identifiers from confidential data, and preventing unauthorized access to data. Unauthorized access to data is prevented through data encryption "across the wire", restricted access to the server, and limitations on data availability as required. The core will not provide data to anyone who has not been specifically authorized to have access and will generally provide only de-indentified data. In cases where identifiable protected health information is required by thr nature of the research, we will provide limited data sets after all appropriate IRB requirements are met and approval has been granted by the IRB. Drs. George and Owzar have experience in applying data sharing policies of the CALGB and of Duke.

7. INCLUSION OF WOMEN AND MINORITIES

The data sets created and maintained by Core B will often have information on gender, ethnicity and race, but we are not the investigators who generated the original data. However, we will be able to report on the distribution of these variables in the data sets, particularly women and minorities.

8. TARGETED/PLANNED ENROLLMENT TABLE

N/A

9. INCLUSION OF CHILDREN

Although most of the studies for which we will be preparing data sets are from studies in adults, it is possible that we will also handle some data from studies involving children. These data will be handled in the same fashion and we will be able to report on the age distribution in the data sets.

10. VERTEBRATE ANIMALS

N/A

11. SELECT AGENT RESEARCH

N/A

12. MULTIPLE PD/PI LEADERSHIP PLAN

N/A

13. CONSORTIUM/CONTRACTUAL ARRANGEMENTS

If the present application is funded, the University of North Carolina at Chapel Hill will execute subcontracts with the consortium institutions (Duke University and North Carolina State University). These inter-institutional agreements will be written consistent with the NIH consortium agreement policy.

14. LETTERS OF SUPPORT

- H. Shelton Earp, MD, Director of the UNC-CH Lineberger Comprehensive Cancer Center, Providing institutional support.
- H. Kim Lyerly, MD, Director of the Duke Comprehensive Cancer Center, Providing institutional support.
- Richard L. Schilsky, MD, Chair of Cancer and Leukemia Group B, Providing institutional support.

UNC
LINEBERGER COMPREHENSIVE
CANCER CENTER
N.C. CANCER HOSPITAL

December 19, 2008

Michael R. Kosorok, PhD
Lead Principal Director/Principal Investigator
"Statistical Methods for Cancer Clinical Trials" Program Project
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Michael:

The strong partnership between UNC Lineberger Cancer Center and the Department of Biostatistics is decades old but the events of the last five years have taken it to another level. Recruitment has assembled an extraordinary group of statistical methodologic researchers, including Joe Ibrahim, Danyu Lin, Fred Wright, and Jason Fine, who contribute substantially to our cancer research. Under your leadership, an impressive era of scientific productivity is underway that will measurably improve clinical trials methodology.

As Director of the UNC Lineberger Comprehensive Cancer Center, I am writing to express my great enthusiasm and support for your application for an NCI Program Project (P01) entitled "Statistical Methods for Cancer Clinical Trials" at the University of North Carolina at Chapel Hill. This is certainly an important initiative for the UNC Lineberger, the Gillings School of Global Public Health, and for the University of North Carolina at Chapel Hill. With all of the recent advances in biomedicine, there remains a serious bottleneck between laboratory discoveries and their utilization in clinical practice. New clinical trials methodology is needed to keep abreast of and take advantage of molecular genetic discovery. I believe that the innovative program you have outlined will make important breakthroughs in solving this fundamental problem and have broad applicability for breast, colon and lung cancer as well as for other cancers and other diseases. I am very supportive of you and your research group utilizing existing clinical trial data sets housed in the Lineberger Comprehensive Cancer Center.

An important aspect of this program project is the collaboration with North Carolina State University and Duke University. This brings together a diverse group of investigators not only in biostatistics but also in medical oncology, health policy, pharmacogenomics, and computer science.

In summary, your program project application has my highest level of support and commitment. I will do all that I can to help you and your colleagues achieve the goals of this forward-looking P01 and, in the process, to help UNC become a leader in the field of cancer clinical trials.

Sincerely yours,

H. Shelton Earp III, MD
Director and Lineberger Professor
Professor of Medicine and Pharmacology

DukeMedicine

Duke Comprehensive Cancer Center
A National Cancer Institute-designated Comprehensive Cancer Center

H. Kim Lyerly, MD
George Barth Geller Professor of Research in Cancer
Director
Duke Comprehensive Cancer Center

January 13, 2009

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As Director of the Duke Comprehensive Cancer Center, I am writing to express my strong support for your Program Project (P01) application entitled "Statistical Methods for Cancer Clinical Trials". The overall scientific goal of this project, to develop highly innovative methods for cancer clinical trials, is highly relevant to the strategic plans of the DCCC. Efficient statistical methods are extremely important in accelerating the development of anti-cancer therapy and in translating results into clinical practice. Developments from your proposed research program can be quickly implemented in cancer research projects here because of your role as the director of the biostatistics unit in the DCCC.

I am enthusiastic about this program and I pledge to help in whatever I can to see that you achieve the goals you have laid out.

Sincerely,

H. Kim Lyerly, M.D.
George Barth Geller Professor of Research in Cancer
Director, Duke Comprehensive Cancer Center

BOX DUMC 2714, Durham, NC 27710
LOC 2424 Erwin Road
Hock Plaza, Suite 601
Durham, NC 27705
TEL 919.684.5613
FAX 919.684.5653
EMAIL lyerl001@mc.duke.edu
URL www.cancer.duke.edu
dukemedicine.org

Cancer and Leukemia Group B
CENTRAL OFFICE OF THE CHAIRMAN

230 W. Monroe Street, Suite 2050
Chicago, IL 60606-4703

TEL (773) 702-9171
FAX (312) 345-0117

www.calgb.org

Richard L. Schilsky, M.D.
Chairman

January 14, 2009

Stephen L George, Ph.D.
Co-Director/Co-PI
"Statistical Methods for Cancer Clinical Trials" Program Project
Duke University School of Medicine
Durham, NC 27705-3833

Re: PAR-09-025 – National Cancer Institute Program Project (P01) Applications

Dear Steve:

As chair of the Cancer and Leukemia Group B (CALGB), I am writing to express my enthusiastic support for your P01 application entitled "Statistical Methods for Cancer Clinical Trials". Indeed, the CALGB will be a major partner in this research through the participation of several clinical co-investigators participating from Duke and UNC, through the involvement of statisticians from the CALGB Statistical Center, which you direct as Group Statistician, and through the sharing of data from selected CALGB studies to illustrate the methods that are developed. The overall goal of the research, to transform the current paradigm for drug discovery and translation to practice, resulting in improved survival and quality of life for cancer patients, is a shared goal of the CALGB. For all of these reasons, it is anticipated that the results from this program can and will be implemented directly and immediately into the design and analysis of CALGB studies, to the benefit of all.

In summary, I enthusiastically support this program and look forward to our partnership in achieving its aims.

Sincerely,

Richard L. Schilsky, M.D.
Chair, Cancer and Leukemia Group B
Professor of Medicine
University of Chicago

15. RESOURCE SHARING PLAN(S)

The services of the Core B will be utilized by all of the projects in this program. Our estimate of the percentage utilization of the core by the five project are as follows:

Project 1 - 30%
2 - 20%
3 - 10%
4 - 30%
5 - 10%

The following is the external component of our resource sharing plan:

(a) Data sharing plan: The de-identified data sets generated in core B will be available to all investigators in this program for addressing the specific aims of the program. However, since this program did not generate the original data, the owners of the original data must agree to any wider use or distribution of the data sets. In case wider use or distribution of a particular data set is appropriate, a prior written data use agreement with the owner(s) of the original data will be required. For example, Core B will assist in preparing a few simple data sets which will be used in the tutorials that are developed in Core C to assist with training users of software developed in this program project.

(b) Sharing model organisms: N/A

(c) GWAS: N/A

CORE C

COMPUTATIONAL RESOURCE AND DISSEMINATION CORE

Core Director: Marie Davidian, PhD

PROJECT SUMMARY (See instructions):

The overall scientific goal of this ambitious Program Project is to develop innovative statistical methods for cancer clinical trials that can help to hasten successful introduction of effective new therapies into practice. The Computational Resource and Dissemination Core (Core C) will carry out several critical functions related to the implementation and dissemination of the statistical methods for the design and analysis of cancer clinical trials developed in the five research projects. The Core will be tasked with developing, in close collaboration with project investigators, efficient, robust code implementing the statistical methods that can be used for evaluation of the methods in extensive simulation studies and for application of the methods to data compiled by Core B and from other sources. The Core will also be tasked with leading and facilitating, in close collaboration with project investigators, development of robust, reliable, user-friendly, and well-documented software applications suitable for public dissemination to practitioners involved in the design and analysis of cancer clinical trials. Core C will adopt best practices for these tasks and provide the necessary information technology and educational infrastructure to disseminate these applications.

RELEVANCE (See instructions):

Before the new statistical methods for design and analysis of cancer clinical trials to be developed in this Program Project can be adopted for use in cancer research, they must be tested and evaluated, and they must be implemented in user-friendly software accessible to practitioners. Core C will collaborate closely with project investigators to facilitate these efforts.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location			
Organizational Name: The University of North Carolina at Chapel Hill			
DUNS: 608195277			
Street 1: Office of Sponsored Research, CB #1350		Street 2: 104 Airport Dr., Suite 2200	
City: Chapel Hill	County: Orange		State: NC
Province:	Country: USA		Zip/Postal Code: 27599-1350
Project/Performance Site Congressional Districts: NC04			

Additional Project/Performance Site Location			
Organizational Name: North Carolina State University			
DUNS: 042092122			
Street 1: Research Admin/ SPARCS		Street 2: 2701 Sullivan Dr.,Admin Serv III, Box 7514	
City: Raleigh	County: Wake		State: NC
Province:	Country: USA		Zip/Postal Code: 27695-7514
Project/Performance Site Congressional Districts: NC-02			

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location

Organizational Name: Duke University

DUNS: 044387793

Street 1: Hock Plaza | Street 2: Box 2716 Med Ct.

City: Durham | County: Durham | State: NC

Province: | Country: USA | Zip/Postal Code: 27705

Project/Performance Site Congressional Districts: NC-004

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Davidian, Marie	davidian2	NC State University	Core Director
Lin, Danyu	DANYU_LIN	UNC-CH	Core Co-Director
Owzar, Kouros	KOWZAR	Duke University	Core Co-Director

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project

Human Embryonic Stem Cells ☒ No ☐ Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

METHODS AND SERVICES TO BE PROVIDED

1 SPECIFIC AIMS

Statistical methods for design and analysis of clinical trials are fundamental tools in cancer research. Before any statistical method can be advocated for use in practice, it must be evaluated to ensure that it yields reliable conclusions over the range of conditions under which it is likely to be employed. Moreover, in order that new, tested methods be adopted by the research community, they must be made available in accessible, friendly software, and demonstrations of their use in applications must be provided. The mission of the Computational Resource and Dissemination Core (Core C) will be to address the following specific aims in support of the goal of this program project to provide new statistical methods for a host of data-analytic challenges in cancer research:

- **Aim 1: Implementation of project methodology.** Core directors and staff will collaborate with project investigators to develop efficient, robust, and reliable code implementing the new methods for use in testing and evaluation of the methods and in analysis of data when such professional development is required.
- **Aim 2: Computational resources in support of project methodology.** Certain generic computational tasks and methods will be required and used by multiple projects. Core directors and staff will collaborate with project investigators to identify such tasks and methods, and core staff will develop multi-use routines and modules for use by project investigators and core programmers. The core will also assist investigators with computational issues, such as development of code embedding multiple languages.
- **Aim 3: Creation and dissemination of public-use software and project methodology.** Core investigators will, in collaboration with the project investigators, develop well-tested, reliable, user-friendly, and well-documented software applications to bring the methodologies developed by the program projects into practical use. Once an application in development is deemed to be mature enough for dissemination to the general user community, the core will post it on the program project public web interface. Each application will have its own page, where a summary of the methodology and the application are provided. The corresponding documentation along with technical references will be posted as well. Core C will also collaborate with project investigators to arrange workshops to demonstrate and educate the user community about the methodologies developed by the projects, including showcasing the applications.

For all Aims, the core will conduct application evaluation, development, and deployment according to best practices and employ a systematic software development process. This includes use of a source code management system to keep track of and document modifications to code, and implementing unit testing for identifying regressions and evaluating the outcome for extreme or pathological test cases. Core C will also provide tools for systematic production of documentation, employ an issue tracking system to keep track of problems and requests reported by the users of the code, and maintain mailing lists to provide updates to the users.

In the sequel, the Computational Resource and Dissemination Core is referred to for brevity as Core C.

2 BACKGROUND AND SIGNIFICANCE

2.1 Background

The overall scientific goal of this ambitious program project is to develop innovative statistical methods for cancer clinical trials that can help hasten successful introduction of effective new therapies into practice. Before the new methods can be recommended for use toward this objective in cancer research, several critical issues must be addressed. First, potential users, such as biostatisticians collaborating on the design and analysis of cancer trials, must be assured that the methods are reliable; e.g., when used to conduct a hypothesis test at level of significance 0.05, a new procedure does in fact reject the hypothesis of no effect at most 5% of the time when it is really true. Second, the methods must be made accessible. Scientists involved in the design and analysis of clinical studies will not employ new methods if to do so is difficult or, worse, requires them to implement the methods themselves. If the new methods developed by the program project are to enjoy widespread adoption,

software implementing them that is straightforward to use, well-documented, and well-tested must be provided. Finally, the utility of new methods in practice must be convincingly demonstrated to potential users in applications for which they were designed through illustrative and detailed case studies that exemplify their use. We now elaborate on each these points.

A statistical method is of no practical use unless it has acceptable *operating characteristics*. For example, for a test statistic, one needs to determine its sampling distribution, which characterizes how values of the statistic vary over all possible data sets of the same size that could have resulted from conducting a study, which is required to calculate p-values. As discussed above, one then needs assurance that the p-value so calculated is reliable. When interest focuses on obtaining an estimate of a quantity such as a treatment effect (e.g., difference in means, odds ratio, hazard ratio), one would like to trust that the estimator used performs well in general; e.g., does not exhibit unacceptable bias when targeting the true value of the quantity of interest. One also requires an estimate of sampling variability of the estimator to calculate standard errors and confidence intervals. Here, for example, one would expect that a 95% confidence interval for the true value of the quantity being estimated, calculated using the method, would contain the true value across all possible data sets 95% of the time. In summary, that a statistical method has sound operating characteristics, including negligible bias, sampling variability that is not too large that can be characterized accurately, and advertised test and interval performance, is a requirement for its widespread use.

In all but the most trivial cases, the operating characteristics of a statistical method cannot be derived exactly by straightforward analytical calculations as they can for the simple methods introduced in introductory statistics courses. Rather, both for methods that are in routine use, such as the logrank test for survival analysis, and for new methods like those being developed in this program project, the usual approach is to appeal to statistical theory, from which approximations to the behavior of the methods, e.g., a description of sampling distribution, can be derived. Such so-called asymptotic approximations are obtained under the condition that the sample size available in a study is infinitely large, but are then used to calculate standard errors, confidence intervals, and p-values as if the approximation were exact. There is thus no guarantee that the approximation will lead to acceptable operating characteristics in practice with finite sample sizes. Moreover, the theory does not establish how large the sample size needs to be in order that the approximation be reasonable.

In the development of new statistical methods, it has thus become routine, if not required, to carry out so-called *simulation studies* to evaluate how well the asymptotic theory can approximate the true operating characteristics of a method in sample sizes like those encountered in practice and to determine how large the sample size needs to be in order to use the approximations with confidence. Similarly, in many cases, there may be more than one method available, e.g., a new method has been proposed that is purported to improve over an existing one. Here, a comparison of properties, such as unbiasedness, sampling variability, and efficiency, a measure derived from these features, is then called for. Although it may be possible to base the comparison on the asymptotic theory, again, how the methods compare in realistic sample sizes is of paramount interest, and simulation studies are the primary mechanism for such evaluations.

A simulation study is an empirical, computational investigation of performance of statistical methods that is designed to reflect conditions actually encountered in practice. Typically, random samples of data are generated according to a mechanism that has been chosen to resemble conditions in actual studies, including sample size; variability of outcomes across subjects; true values of quantities of interest such as proportions of responders, hazard ratios, and degrees of censoring; and other factors. Ideally, a method should be robust, that is, yield acceptable operating characteristics over a range of such conditions. Accordingly, new methods are ordinarily subject to an extensive battery of simulation scenarios to demonstrate robustness, and to establish under what conditions the asymptotic approximations are reasonable and the new method outperforms existing ones.

Fortunately, with the advent of affordable high-performance computing (HPC) resources, large-scale, computationally intensive empirical studies to investigate the operating characteristics of complex statistical methods under a wide array of scenarios are readily facilitated. Such studies are considered an essential component of research developing new statistical methods; indeed most biostatistical journals ordinarily demand that articles proposing new methods include a report of the results of relevant simulation studies.

In addition to operating characteristics, a second issue in the development of new, complex statistical methods is their implementation itself. For example, calculation of an estimator for a quantity of interest may involve an iterative algorithm embedding a complex series of manipulations that require the use of sophisticated optimization techniques and numerical approximations, and hence may be computationally intensive and sensitive

to conditions such as sample size and anomalies in data. The speed with which such algorithms "converge" to a result and the extent to which they yield acceptable results are critical issues. Simulation studies are also used to evaluate the performance and speed of the numerical algorithms under different scenarios.

Once a method has been implemented and thoroughly evaluated, it must be made accessible to the user community in order to enjoy successful adoption in practice. To this end, a software application must be developed. An application is unlikely to be adopted for routine use if it fails to be perceived as reliable and well-tested. It is also unlikely to be adopted if it is not straightforward to use or lacks proper documentation.

Often, biostatistical researchers implement methods they have developed for their own purposes; e.g., to facilitate simulation studies of their performance and illustrative analyses demonstrating their use to be reported in journal articles. However, although they may make the associated code publicly available, the code is often not sufficiently general, is poorly documented, or is inefficient because it was not developed with widespread use in mind. Many researchers are admittedly not trained as programmers and do not have the expertise or time to adapt their code so that it is broadly accessible; accordingly, it is unlikely that their methods will be widely used. Professional programmers, who have specific expertise in the development of efficient and robust code and in the creation of user interfaces and documentation that facilitate intuitive and straightforward use of the methods, are best equipped to develop software that is suitable for public dissemination.

2.2 Overview

Core C will be tasked with leading and facilitating, in collaboration with the project investigators, efforts to address all of the foregoing issues through three key functions, described in greater detail in Section **4**. As noted at the outset, Core C will assist project investigators with the development and testing of efficient, robust code implementing project methods when required that can be used to facilitate simulation studies to evaluate operating characteristics and numerical performance of methods and data analyses. Core C will also provide resources such as multi-use generic routines implementing tasks and methods that are used by several projects and assistance with coding issues. Finally, Core C will develop software for public dissemination to the user community.

In carrying out these functions, the core will adopt best practices for its processes to ensure that the code, simulations and case studies, and software applications are developed and implemented in a professional fashion. A source code management system will be employed to manage and annotate the software projects. The applications will be tested under normal and pathological conditions and unit testing will be implemented to ensure that the applications are robust and reliable. Automated documentation systems will be used throughout the development process. The intellectual output of the core will be stored, disseminated and protected using state-of-the-art information technology resources. The computing systems available to the investigators have multiple processing cores ranging from two cores to a few thousand. The core will implement procedures for harnessing the power of these multi-core system which make it possible to spread the work over several cores whereby greatly reducing the time needed to complete the task. The applications will be built using open source software as much as possible, not only to provide greater flexibility to the developers, as the underlying code can be customized to ones need, but also to eliminate licensing issues for users. Implementations of some project methodology in SAS will also be developed as deemed necessary. Applications will also be built, to the extent possible, using cross-platform software to ensure that they can be run by Linux, Windows and MacOS users. Software will be packaged with installer routines to facilitate easy set-up by a variety of users. Further details are presented in Section **4**.

3 CORE PERSONNEL AND STRUCTURE

3.1 Core Leadership and Experience

The Core C leadership is comprised of leaders in the field of clinical biostatistics from all three institutions with wide-ranging expertise and keen interest in statistical computing and the development of software applications for statistical methodology. This highly qualified group of individuals will oversee the activities of the core Staff, who are profiled below, and facilitate interactions between the core and project investigators.

Marie Davidian, PhD, will serve as core Director. Dr. Davidian is William Neal Reynolds Professor of Statistics at North Carolina State University (NCSU) and will also serve as the NCSU Program Director/Principal Investigator (PD/PI) for the program project. Dr. Davidian is internationally known for her research on methods for longitudinal data analysis, design and analysis of clinical trials, and methods for causal inference and handling missing data. Dr. Davidian will be responsible for overall administration of core activities and for overseeing core efforts associated with the core Aim 1 related to shared computational resources.

Dr. Davidian has considerable experience in the design and execution of large scale simulation studies involving complex methods and data generating mechanisms, including a current effort in another project sponsored by NIAID involving simulations of within-host HIV infection dynamics across a population treated according to different adaptive treatment strategies in connection with the design of treatment strategies and clinical trials. She has developed and taught a course discussing design and execution of simulation studies for PhD students at NCSU. Dr. Davidian was one of the invited participants at a National Cancer Institute (NCI) workshop on "Barriers to Producing Well-Tested, User-Friendly Software for Cutting-Edge Statistical Methodology" held in May 2008; see below. As a PD/PI for the overall program project, she will be well-positioned to ensure that the core is responsive and adapts to needs and issues of the entire project as they arise. Her experience in this regard makes her well-qualified to oversee these efforts.

Danyu Lin, PhD, will serve as core co-Director. Dr. Lin is the Dennis Gillings Distinguished Professor of Biostatistics at the University of North Carolina at Chapel Hill (UNC-CH) and a co-Director of the overall program project. Dr. Lin is an internationally recognized biostatistician who has made numerous methodological contributions to the designs and analysis of clinical trials and genetic studies. He will oversee the software development and dissemination activities of the core associated with core Aim 3.

Dr. Lin has developed and published computer programs implementing the statistical methods he has developed. Several of his methods are available in commercial statistical software packages, including SAS, S-Plus, and Stata, and widely used by a variety of researchers. He has served as a consultant to SAS; Insightful, Inc.; and Stata, Inc. Dr. Lin has posted several computer programs for his recent methods on genetic data analysis at his website, and his software interface HAPSTAT, which was developed by a professional computer programmer under Dr. Lin's supervision, has been downloaded by more than 100 researchers and used in several genetic association studies. His more recent software interface SNPMStat is also very popular. Computer programs implementing many of his recent methods for survival analysis and clinical trials are posted at the website of his main collaborator, Donglin Zeng. Dr. Lin is a strong advocate of developing software for new statistical methods and was one of the 6 invited speakers at the NCI workshop on "Barriers to Producing Well-Tested, User-Friendly Software for Cutting-Edge Statistical Methodology" noted above. This experience makes Dr. Lin highly qualified to lead the application development and dissemination efforts.

Kouros Owzar, PhD, will serve as core co-Director. Dr. Owzar is Assistant Professor in the Department of Biostatistics and Bioinformatics at the Duke University Medical Center. His research interests are in the areas of pharmacogenomics; survival analysis; statistical computing; and statistical dependence, specifically copulas. Dr. Owzar will oversee the implementation efforts of core Aim 1.

He currently serves as the director of the Cancer and Leukemia Group B (CALGB) Bioinformatics Unit, director of the Radiation Countermeasures Center of Research Excellence Biostatistics and Computational Biology Core, and as chair of computing for the Department of Biostatistics and Bioinformatics. . Dr. Owzar is a developer of the `distR` package which is available at `http://r-forge.r-project.org/projects/distr/`. This package provides a conceptual treatment of random variables by means of R S4-classes. Other packages related to `distR` are `distTest`, `distSim` and `distMod`. These can be used to provide conceptual treatments to testing, simulation and modeling. All of these packages will be extensive used by Core C. Dr. Owzar and his CALGB colleagues are also in the process of developing an R package for analysis of SNP and microarray data. These activities make Dr. Owzar very well-qualified to lead the implementation efforts.

The core leadership will meet weekly by conference call or face-to-face if deemed necessary at one the institutions or at the National Institute of Statistical Sciences (NISS) building in Research Triangle Park. Core leaders will discuss investigator proposals for core assistance, as outline in Section **4**, performance and progress of the programmers (see below), and other issues related to the core functions and activities. Core leadership will also participate in meetings with the core Staff, as discussed below.

3.2 Core Staff

Core Programmmers. Core C will recruit three programmers, one at each institution, to support the tasks associated with each of the core Aims. Some of the programs and applications will require implementation of technically advanced and complicated numerical novel algorithms. For these cases, programming expertise in numerical algorithms is needed. For many computationally intensive applications, such as vectorizing a loop, parallelizing an embarrassingly parallel simulation study or processing large genomic data, understanding of the underlying algorithm is of lesser importance. For these cases, the ability to write efficient code and to be able to tap into multi-core computing resources is of relatively greater importance. Public-use software disseminated by the program project is intended to be used by practitioners and not only expert users. This will necessitate equipping the application with a graphical user interface or a web interface. Given that users may use a variety of operating systems such as Linux, Microsoft Windows or Mac OS, it is imperative to develop cross platform application so as to minimize if not eliminate the need for writing platform specific applications. To address these needs, three different types of programmers will be recruited by the core. One of the programmers will have primary expertise in algorithms. The second programmer will have expertise in writing efficient code, parallelization, and embedding foreign languages. The third programmer will have expertise in creating user interfaces.

The core programmers will be in constant communication with one another via email. They will have a standing weekly conference call to discuss issues and progress. The main focus of these weekly calls will be on technical issues. This call will be attended by at least one member of the core leadership, the core director or one of the core co-Directors. Additionally, there will be a monthly conference call to be attended by the core leadership and the programmers to discuss more general issues. The programmers will take turns writing and posting minutes of these calls. These minutes will include follow-up items for each member of the core staff. Project investigators and the systems administration personnel described below will be invited to attend these calls as needed. Core programmers will also meet face-to-face as needed at one the institutions or at the NISS facilities. The core leadership will periodically review the progress of the programmers and provide feedback. The core leadership will also provide mentoring to the core programmers.

Core Management and Systems Administration. The NCSU Department of Statistics will provide systems management and administration support to the core. Professional systems administration staff at NCSU will manage and host the information technology resources needed to support the mission of the core; as discussed in Section **4.6**, the server hosting core code, software, documentation, and related materials will be housed at NCSU. Mr. Terry Byron is the Systems Administrator of the Department of Statistics at NCSU and is in charge of supporting the information technology needs of over 35 faculty and 170 graduate students. Mr. Byron has over 18 years experience in the planning, acquisition, and maintenance of academic computing systems, and he has a close working relationship with Dr. Davidian, with whom he has collaborated on other projects involving acquisition and management of computing resources. Mr. Byron will be assisted by Mr. Chris Waddell, who is also a full-time Systems Administrator in the Department. Mr. Waddell has 7 years experience, with 3 of these in the Department. Together, Mr. Byron and Mr. Waddell will manage the information technology needs of the core.

Dr. Davidian will meet weekly with Mr. Byron to review issues related to these functions.

4 METHODS AND SERVICES TO BE PROVIDED

Core C will provide an extensive host of services in support of the program research projects. We summarize each of these services and the methodology used in providing these in this section.

4.1 Aim 1: Implementation of Project Methodology

Simulation studies to test and evaluate performance of methods and the numerical algorithms involved in their implementation will be a key activity across all projects, as will be application of the methods to data from Core B and other sources demonstrating their use. In some cases, the nature of the methodology may be such that the investigators themselves can develop efficient and robust implementations for these purposes. In other

situations, the methods may be sufficiently complex and/or time-consuming to compute that implementation by a professional programmer with expertise in developing efficient code and/or algorithms is required. For example, in settings such as Aim 1 of Project 2 or Aim 1 of Project 5, where extensive empirical studies will be a main focus of the research, the need for code that of necessity must be fast and efficient is critical, given the computationally intensive nature of the studies to be conducted. The core will assist investigators faced with these challenges with the development and testing of of efficient, robust code implementing methods that the investigators can use to facilitate simulation studies and data analyses.

Project investigators requiring assistance will contact Dr. Owzar, the co-Director for this function, and work with him to develop a software development plan outlining in detail the required scope of the project. Code developed by the investigators to be refined or adapted by the core to efficient implementations will be provided. Once the development plan has been approved by Dr. Owzar and the core Director, Dr. Davidian, one or more core programmers under the supervision of Drs. Davidian, Lin or Owzar, and in collaboration with the investigators, will carry out the programming tasks involved.

4.2 Aim 2: Computational Resources in Support of Methodology

Certain computational tasks and methods will be common to several projects. For example, several of the projects involve techniques such as bootstrapping, and the intensive, large-scale simulations to be carried out in several of them will require parallel processing on a HPC cluster to be feasible. Core C will work with project investigators to identify such tasks and methods and develop well-tested and well-documented multi-use routines and modules to be made available as a shared resource across projects. Investigators developing their own code as well as core programmers will be able to exploit these routines and modules in developing their own applications. Moreover, the availability of such resources, which can be used by multiple investigators and projects, will minimize duplicative development efforts and minimize the chance for errors in testing and evaluation. Because a number of software tools will be used by investigators and core staff, Core C will also provide assistance with issues such as calling modules and routines written in one language from another.

At the beginning of the project, a meeting of all project Leaders and co-Leaders will be held at which an initial list of desired such generic, multi-use routines and modules will be developed. The Core C leadership will review the list and, in consultation with the Project Leaders and co-Leaders, develop a final list of initial resources to be developed. One or more core programmers under the supervision of Drs. Davidian, Lin and Owzar and in collaboration with appropriate investigators, will carry out the programming tasks involved. Subsequently, project investigators identifying needs for such resources will contact Dr. Davidian, who will oversee this function, and work with her to develop a development proposal. Upon approval of the proposal by the core leadership, in consultation with Project Leaders, core programmers will execute the plan.

4.3 Aim 3: Creation and Dissemination of Public-Use Software and Project Methodology

A key function of the core will be the design and development of robust, reliable, user-friendly, and well-documented software applications to bring the methods developed by the program project into practice. Once the operating characteristics and performance of a method have been extensively evaluated in simulation studies by the investigators and found to be satisfactory, investigators may request development of a public-use software application. After development, when an application is deemed to be sufficiently mature, it will be released to the users on the program project web portal dedicated to public use software developed by the project. The core will be available to assist investigators in arranging workshops, webinars, and other activities to demonstrate the software to the user community.

Investigators wishing to initiate development of public-use software will submit a development proposal to Dr. Lin, who will oversee this function. Upon approval of the proposal by the core leadership, in consultation with Project Leaders, core programmers will execute the plan. The entire development process will be based on an intimate collaboration between Core C Directors and staff and the project investigators. The investigators and programmers responsible for developing the application will conduct regular conference calls and face-to-face meetings as deemed necessary as development proceeds. A monthly conference call with a member of the core leadership will be conducted to keep the leadership apprised of progress and any unanticipated issues.

Core C will host a well-maintained web presence to provide access and information about its applications, activities, and accomplishments. Each software application will have its own dedicated page that will provide a summary of the application, links for download, installation information, and other documentation. Each released version of the application will be posted along with information about any new features added or bugs fixed since the last released version. It will also provide information about how to contact the developers through the issue tracking system to report problems or leave feedback. Software downloads will be tracked through this web page to give an indication about the usage.

4.4 Development Process

In the implementation of the methods, Core C will employ a standard software development process. This process will begin by working with the project investigators to understand the requirements and scope of the application to be developed. The next step will be to find out if the application needs to developed from scratch or if it can be developed by extending or modifying an existing application. The next step is to specify how the software will be developed and decide on the architecture. After these steps have been completed, the actual coding will begin. Testing and documentation will be carried out simultaneously. The software will then be under continual maintenance by correcting bugs or adding features provided by users.

Best Practices. Core C will conduct all of its activities according to so called best practices. It is committed to providing code and software that has been developed systematically in consultation with the project investigators and has been tested against nice as well as extreme or pathological cases. It is also committed to develop code that is well documented for both expert users who seek to understand not only how to use the software but also understand the underlying methodology and for non-expert users who are content with using the software as a black box.

Software Tools. Core C investigators have extensive experience in using statistical software and programming languages to develop code and applications for support of their statistical research and clinical design. The R statistical environment has proven to be a superb platform for statistical computing and application development. Its facilities for producing powerful graphics in a straightforward and flexible fashion are unmatched. As R is a interpreted language, it is not optimized for looping. The Core C investigators have extensive experience embedding the compiled languages C/C++ and Fortran in R. This approach will allow the investigator to code computationally intensive parts of the application using one of these compiled languages and the dynamically load the the objected as shared libraries.

The Python programming language has enjoyed great success in the quantitative research field. It offers a wide range of modules for manipulating text files and expressions. More importantly, it provides facilities for interfacing Another useful feature is that one can embed R into Python using Rpy. There are also interfaces for embedding C/C++ and Fortran into Python.

The Core C investigators have extensive experience using parallelization programs such as MPI and OpenMP. Given that many of these simulations are embarrassingly parallel, the use of these programs is very important as it will allow the investigators to take fully take advantage of multi-core HPC resources, which will be critical for large-scale computational efforts.

The investigators also have extensive experience and access to the Maxima, Mathematica, and Maple symbolic processing software. One important feature of these software is that they provide powerful facilities for manipulating complicated expressions algebraically. Once the expression is manipulated into the desired form, it can then be exported into either C or Fortran for embedding into R. The core investigators also have access and expertise with Matlab and SAS, and will use these software as needed. For some project methodology of interest to users who may be involved in the development of regulatory submissions, such as that in Aim 1 of Project 2, it will be important to provide public-use implementations in SAS where possible. The need for SAS implementation will be determined through consultation between project investigators and the Core C leadership.

For some applications, especially for those geared toward practitioners, the application may be interfaced through a graphical user interfaces (GUI). A flexible approach to this problem is to use wxPython which is a GUI toolkit for Python. It is a cross-platform toolkit that will allow us to run the application on multiple platforms (e.g., Linux, Microsoft Windows and Mac OS) without modification. One comprehensive approach for developing software applications which would benefit from having a GUI is be to develop the application using Python by embedding R, C/C++ and Fortran code as necessary and then equip it with a wxPython interface. Needless

to say, the C/C++ and Fortran libraries need to be recompiled depending on the platform. However, careful programming should greatly minimize if not eliminate the need for rewriting the code.

It should be noted that all of the software listed above, with the exception of Matlab, Mathematica and Maple, are or expected to soon be open-source software. For the application development, Core C will use open source software to avoid licensing issues for the users.

Source Code Management. Source code management is imperative for proper application development. We will use the subversion system for source code management. The clients for interfacing with subversion are available on several platforms including Linux, Windows and MacOS. This system is used among other projects by R and Bioconductor. It allows to check out and commit locally or remotely via ssh.

The programmers will be required to use this system not only to commit code modifications but also to document their modifications. This is important as it will facilitate collaboration among programmers and also put a new programmer in a better position to get involved as the documentation will contribute to understanding the chronological story of the project. The log history will also allow the core leaders to monitor the progress of each software project and as such will facilitate the management and evaluation of the project programmers.

The subversion code repository will hosted on a dedicated server and be backed up to protect against catastrophic hardware failure, corruption of the repository or unrecoverable user error. Additional details, are provided in the description of the information technology infrastructure.

Unit Testing and Reproducibility. Even a seemingly minor harmless code modification, may have catastrophic unintended consequences to the cases. Furthermore, the robustness of a software cannot be adequately judged when testing it exclusively against nice cases. The core will employ the concept of unit testing for its software projects. A number of normal, extreme and pathological test cases will be developed. Once we have ascertained, in consultation with the project investigators, that the outcome of these tests are what is to be expected, the test outcomes are archived. Anytime the code for a software project is modified, all of the test cases for that project are rerun. The outcome of these test cases will then be compared to the archived versions. Any discrepancies are noted and investigated. Whenever a bug is reported along with a reproducible example, the example is added as a new test case. Implementation of unit testing requires considerable effort but given that the software is to be used for analyses of data from clinical trials it is incumbent upon us to implement systematic and robust testing procedures to protect against regressions. It is also well known that unit testing serves as a live documentation and as such will improve the overall quality of the software. It should be noted that for many programming environments such as R and Python, there are systems to facilitate the implementation of unit testing.

Depending on the complexity and nature of the software application, in consultation with the investigators who developed the methods implemented, the core may invite potential users, who may be experts in the particular methodological area or practitioners with strong interest in the methods, to serve as beta-testers for a relatively mature version of the software prior to its release to the public. The beta-testers will be asked to use the software, report any glitches and difficulties encountered, and offer recommendations on refinements that may improve accessibility.

Documentation. The project is tasked with developing software that is useful to both researchers and practitioners. To this end, the provision of adequate documentation is imperative. Proper documentation not only improves the quality of the software, as it forces the developer to think through what the code is expecting and returning, but also increases the usage as well documented software is considered to be more user-friendly.

For R packages, the native R package documentation system will be used. For other languages, we will use existing system for technical documentation systems such as DocBook.

In addition to standard documention, tutorials providing demonstrations of the use of the software on data sets, will be created. The core will facilitate this activity by developing standard templates for these resources. The core will work with the relevant investigators to assist them in working with the programmers responsible for the software to develop these tutorials.

Issue Tracking System. Core C will implement an electronic issue tracking system to allow project investigators to report issues and request support. This will not only allow us to track the progress of the support requests but also allow the leadership to assess the responsiveness of the core staff to issues.

Mailing List. Core C will maintain a mailing list to transmit communications efficiently to the project investigators and personnel. This will ensure that project investigators from all projects are aware of the core's activities and milestones. The mailing membership list will be actively maintained to ensure that new members are added in

an expeditious manner and that individuals no longer involved with the projects are removed.

4.5 Training

To assist the investigators and research assistants working on the individual projects, who will be developing and running code for research purposes, to write more efficient and robust programs and execute them effectively, the core will host periodic training workshops. For example, sessions on parallel computing, Python programming, efficient optimization algorithms will be organized, led by the core leaders, programmers, or speakers recruited from outside the program project, as appropriate. The core will also develop training materials covering such topics and post them and presentation materials on the core web site for access by project personnel.

4.6 Computing and Information Technology Resources

Collectively, the three participating institutions boast extensive state-of-the-art HPC resources and general information technology resources to support the mission of Core C. A summary of these resources is provided next.

HPC Computing Resources. Investigators at Duke, have access to departmental HPC resources consisting of two 16-core AMD Opteron 8222SE servers with 64GB of RAM (expandable to 128GB) each and one 8-core AMD Opteron AMD 8222SE server with 32GB of RAM (expandable to 64GB of RAM). Each server has 3TB of storage in RAID 10. These servers are managed by the Duke University Office of Information technology (OIT). These servers provide parallelization facilities using MPI (up to 40 cores) and OpenMP (up to 16 cores). Additionally, the Duke core investigators will have access to a server grade two-way dual-core Opteron workstation with 16GB of RAM with RAID, a two-way quad-core Xeon workstation with 16GB of RAM and a dual-core mobile workstation with 8GB of RAM.

Investigators at NCSU have access to multiple HPC computing resources to serve their computing needs. The Department of Statistics maintains a 20-core Beowulf cluster utilizing 3.6GHz Xeon processors with 4GB of RAM (expandable to 24GB) each and a 12-core Windows AD cluster utilizing 2.3GHz Intel processors with 4GB of RAM (expandable to 24GB) each. These Department clusters are customized according to the needs of Department investigators. There are additional Linux HPC computing resources available at both the College of Physical and Mathematical Sciences and University levels that are also available, each having similar computing power and performance. Each investigator has access at minimum to an Intel 2-core workstation using an Intel 3.0GHz processor with 4GB of RAM as a primary desktop resource.

Investigators at UNC-CH, have access to departmental HPC resources consisting of a total of 126 job slots for serial jobs and 8TB of available space on a IBM DS4300 SAN. These include 11 two-way Intel Xeon servers each with 4GB of RAM, 8 two-way Xeon quad-core servers with 8GB of RAM and 4 two-way Xeon quad-core servers with 16GB of RAM. Additionally, the investigators at UNC-CH will have access to university-wide HPC resources. The investigators at UNC-CH have access to a wide range of HPC resources through the UNC-CH Information Technology Services (ITS) research computing division. These include a 4160-core Dell Linux cluster with Infiniband interconnects, a 700-processor general purpose Linux cluster, a 128-processor SGI Altix with 512 GB of memory and an 8-processor login/interactive front-end, four 16-way Power5-based servers, and a 32-processor IBM P690. The Dell cluster has ranked in the top 100 of the Top500 supercomputer list since 2006. Storage for research data accessed on the above systems includes more than 100 terabytes of disk, comprising locally attached disks; network-attached shared scratch space, and network-attached shared file systems. In addition, an archival mass storage system, with a capacity of more than 700 TB, is available.

Application and Source Code Repository Server. The core will acquire and manage a web server to host the software applications developed by the project for dissemination to the public and to host the source code repository. This will be a Dell PowerEdge 2950 4-Core server utilizing 2.5GHz AMD processors with 16GB of RAM (expandable to 128GB) or better. It will be housed in a server room in the newly constructed Mathematics and Statistics Building at NCSU, scheduled to be dedicated in May 2009. The new server room will have automated fire detection and suppression, a separate cooling system, key card security, high speed fiber link, and expandable UPS.

Backup Facilities. Core C is tasked with the stewardship of the source code repository. Although the RAID array of the server will provide some level of redundancy, it will not be relied on as a backup to protect against

catastrophic hardware failure, software corruption, or user errors. The core will employ a tape-based system for archiving the code repository. The tape media will be stored remotely in a different building house on the NCSU campus. The backups will be performed nightly (Monday through Friday) by Mr. Byron or Mr. Waddell and archived for at least one month before rotating. If it is determined that longer archival is required, adjustments can be easily made with the purchase of additional tape media.

5 INCLUSION ENROLLMENT REPORT

N/A

6 BIBLIOGRAPHY AND REFERENCES CITED

N/A

7 PROTECTION OF HUMAN SUBJECTS

Although the proposed research indirectly involves human subjects through the preparation, in Core B, of de-identified data sets from identifiable patient data sources, the investigators on Core C will have access only to the de-identified data. Thus, the investigators on Core C will have no access to any identifiable patient information.

8 INCLUSION OF WOMEN AND MINORITIES

The methods we develop will be applicable to studies with both women and minorities and also to studies which examine treatment differences adjusted for gender, ethnicity and race. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, many of the existing data sets to be studied and provided by Core B include women and minorities, although we will not be generating any new data involving human subjects.

9 TARGETED/PLANNED ENROLLMENT TABLE

N/A

10 INCLUSION OF CHILDREN

The methods we develop will be applicable to studies with children and also to studies which examine treatment differences adjusted for age. This is accomplished through the general formulation of the statistical designs, models and methods studied that allow for many possible kinds of risk factors. Moreover, some of the existing data sets to be studied and provided by Core B may include children, although we will not be generating any new data involving human subjects.

11 VERTEBRATE ANIMALS

N/A

12 SELECT AGENT RESEARCH

N/A

13 MULTIPLE PD/PI LEADERSHIP PLAN

N/A

14 CONSORTIUM/CONTRACTUAL ARRANGEMENTS

If the present application is funded, the University of North Carolina at Chapel Hill will execute subcontracts with the consortium institutions (Duke University and North Carolina State University). These inter-institutional agreements will be written consistent with the NIH consortium agreement policy.

15 LETTERS OF SUPPORT

N/A

16 RESOURCE SHARING PLAN(S)

The services of Core C will be used by all of five individual research projects in this program. Our estimate is that the percentage of core usage by the five projects will be roughly the same, that is,

Project 1 - 20%
2 - 20%
3 - 20%
4 - 20%
5 - 20%

The following is the external component of our resource sharing plan:

(a) Data sharing plan: The data-related resources generated by the proposed research consists of new statistical methodology, software packages for implementation of the methodology, and tutorials for the software. The statistical methodology will be shared through peer reviewed publications and national meetings and through other standard means. All accepted publications will be deposited in PubMed Central in accordance with the NIH Public Access Policy. Summaries of the methodology, the software and tutorials will be shared through a public web site managed by Core A, while Core C will assist in preparation of the software and tutorials for dissemination. Core C will use de-identified data prepared by Core B to test the methods and to create demonstrations of use of the methods to be included in tutorials. Core C will not be involved in sharing of these data; this function will be addressed by Core B.

(b) Sharing model organisms: N/A

(c) GWAS: N/A

PERCENTAGE DISTRIBUTION OF SCIENTIFIC CORE RESEARCH RESOURCES TO PROJECTS

Project	Project 1	Project 2	Project 3	Project 4	Project 5	Total (100 %)
Core A: Administrative Core	20	20	20	20	20	100
Core B: Data Compilation Core	30	20	10	30	10	100
Core C: Computational Resource and Dissemination Core	20	20	20	20	20	100