On the relative efficiency of using summary statistics versus individual level data in meta-analysis

By D. Y. LIN and D. ZENG

Department of Biostatistics, CB# 7420, University of North Carolina, Chapel Hill, North Carolina 27599-7420, U.S.A.

lin@bios.unc.edu dzeng@bios.unc.edu

SUMMARY

Meta-analysis is widely used to synthesize the results of multiple studies. Although meta-analysis is traditionally carried out by combining the summary statistics of relevant studies, advances in technologies and communications have made it increasingly feasible to access the original data on individual participants. In the present paper, we investigate the relative efficiency of analyzing original data versus combining summary statistics. We show that, for all commonly used parametric and semiparametric models, there is no efficiency gain by analyzing original data if the parameter of main interest has a common value across studies, the nuisance parameters have distinct values among studies, and the summary statistics are based on maximum likelihood. We also assess the relative efficiency of the two methods when the parameter of main interest has different values among studies or when there are common nuisance parameters across studies. We conduct simulation studies to confirm the theoretical results and provide empirical data from a genetic association study.

Some key words: Cox regression; Evidence-based medicine; Genetic association; Individual patient data; Information matrix; Linear regression; Logistic regression; Maximum likelihood; Profile likelihood; Research synthesis.

1. INTRODUCTION

Meta-analysis, the combination of results from a series of independent studies, is gaining popularity in many fields, including medicine, psychology, epidemiology, education, genetics and ecology. In particular, meta-analysis publications in medical research have grown enormously over the last three decades, due to greater emphasis on evidence-based medicine and the need for reliable summarization of the vast and expanding volume of clinical research (e.g., Sutton et al., 2000; Whitehead, 2002). Most of the recent discoveries on genetic variants influencing complex human diseases were made possible through meta-analysis of multiple studies (e.g., Lohmueller et al., 2003; Zeggini et al., 2008).

Traditionally, meta-analysis is carried out by combining the summary statistics of relevant studies, which are available in journal articles. With improving technologies and communications and increasing recognition of the benefits of meta-analysis, it is becoming more feasible to obtain the raw or original data on individual participants (e.g., Sutton et al., 2000). Indeed, meta-analysis of individual patient data is regarded as the gold standard in systematic reviews of randomized clinical trials (e.g., Chalmbers et al., 1993). Recently, a number of networks or consortia have been created to share original data from genetic association studies (e.g., Kavvoural & Ioannidis 2008; The Psychiatric GWAS Consortium Steering Committee, 2009). In general, obtaining original data is difficult, costly and time-consuming. A question naturally arises as to how much efficiency gain can be achieved by analyzing original data over combining summary statistics.

A partial answer to this question was provided by Olkin & Sampson (1998), who showed that, in the case of comparing multiple treatments and a control with respect to a continuous outcome, the traditional meta-analysis based on estimated treatment contrasts is equivalent to the least-squares regression analysis of individual patient data if there are no study-by-treatment interactions and the error variances are constant across trials. Mathew & Nordström (1999) stated that the equivalence holds even if the error variances are different across trials. There has been no theoretical investigation beyond this special setting.

Empirically, meta-analysis using original data has been found to be generally similar but not identical to meta-analysis using summary statistics (e.g., Whitehead, 2002, Ch. 5).

In the present paper, we provide a systematic investigation into the relative efficiency of using summary statistics versus original data in fixed-effects meta-analysis, which assumes a common effect among studies. We prove that the two types of meta-analysis are asymptotically equivalent for all commonly used parametric and semiparametric models provided that the effect sizes are indeed the same for all studies, the nuisance parameters are distinct across studies, and maximum likelihood estimation is used in the calculations of summary statistics and in the joint analysis of original data. We also investigate the relative efficiency of the two methods when the effect sizes are different among studies or when there are common nuisance parameters across studies. We illustrate the theoretical results with simulated and empirical data.

2. THEORETICAL RESULTS

2.1. Main Results

Suppose that there are K independent studies, with n_k participants for the kth study. The original data consist of (Y_{ki}, X_{ki}) $(k = 1, \ldots, K; i = 1, \ldots, n_k)$, where Y_{ki} is the response variable for the ith participant of the kth study, and X_{ki} is the corresponding vector of explanatory variables. The response variable can be continuous or discrete, univariate or multivariate. Under fixed-effects models, the conditional density of Y_{ki} given X_{ki} takes the form $f(y, x; \beta, \eta_k)$, where β is a vector of parameters common to all K studies, and η_k is a vector of parameters specific to the kth study. A simple example is the linear regression model for the normal response variable:

$$Y_{ki} = \alpha_k + \beta^{\mathrm{T}} X_{ki} + \epsilon_{ki}, \quad k = 1, \ldots, K; \ i = 1, \ldots, n_k,$$

where ϵ_{ki} is normal with mean zero and variance σ_k^2 (Whitehead, 2002, §5.2.1). In this case, $f(y, x; \beta, \eta_k) = (2\pi\sigma_k^2)^{-1/2} \exp\{-(y - \alpha_k - \beta^{\mathrm{T}} x)^2 / 2\sigma_k^2\}$, and $\eta_k = (\alpha_k, \sigma_k^2)$. Additional examples are given in §2.4. We wish to make inference about β.

Meta-analysis is usually performed on a scalar parameter. We allow vector-valued β for two reasons. First, there are important applications in which the effects of interest, such as treatment differences in a multi-arm clinical trial or co-dominant effects of a genetic variant, are truly multivariate. Second, if the nuisance parameters, e.g., intercepts or confounding effects, have the same values among the K studies, then performing meta-analysis jointly on the effects of interest and the common nuisance parameters can improve statistical efficiency, as will be discussed in §2.2. Of course, our formulation includes scalar β as a special case.

Let $\widehat{\beta}_k$ be the maximum likelihood estimator (MLE) of β by maximizing the kth study likelihood:

$$L_k(\beta, \eta_k) := \prod_{i=1}^{n_k} f(Y_{ki}, X_{ki}; \beta, \eta_k),$$

and let $\widetilde{\beta}$ be the MLE of β by maximizing the joint likelihood

$$L(\beta, \eta_1, \ldots, \eta_K) := \prod_{k=1}^{K} L_k(\beta, \eta_k).$$

The profile likelihood functions for β based on $L_k(\beta, \eta_k)$ and $L(\beta, \eta_1, \ldots, \eta_K)$ are, respectively,

$$pl_k(\beta) := \sup_{\eta_k} L_k(\beta, \eta_k),$$

and

$$pl(\beta) := \sup_{\eta_1, \ldots, \eta_K} L(\beta, \eta_1, \ldots, \eta_K).$$

The corresponding observed profile information matrices are $\mathcal{I}_k(\beta) := -\partial^2 \log pl_k(\beta)/\partial\beta^2$ and $\mathcal{I}(\beta) := -\partial^2 \log pl(\beta)/\partial\beta^2$. The maximizer of the profile likelihood is the same as the MLE in that $\widehat{\beta}_k = \text{argmax}\, pl_k(\beta)$ and $\widetilde{\beta} = \text{argmax}\, pl(\beta)$. Write $n = \sum_k n_k$ and assume that $n_k/n \to c_k \in (0, 1)$ as $n \to \infty$. Assume also that the regularity conditions for profile likelihood as stated in Murphy & van der Vaart (2000) hold. Then $n_k \mathcal{I}_k^{-1}(\widehat{\beta}_k)$ and $n\mathcal{I}^{-1}(\widetilde{\beta})$ are consistent estimators of the covariance matrices of $n_k^{1/2}(\widehat{\beta}_k - \beta)$ and $n^{1/2}(\widetilde{\beta} - \beta)$, respectively.

Remark 1. For survival data and other censored data, the likelihood needs to be modified. If Y_{ki} is right censored at $\widetilde{Y}_{ki}$, then we replace $f(Y_{ki}, X_{ki}; \beta, \eta_k)$ in the likelihood by $S(\widetilde{Y}_{ki}, X_{ki}; \beta, \eta_k)$, where $S(y, x; \beta, \eta_k) = \int_y^\infty f(u, x; \beta, \eta_k) du$.

Remark 2. Our framework allows different likelihood functions among studies, and the statistical models are not necessarily regression models. In meta-analysis of diagnostic accuracy data, the likelihood function for each study pertains to the multinomial distribution of a 2×2 contingency table, and β can be the sensitivity or specificity or both.

In traditional meta-analysis, one collates summary statistics $\widehat{\beta}_k$ and $\widehat{\mathrm{var}}(\widehat{\beta}_k) := 1/\mathcal{I}_k(\widehat{\beta}_k)$ $(k = 1, \ldots, K)$ for a scalar parameter β. The well-known inverse-variance estimator of β is

$$\widehat{\beta} := \frac{\sum_{k=1}^{K} \widehat{\beta}_k / \widehat{\mathrm{var}}(\widehat{\beta}_k)}{\sum_{k=1}^{K} 1/\widehat{\mathrm{var}}(\widehat{\beta}_k)}, \tag{1}$$

and its variance is estimated by

$$\widehat{\mathrm{var}}(\widehat{\beta}) := \frac{1}{\sum_{k=1}^{K} 1/\widehat{\mathrm{var}}(\widehat{\beta}_k)}.$$

To allow vector-valued β, we propose a multivariate version of estimator (1):

$$\widehat{\beta} = \left\{ \sum_{k=1}^{K} \mathcal{I}_k(\widehat{\beta}_k) \right\}^{-1} \sum_{k=1}^{K} \mathcal{I}_k(\widehat{\beta}_k) \widehat{\beta}_k, \tag{2}$$

whose covariance matrix is estimated by

$$\widehat{\mathrm{var}}(\widehat{\beta}) := \left\{ \sum_{k=1}^{K} \mathcal{I}_k(\widehat{\beta}_k) \right\}^{-1}. \tag{3}$$

If original data are available, one can estimate β by the MLE $\widetilde{\beta}$, whose covariance matrix is estimated by $\widehat{\mathrm{var}}(\widetilde{\beta}) := \mathcal{I}^{-1}(\widetilde{\beta})$. It is easy to see that $pl(\beta) = \prod_{k=1}^{K} pl_k(\beta)$, which implies that $\mathcal{I}(\beta) = \sum_{k=1}^{K} \mathcal{I}_k(\beta)$. Thus,

$$\widehat{\mathrm{var}}(\widetilde{\beta}) = \left\{ \sum_{k=1}^{K} \mathcal{I}_k(\widetilde{\beta}) \right\}^{-1}. \tag{4}$$

Equations (3) and (4) show that $\widehat{\mathrm{var}}(\widehat{\beta})$ and $\widehat{\mathrm{var}}(\widetilde{\beta})$ take the same form, the only difference being that the $\mathcal{I}_k(\beta)$ are evaluated at the $\widehat{\beta}_k$ in the former and at $\widetilde{\beta}$ in the latter. Under standard regularity conditions, $\widehat{\beta}_k$ $(k = 1, \ldots, K)$ and $\widetilde{\beta}$ converge to β, and $n_k^{-1}\mathcal{I}_k(\beta)$ $(k = 1, \ldots, K)$ converge to constant matrices. It follows that $n^{1/2}(\widehat{\beta} - \beta)$ and $n^{1/2}(\widetilde{\beta} - \beta)$ have the same limiting normal distribution. Thus, using summary statistics has the same asymptotic efficiency as using original data.

2.2. Common Nuisance Parameters

According to the results of the last section, meta-analysis based on summary statistics has the same asymptotic efficiency as the MLE of full data if the former analysis is performed jointly on all common parameters. It is generally difficult to obtain multivariate summary statistics, especially in retrospective meta-analysis of published results. Thus, it is important to determine the efficiency loss of meta-analysis based on the univariate summary statistics for the effect of main interest when there are other common effects, which are referred to as common nuisance parameters.

Suppose that β is a scalar parameter representing a common effect of main interest and that a subset of η_k, denoted by γ, is a vector of common nuisance parameters. Denote the profile information matrices for (β, γ) based on the kth study data and all the data as

$$\mathcal{I}_k = \begin{pmatrix} \mathcal{I}_{k\beta\beta} & \mathcal{I}_{k\beta\gamma} \\ \mathcal{I}_{k\gamma\beta} & \mathcal{I}_{k\gamma\gamma} \end{pmatrix},$$

and

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\gamma} \\ \mathcal{I}_{\gamma\beta} & \mathcal{I}_{\gamma\gamma} \end{pmatrix},$$

respectively. The variance of $\widehat{\beta}_k$ is approximately $(\mathcal{I}_{k\beta\beta} - \mathcal{I}_{k\beta\gamma}\mathcal{I}_{k\gamma\gamma}^{-1}\mathcal{I}_{k\gamma\beta})^{-1}$, so the variance of $\widehat{\beta}$ is approximately $\{\sum_k(\mathcal{I}_{k\beta\beta} - \mathcal{I}_{k\beta\gamma}\mathcal{I}_{k\gamma\gamma}^{-1}\mathcal{I}_{k\gamma\beta})\}^{-1}$. The variance of $\widetilde{\beta}$ is approximately $\left(\mathcal{I}_{\beta\beta} - \mathcal{I}_{\beta\gamma}\mathcal{I}_{\gamma\gamma}^{-1}\mathcal{I}_{\gamma\beta}\right)^{-1}$. Thus, the relative efficiency of $\widehat{\beta}$ to $\widetilde{\beta}$ is approximately

$$\frac{\sum_k \mathcal{I}_{k\beta\beta} - \sum_k \mathcal{I}_{k\beta\gamma}\mathcal{I}_{k\gamma\gamma}^{-1}\mathcal{I}_{k\gamma\beta}}{\mathcal{I}_{\beta\beta} - \mathcal{I}_{\beta\gamma}\mathcal{I}_{\gamma\gamma}^{-1}\mathcal{I}_{\gamma\beta}}.$$

Because $\mathcal{I} = \sum_k \mathcal{I}_k$, the relative efficiency can also be expressed as

$$\frac{\sum_k \mathcal{I}_{k\beta\beta} - \sum_k \mathcal{I}_{k\beta\gamma}\mathcal{I}_{k\gamma\gamma}^{-1}\mathcal{I}_{k\gamma\beta}}{\sum_k \mathcal{I}_{k\beta\beta} - (\sum_k \mathcal{I}_{k\beta\gamma})(\sum_k \mathcal{I}_{k\gamma\gamma})^{-1}(\sum_k \mathcal{I}_{k\gamma\beta})}.$$

It follows from Lemma 1 of Appendix A that

$$\sum_k \mathcal{I}_{k\beta\gamma}\mathcal{I}_{k\gamma\gamma}^{-1}\mathcal{I}_{k\gamma\beta} \geq (\sum_k \mathcal{I}_{k\beta\gamma})(\sum_k \mathcal{I}_{k\gamma\gamma})^{-1}(\sum_k \mathcal{I}_{k\gamma\beta}).$$

Thus, the relative efficiency is always less than or equal to 1. It also follows from Lemma 1 that the relative efficiency is 1 if and only if $\mathcal{I}_{k\beta\gamma}\mathcal{I}_{k\gamma\gamma}^{-1} = \mathcal{I}_{l\beta\gamma}\mathcal{I}_{l\gamma\gamma}^{-1}$ $(\forall k \neq l)$. Note that

$\mathcal{I}_{k\beta\gamma}\mathcal{I}_{k\gamma\gamma}^{-1} \approx -\text{var}(\widehat{\beta}_k)^{-1}\text{cov}(\widehat{\beta}_k, \widehat{\gamma}_k)$, where $\widehat{\beta}_k$ and $\widehat{\gamma}_k$ are the MLEs of β and γ based on the kth study data. Thus, the relative efficiency is 1 if and only if $\text{var}(\widehat{\beta}_k)^{-1}\text{cov}(\widehat{\beta}_k, \widehat{\gamma}_k)$ are the same among the K studies. Obviously, the latter condition is satisfied if $\text{cov}(\widehat{\beta}_k, \widehat{\gamma}_k) = 0$ $(k = 1, \ldots, K)$. The foregoing conclusions also hold for multivariate β.

2.3. Unequal Effect Sizes

The fixed-effects meta-analysis assumes that the effect sizes β_k are the same across studies. This assumption does not affect the type I error of hypothesis testing since all effect sizes are the same under the null hypothesis. However, it is of practical importance to determine the relative power of using summary statistics versus original data when the β_k are unequal.

Write $U_k(\beta) = \partial \log pl_k(\beta)/\beta$ and $U(\beta) = \partial \log pl(\beta)/\beta$. By definition, $U_k(\widehat{\beta}_k) = 0$ $(k = 1, \ldots, K)$ and $U(\widetilde{\beta}) = 0$. Since $U(\beta) = \sum_{k=1}^K U_k(\beta)$, we have $\sum_{k=1}^K \left\{U_k(\widehat{\beta}_k) - U_k(\widetilde{\beta})\right\} = 0$. By the mean-value theorem, $\sum_{k=1}^K \mathcal{I}_k(\beta_k^*)(\widehat{\beta}_k - \widetilde{\beta}) = 0$, where β_k^* lies between $\widehat{\beta}_k$ and $\widetilde{\beta}$. In other words,

$$\widetilde{\beta} = \left\{\sum_{k=1}^K \mathcal{I}_k(\beta_k^*)\right\}^{-1} \sum_{k=1}^K \mathcal{I}_k(\beta_k^*)\widehat{\beta}_k. \tag{5}$$

Comparison of (5) to (2) reveals that $\widetilde{\beta}$ is the same kind of weighted combination of the $\widehat{\beta}_k$ as $\widehat{\beta}$, with the weights $\mathcal{I}_k(\beta)$ evaluated at the β_k^* rather than the $\widehat{\beta}_k$. As shown in §2.1, the only difference between $\widehat{\text{var}}(\widetilde{\beta})$ and $\widehat{\text{var}}(\widehat{\beta})$ lies in the evaluation of the $\mathcal{I}_k(\beta)$ at $\widetilde{\beta}$ versus the $\widehat{\beta}_k$. Hence, meta-analysis based on summary statistics and meta-analysis of original data will have similar power provided that the $\mathcal{I}_k(\beta)$ do not changes their values drastically when β varies between the $\widehat{\beta}_k$ and $\widetilde{\beta}$. When the β_k are unequal, the limits of $\widehat{\beta}$ and $\widetilde{\beta}$ pertain to weighted combinations of the β_k, rather than a common parameter is a statistical model.

We consider the local alternatives $H_{1n} : \beta_k^{(n)} = \beta + O(n^{-1/2})$, $k = 1, \ldots, K$. Under H_{1n}, the estimators $\widehat{\beta}_k$ $(k = 1, \ldots, K)$ converge in probability to β. We show in Appendix B that $\widetilde{\beta}$ also converges in probability to β under H_{1n}. It follows that, for each k, the weights $n^{-1}\mathcal{I}_k(\widehat{\beta}_k)$, $n^{-1}\mathcal{I}_k(\widetilde{\beta})$ and $n^{-1}\mathcal{I}_k(\beta_k^*)$ all converge in probability to the same constant matrix. Thus, meta-analysis based on summary statistics and meta-analysis of original data have the same asymptotic power against H_{1n}.

For hypothesis testing, one can use score statistics in meta-analysis. For testing the null hypothesis $H_0 : \beta = \beta_0$, the score statistics based on summary statistics and original data are

$$\left\{\sum_{k=1}^{K} U_k(\beta_0)\right\}^{\mathrm{T}} \left\{\sum_{k=1}^{K} \mathcal{I}_k(\beta_0)\right\}^{-1} \left\{\sum_{k=1}^{K} U_k(\beta_0)\right\},$$

and $U^{\mathrm{T}}(\beta_0)\mathcal{I}^{-1}(\beta_0)U(\beta_0)$, respectively. The two statistics are numerically identical since $U(\beta) = \sum_{k=1}^{K} U_k(\beta)$ and $\mathcal{I}(\beta) = \sum_{k=1}^{K} \mathcal{I}_k(\beta)$. This equivalence holds whether the true effect sizes are equal or not.

2.4. Special Cases

In this section, we consider three most common cases of meta-analysis: linear regression for continuous response, logistic regression for binary response, and Cox regression for potentially censored survival time. We will pay particular attention to the form of the observed profile information matrices $\mathcal{I}_k$ because, as shown in §§2.1 and 2.3, the only difference between $\widehat{\beta}$ and $\widetilde{\beta}$ lies in the argument of the profile information matrix. We use X to denote the explanatory variables of main interest and Z to denote the unit component and possibly other explanatory variables or covariates that are measured at the individual level. The numbers and types of covariates need not be the same among the K studies.

Example 1 (Linear regression). Assume that the distribution of Y_{ki} conditional X_{ki} and Z_{ki} is normal with mean $\beta^{\mathrm{T}}X_{ki} + \gamma_k^{\mathrm{T}}Z_{ki}$ and variance σ_k^2. The observed profile information matrix for β based on the kth study data is $\mathcal{I}_k(\sigma_k^2) := D_k/\sigma_k^2$, where

$$D_k = \sum_{i=1}^{n_k} X_{ki}^{\otimes 2} - \sum_{i=1}^{n_k} X_{ki}Z_{ki}^{\mathrm{T}}\Big(\sum_{i=1}^{n_k} Z_{ki}^{\otimes 2}\Big)^{-1}\sum_{i=1}^{n_k} Z_{ki}X_{ki}^{\mathrm{T}},$$

and $a^{\otimes 2} = aa^{\mathrm{T}}$. The MLEs of σ_k^2 based on the kth study data and all data are, respectively,

$$\widehat{\sigma}_k^2 = n_k^{-1}\sum_{i=1}^{n_k}(Y_{ki} - \widehat{\beta}_k^{\mathrm{T}}X_{ki} - \widehat{\gamma}_k^{\mathrm{T}}Z_{ki})^2,$$

and

$$\widetilde{\sigma}_k^2 = n_k^{-1}\sum_{i=1}^{n_k}(Y_{ki} - \widetilde{\beta}^{\mathrm{T}}X_{ki} - \widetilde{\gamma}_k^{\mathrm{T}}Z_{ki})^2,$$

where $\widehat{\beta}_k$ and $\widehat{\gamma}_k$ are the least-squares estimators of β and γ_k based on the kth study data, and $\widetilde{\beta}$ and $\widetilde{\gamma}_k$ are the MLEs of β and γ_k based on all data. By definition, $\widehat{\beta} = \{\sum_k \mathcal{I}_k(\widehat{\sigma}_k^2)\}^{-1} \sum_k \mathcal{I}_k(\widehat{\sigma}_k^2)\widehat{\beta}_k$ and $\widehat{\mathrm{var}}(\widehat{\beta}) = \{\sum_k \mathcal{I}_k(\widehat{\sigma}_k^2)\}^{-1}$. It is easy to show that $\widetilde{\beta} = \{\sum_k \mathcal{I}_k(\widetilde{\sigma}_k^2)\}^{-1} \sum_k \mathcal{I}_k(\widetilde{\sigma}_k^2)\widehat{\beta}_k$ and $\widehat{\mathrm{var}}(\widetilde{\beta}) = \{\sum_k \mathcal{I}_k(\widetilde{\sigma}_k^2)\}^{-1}$. Thus, $\widehat{\beta}$ and $\widetilde{\beta}$, and their variance estimators differ only in whether $\mathcal{I}_k(\sigma_k^2)$ is evaluated at $\widehat{\sigma}_k^2$ or $\widetilde{\sigma}_k^2$. In general, $\widehat{\sigma}_k^2$ and $\widetilde{\sigma}_k^2$ are approximately the same, so that the results of meta-analysis using summary statistics and using original data are similar. Under the assumed model, both $\widehat{\sigma}_k^2$ and $\widetilde{\sigma}_k^2$ converge to σ_k^2, so that using summary statistics is asymptotically equivalent to using original data. If the assumed model is incorrect, then the two estimators may converge to different constants.

The setting considered by Olkin & Sampson (1998) and Mathew & Nordström (1999) is a special case of our model in which X consists of treatment indicators and $Z = 1$. In this setting, D_k is a function of group sizes only. Under the assumption that $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2$, one can define $\widehat{\beta}$ as $(\sum_k D_k)^{-1} \sum_k D_k \widehat{\beta}_k$, which also turns out to be the expression of $\widetilde{\beta}$ under the common variance assumption. Thus, using summary statistics is numerically identical to using original data, which is the finding of Olkin & Sampson (1998). Mathew & Nordström (1999) stated that the equivalence continues to hold even if the error variances are unequal. They used the true values of the σ_k^2 in their definition of $\widehat{\beta}$. Since the σ_k^2 need to be estimated from the data, the equivalence holds only asymptotically rather than numerically.

Example 2 (Logistic regression). Assume that

$$\mathrm{pr}(Y_{ki} = 1 | X_{ki}, Z_{ki}) = \frac{e^{\beta^{\mathrm{T}} X_{ki} + \gamma_k^{\mathrm{T}} Z_{ki}}}{1 + e^{\beta^{\mathrm{T}} X_{ki} + \gamma_k^{\mathrm{T}} Z_{ki}}}, \quad k = 1, \ldots, K; \; i = 1, \ldots, n_i.$$

Write $\theta_k = (\beta, \gamma_k)$. The observed profile information matrix for β based on the kth study data is

$$\mathcal{I}_k(\theta_k) := \sum_{i=1}^{n_k} v_{ki}(\theta_k) X_{ki}^{\otimes 2} - \left\{ \sum_{i=1}^{n_k} v_{ki}(\theta_k) X_{ki} Z_{ki}^{\mathrm{T}} \right\} \left\{ \sum_{i=1}^{n_k} v_{ki}(\theta_k) Z_{ki} Z_{ki}^{\mathrm{T}} \right\}^{-1} \left\{ \sum_{i=1}^{n_k} v_{ki}(\theta_k) Z_{ki} X_{ki}^{\mathrm{T}} \right\},$$

where $v_{ki}(\theta_k) = e^{\beta^{\mathrm{T}} X_{ki} + \gamma_k^{\mathrm{T}} Z_{ki}} / (1 + e^{\beta^{\mathrm{T}} X_{ki} + \gamma_k^{\mathrm{T}} Z_{ki}})^2$. Note that $\mathcal{I}_k$ depends on θ_k only through the $v_{ki}(\theta_k)$. Clearly, $v_{ki}(\theta_k) = \mathrm{pr}(Y_{ki} = 1)\{1 - \mathrm{pr}(Y_{ki} = 0)\}$, which is not sensitive to the value of θ_k unless $\mathrm{pr}(Y_{ki} = 1)$ is extreme. Thus, the results of meta-analysis using summary

statistics and using original data are generally similar, regardless of whether the effect sizes are equal or not.

Example 3 (Cox regression). The Cox (1972) proportional hazards model specifies that the hazard function of the survival time Y_{ki} conditional on covariates X_{ki} takes the form

$$\lambda(y|X_{ki}) = \lambda_{k0}(y)e^{\beta^{\mathrm{T}}X_{ki}}, \quad k = 1, \ldots, K; \ i = 1, \ldots, n_i,$$

where the $\lambda_{k0}(\cdot)$ are arbitrary baseline hazard functions. In the presence of right censoring, the data consist of $(\widetilde{Y}_{ki}, \Delta_{ki}, X_{ki})$ $(k = 1, \ldots, K; i = 1, \ldots, n_k)$, where $\widetilde{Y}_{ki} = \min(Y_{ki}, C_{ki})$, $\Delta_{ki} = I(Y_{ki} \leq C_{ki})$, C_{ki} is the censoring time on Y_{ki}, and $I(.)$ is the indicator function. The observed profile information matrix for β based on the kth study data is $\mathcal{I}_k(\beta) = \sum_{i=1}^{n_k} \Delta_{ki} V_k(\beta; \widetilde{Y}_{ki})$, where

$$V_k(\beta; y) = \frac{\sum_{j=1}^{n_k} I(\widetilde{Y}_{kj} \geq y)e^{\beta^{\mathrm{T}}X_{kj}}X_{kj}^{\otimes 2}}{\sum_{j=1}^{n_k} I(\widetilde{Y}_{kj} \geq y)e^{\beta^{\mathrm{T}}X_{kj}}} - \left\{ \frac{\sum_{j=1}^{n_k} I(\widetilde{Y}_{kj} \geq y)e^{\beta^{\mathrm{T}}X_{kj}}X_{kj}}{\sum_{j=1}^{n_k} I(\widetilde{Y}_{kj} \geq y)e^{\beta^{\mathrm{T}}X_{kj}}} \right\}^{\otimes 2}.$$

Note that $V_k(\beta; y)$ is an empirical covariance matrix of X and is not sensitive to the value of β. Thus, the results of meta-analysis using summary statistics and using original data are similar whether the effect sizes are equal or not.

3. NUMERICAL RESULTS

3.1. Simulation Studies

We conducted simulation studies to assess how well the asymptotic efficiency results of §2 approximate realistic situations. We mimicked meta-analysis of randomized clinical trials with a binary outcome and simulated data from the following logistic regression model

$$\mathrm{pr}(Y_{ki} = 1|X_{ki}) = \frac{e^{\alpha_k + \beta_k X_{ki}}}{1 + e^{\alpha_k + \beta_k X_{ki}}}, \quad k = 1, \ldots, K; \ i = 1, \ldots, n_k, \tag{6}$$

where X_{ki} is the treatment indicator, and the proportion of subjects receiving treatment 1 in the kth trial is p_k. The first set of simulation studies was focused on $K = 2$. We set $\alpha_1 = 0$ and $\alpha_2 = -1$, and chose various values of β_1, β_2, p_1, p_2, n_1 and n_2. For each combination of

the simulation parameters, we generated 1 million data sets; for each data set, we performed the two types of meta-analysis, i.e., the one based on summary statistics versus the one based on original data. The results are summarized in Table 1.

When the effect sizes are the same between the two studies, i.e., $\beta_1 = \beta_2$, both $\widehat{\beta}$ and $\widetilde{\beta}$ are virtually unbiased, and their standard errors are virtually identical to each other. Consequently, the two types of meta-analysis have essentially the same power. When the effect sizes are different between the two studies, i.e., $\beta_1 \neq \beta_2$, there are some noticeable differences between the two methods. In many cases, the meta-analysis based on original data appears slightly more powerful than the meta-analysis based on summary statistics. There are also cases in which the latter is slightly more powerful than the former.

In the second set of studies, we simulated K trials of size n from (6) with $\alpha_k = -1$, $\beta_k = 1$ and $p_k = 0.5$ $(k = 1, \ldots, K)$. For both meta-analysis of summary statistics and meta-analysis of original data, the K intercepts might be assumed to the same or allowed to be different. To impose a common intercept in meta-analysis of summary statistics, we used the bivariate summary statistics for the (α_k, β_k). Table 2 displays the relative efficiency results based on 10000 replicates that have no zero cell counts. Meta-analysis of summary statistics appears to be as efficient as or slightly more efficient than meta-analysis of original data when the two methods make the same modelling assumptions. For $n = 10$, meta-analysis of original data with a common intercept is a bit more efficient than meta-analysis of summary statistics with different intercepts.

3.2. Major Depression Data

Major depression is a complex common disease with enormous public health significance. The lifetime prevalence of this disorder is approximately 15% and is two-fold higher in women than men. Recently, a genome-wide association study was conducted to identify single nucleotide polymorphisms, SNPs, that are associated with major depression (Sullivan et al., 2009). Using a case-control sample of 1,738 cases and 1,802 controls, the investigators found strong signals in a region surrounding the gene piccolo, PCLO. The investigators

then attempted to replicate the results with five independent case-control samples. For our illustration, we exclude the two replication samples that do not have information on sex, which is an important predictor of major depression. The remaining three replication samples have 1907, 2489 and 2005 subjects, with a total of 3135 cases and 3266 controls.

A total of 30 SNPs in the PCLO region were genotyped in the replication samples. For each SNP, we fit the logistic regression model

$$\text{pr}(Y_{ki} = 1 | X_{ki}, Z_{ki}) = \frac{e^{\alpha_k + \beta X_{ki} + \gamma_k Z_{ki}}}{1 + e^{\alpha_k + \beta X_{ki} + \gamma_k Z_{ki}}}, \quad k = 1, \ldots, 3; \ i = 1, \ldots, n_k, \tag{7}$$

where Y_{ki} is the case-control status of the ith subject in the kth sample, and X_{ki} and Z_{ki} are the corresponding genotype score and sex indicator; the genotype score is the number of copies of the less frequent nucleotide of the SNP that the subject carries. The estimates of the genetic effects and their standard error estimates vary substantially among the three replication samples. For meta-analysis of summary statistics, the estimates of the genetic effects pre-adjusted for sex from the three samples are combined according to formula (1).

To perform meta-analysis of original data, we allow the α_k in (7) to be different among the three samples so as to reflect the unequal case-control ratios; the γ_k may be the same or different among the three samples. Figure 1 compares the results of meta-analysis based on summary statistics versus original data when the γ_k are allowed to be different. The two meta-analysis methods have the same modeling assumptions. The meta-analysis estimates of the genetic effects based on summary statistics are essentially identical to their counterparts based on original data, and so are the two sets of standard error estimates; therefore, the two types of meta-analysis yield virtually identical test statistics and p-values. Figure 2 compares the results of the two methods when the γ_k are assumed to be the same among the three replication samples in the meta-analysis of original data; this assumption is not used in the meta-analysis of summary statistics since three separate sub-models are fitted. Now the results are not always the same between the two types of analysis, but the differences are very small. Incidentally, the covariance estimates between the $\widehat{\beta}_k$ and $\widehat{\gamma}_k$ are virtually zero, so the differences should be small in light of the results of §2.2.

4. REMARKS

The theoretical results of the present paper are much broader than those of Olkin & Sampson (1998) and Mathew & Nordström (1999), even in the special setting considered by those authors; we have clarified the conditions for the equivalence results stated in those two papers and examined the consequences of violating the underlying assumptions. We have considered more general models for continuous response variables, as well as general parametric and semiparametric models for other response variables.

Our work has important practical implications. There is an ongoing debate on whether the benefits of using original data outweigh the extra cost of taking this approach. The statistical issues surrounding this debate have not been well understood. There is a common perception that analyzing original data is statistically more efficient than combining summary statistics. We have demonstrated both theoretically and numerically that this perception is false. Meta-analysis based on summary statistics will reduce resource utilization, simplify data collection and analysis, and avoid bias and efficiency loss due to exclusion of studies without original data.

By accessing original data, one can enhance comparability among studies with respect to inclusion/exclusion criteria, definitions of variables, creations of subgroups and adjustments of covariates, ensure estimation of the same parameter by the same statistical method, and perform model building and diagnostics. These benefits can still be achieved if all participating investigators follow a common set of guidelines on quality control and statistical analysis and then submit their summary statistics to the meta-analyst. Providing summary statistics is logistically much simpler than transferring original data. Indeed, human subject protection and other study policies often prohibit investigators from releasing original data.

One reason for obtaining original data is to use individual-level covariates. It is widely recognized that using study-level covariates can yield highly biased and inefficient meta-analysis (Berlin et al., 2002; Lambert et al., 2002). We have shown both theoretically and numerically that there is no bias or efficiency loss if the covariate adjustments are made

within each study and the covariate-adjusted effect estimates are combined via formula (2).

In our context, the asymptotics pertains to individual study sizes n_k. For commonly used parametric and semiparametric models, such as linear, logistic and Cox regression models, the asymptotic approximations are accurate even for small n_k. When the data for individual studies are very sparse, the parameter estimates may be undefined or unreliable. In that situation, analysis of original data will encounter the same difficulties if it is stratified by studies but will tend to be more stable if it is unstratified. The unstratified analysis can be very misleading if the underlying populations are different among studies.

We have focused on fixed-effects models, which assume a common value for the parameter of main interest among studies. An alternative approach is to employ random-effects models, in which the parameter of main interest is treated as a random variable with different realizations across studies (DerSimonian & Laird, 1986). It is technically more challenging to deal with random-effects models than fixed-effects models. Indeed, the properties of meta-analysis under random-effects models have not be rigorously investigated. Our preliminary investigations reveal that the conclusions of §2 hold for random-effects models under certain conditions. The results will be communicated in a separate report.

ACKNOWLEDGEMENT

This research was supported by the U. S. National Institutes of Health. The authors thank the Editor and two referees for helpful comments.

APPENDIX A

Some useful matrix results

Lemma 1. For any matrices $A_{p\times q}$, $B_{q\times q}$, $C_{p\times q}$ and $D_{q\times q}$ with $B > 0$ and $D > 0$,

$$AB^{-1}A^{\mathrm{T}} + CD^{-1}C^{\mathrm{T}} \geq (A+C)(B+D)^{-1}(A+C)^{\mathrm{T}}. \qquad (A.1)$$

The equality holds if and only if $AB^{-1} = CD^{-1}$.

Proof. Since $B > 0$ and $D > 0$, we can find a non-singular matrix P such that $B = P\text{diag}\{\lambda_1, \ldots, \lambda_q\}P^{\mathrm{T}}$ and $D = P\text{diag}\{\mu_1, \ldots, \mu_q\}P^{\mathrm{T}}$, where $\lambda_i > 0$ and $\mu_i > 0$ $(i = 1, \ldots, q)$. By redefining A as $A(P^{\mathrm{T}})^{-1}$ and C as $C(P^{\mathrm{T}})^{-1}$, it suffices to prove the lemma when $B = \text{diag}\{\lambda_1, \ldots, \lambda_q\}$ and $D = \text{diag}\{\mu_1, \ldots, \mu_q\}$. Let a_i be the ith column of A and c_i be the ith column of C. Then (A.1) becomes

$$\sum_{i=1}^{q} \left(\lambda_i^{-1} a_i a_i^{\mathrm{T}} + \mu_i^{-1} c_i c_i^{\mathrm{T}}\right) \geq \sum_{i=1}^{q} (\lambda_i + \mu_i)^{-1} (a_i + c_i)(a_i + c_i)^{\mathrm{T}}.$$

We wish to show that $\lambda_i^{-1} a_i a_i^{\mathrm{T}} + \mu_i^{-1} c_i c_i^{\mathrm{T}} \geq (\lambda_i + \mu_i)^{-1}(a_i + c_i)(a_i + c_i)^{\mathrm{T}}$ or equivalently $\lambda_i^2 c_i c_i^{\mathrm{T}} + \mu_i^2 a_i a_i^{\mathrm{T}} \geq \lambda_i \mu_i (a_i c_i^{\mathrm{T}} + c_i a_i^{\mathrm{T}})$. The desired inequality holds if, for any x,

$$(\lambda_i x^{\mathrm{T}} c_i)^2 + (\mu_i x^{\mathrm{T}} a_i)^2 \geq 2\lambda_i \mu_i (x^{\mathrm{T}} a_i)(x^{\mathrm{T}} c_i),$$

which is obvious from the Cauchy-Schwartz inequality. The foregoing inequality becomes an equality if and only if $\lambda_i c_i = \mu_i a_i$. Thus, the equality in (A.1) holds if and only if $AB^{-1} = CD^{-1}$.

APPENDIX B

Consistency of MLE under local alternatives

By the profile likelihood theory (Murphy & van der Vaart, 2000),

$$\log pl_k(\beta) = \log pl_k(\widehat{\beta}_k) - \frac{1}{2}(\beta - \widehat{\beta}_k)^{\mathrm{T}} \mathcal{I}_k(\beta_k)(\beta - \widehat{\beta}_k) + o_p(n^{1/2}\|\beta - \beta_k\| + 1)^2$$

for β in a neighborhood of the true value of β_k. Denote the true value of β by β_0. Because $\widehat{\beta}_k - \beta_k = O_p(n^{-1/2})$, $\beta_k - \beta_0 = O(n^{-1/2})$ and $\mathcal{I}_k(\beta_k) = \mathcal{I}_k(\widehat{\beta}_k) + o_p(n)$, we have

$$\sum_{k=1}^{K} \log pl_k(\beta) = \sum_{k=1}^{K} \log pl_k(\widehat{\beta}_k) - \frac{1}{2} \sum_{k=1}^{K} (\beta - \widehat{\beta}_k)^{\mathrm{T}} \mathcal{I}_k(\widehat{\beta}_k)(\beta - \widehat{\beta}_k) + o_p(n^{1/2}\|\beta - \beta_0\| + 1)^2$$

for β in a neighborhood of β_0. By definition, $\widehat{\beta}$ maximizes

$$-\frac{1}{2} \sum_{k=1}^{K} (\beta - \widehat{\beta}_k)^{\mathrm{T}} \mathcal{I}_k(\widehat{\beta}_k)(\beta - \widehat{\beta}_k).$$

It then follows from the Taylor series expansion that

$$-\frac{1}{2}\sum_{k=1}^{K}(\widehat{\beta}-\widehat{\beta}_k)^{\mathrm{T}}\mathcal{I}_k(\widehat{\beta}_k)(\widehat{\beta}-\widehat{\beta}_k) \geq -\frac{1}{2}\sum_{k=1}^{K}(\beta-\widehat{\beta}_k)\mathcal{I}_k(\widehat{\beta}_k)(\beta-\widehat{\beta}_k)+\alpha n\|\widehat{\beta}-\beta\|^2$$

for some positive constant α. Thus,

$$\log pl(\widehat{\beta})-\log pl(\beta) \geq \alpha n\|\widehat{\beta}-\beta\|^2+o_p(n\|\beta-\beta_0\|^2+n\|\widehat{\beta}-\beta_0\|^2+1).$$

Because $\|\widehat{\beta}-\beta_0\| = O(n^{-1/2})$, the foregoing inequality implies that $pl(\widehat{\beta}) > pl(\beta)$ for any β such that $\|\beta-\widehat{\beta}\| = n^{-1/2}M$ for a large M. Hence, there exists a local maximum within the $n^{-1/2}M$-neighborhood of $\widehat{\beta}$. We define that estimator as $\widetilde{\beta}$ and conclude that $\widetilde{\beta}-\beta_0 = O_p(n^{-1/2})$.

REFERENCES

Berlin, J. A., Santanna, J., Schmid, C .H., Szczech, L.A., and Feldman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statist. Med.* **21**, 371–87.

Chalmers, I., Sandercrock, P. & Wennberg, J. (1993). The Cochrane collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals New York Acad. Sci.* **703**, 156–65.

Cox, D. R. (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–88.

Lambert, P. C, Sutton, A. J., Abrams, K. R., and Jones, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J. Clin. Epidemiol.* **55**, 86–94.

Lohmueller, K. E., Pearce, C. L., Pike, et al. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* **33**, 177–82.

Kavvoural, F. K. & Ioannidis, J. P. A. (2008). Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Human Genetics* **123**, 1–14.

Mathew, T. & Nordstrom, K. (1999). On the equivalence of meta-analysis using literature and using individual patient data. *Biometrics* **55**, 1221–3.

Murphy, S. A. & van der Vaart, A. W. (2000). On the profile likelihood. *J. Am. Statist. Assoc.* **95**, 449–65.

Olkin, I. & Sampson, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics* **54**, 317–22.

Sullivan, P. F., de Geus, E. J. C., Willemsen, G., et al. (2009). Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Molecular Psychiatry* **14**, 359–75.

Sutton, A. J., Abrams, K. R., Jones, Sheldon, T. A. & Song, F. (2000). *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley & Sons.

The Psychiatric GWAS Consortium Steering Committee. (2009). A framework for interpreting genome-wide association studies of psychiatric disorders. *Molecular Psychiatry* **14**, 10–17.

Whitehead, A. (2002). *Meta-Analysis of Controlled Clinical Trials*. Chichester: Wiley & Sons.

Zeggini, E., Scott, L. J., Saxena, R., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics* **40**, 638–45.

Table 1. Mean parameter estimates, standard errors (SE) and powers at the .05 *significance level for meta-analysis based on summary statistics versus original data*

						Summary statistics			Original data		
β_1	β_2	p_1	p_2	n_1	n_2	Mean	SE	Power	Mean	SE	Power
.5	.5	.5	.5	200	200	.503	.211	.665	.504	.212	.669
				200	400	.502	.174	.827	.503	.174	.828
				400	200	.501	.170	.842	.502	.171	.843
		.2	.5	200	200	.502	.236	.566	.506	.237	.575
				200	400	.501	.186	.769	.504	.187	.774
				400	200	.502	.198	.722	.504	.199	.726
		.5	.2	200	200	.504	.229	.595	.503	.230	.593
				200	400	.502	.195	.729	.501	.196	.729
				400	200	.504	.179	.803	.502	.180	.801
.2	.8	.5	.5	200	200	.483	.205	.643	.488	.207	.656
				200	400	.586	.169	.933	.589	.171	.935
				400	200	.383	.166	.629	.388	.167	.643
		.2	.5	200	200	.553	.231	.666	.559	.234	.676
				200	400	.645	.183	.945	.649	.185	.945
				400	200	.447	.193	.633	.453	.196	.647
		.5	.2	200	200	.427	.228	.474	.427	.231	.471
				200	400	.530	.194	.786	.530	.196	.779
				400	200	.339	.178	.485	.339	.179	.481
.8	.2	.5	.5	200	200	.521	.214	.678	.526	.217	.687
				200	400	.417	.176	.652	.422	.178	.664
				400	200	.619	.172	.949	.622	.174	.950
		.2	.5	200	200	.443	.236	.438	.457	.240	.476
				200	400	.350	.188	.442	.362	.190	.477
				400	200	.550	.198	.780	.559	.201	.798
		.5	.2	200	200	.588	.226	.710	.580	.229	.710
				200	400	.479	.192	.668	.476	.195	.678
				400	200	.673	.178	.961	.665	.180	.958

Table 2. Relative efficiency of using summary statistics versus original data in meta-analysis †

	$K = 20$			$K = 50$			$K = 100$		
n	RE1	RE2	RE3	RE1	RE2	RE3	RE1	RE2	RE3
10	1.108	0.968	1.093	1.113	0.966	1.097	1.112	0.966	1.098
20	1.161	1.010	1.112	1.160	1.006	1.113	1.156	1.004	1.110
50	1.092	1.017	1.065	1.096	1.021	1.068	1.096	1.022	1.069
100	1.049	1.007	1.030	1.047	1.007	1.029	1.046	1.007	1.030

† RE1 is the relative efficiency of meta-analysis of summary statistics with a common intercept to meta-analysis of original data with a common intercept; RE2 is the relative efficiency of meta-analysis of summary statistics with different intercepts to meta-analysis of original data with a common intercept; RE3 is the relative efficiency of meta-analysis of summary statistics with different intercepts to meta-analysis of original data with different intercepts.

Figure Legends

Figure 1 Meta-analysis of genetic effects based on original data versus summary statistics for the major depression study: estimates of log odds ratios, standard error estimates, test statistics, and $-\log_{10}$(p-values) are shown for 30 SNPs in the PCLO region; the effects of sex are allowed to be different among the three replication samples.

Figure 2 Meta-analysis of genetic effects based on original data versus summary statistics for the major depression study: estimates of log odds ratios, standard error estimates, test statistics, and $-\log_{10}$(p-values) are shown for 30 SNPs in the PCLO region; the meta-analysis of original data assumes a common effect of sex among the three replication samples, whereas the meta-analysis based on summary statistics does not.

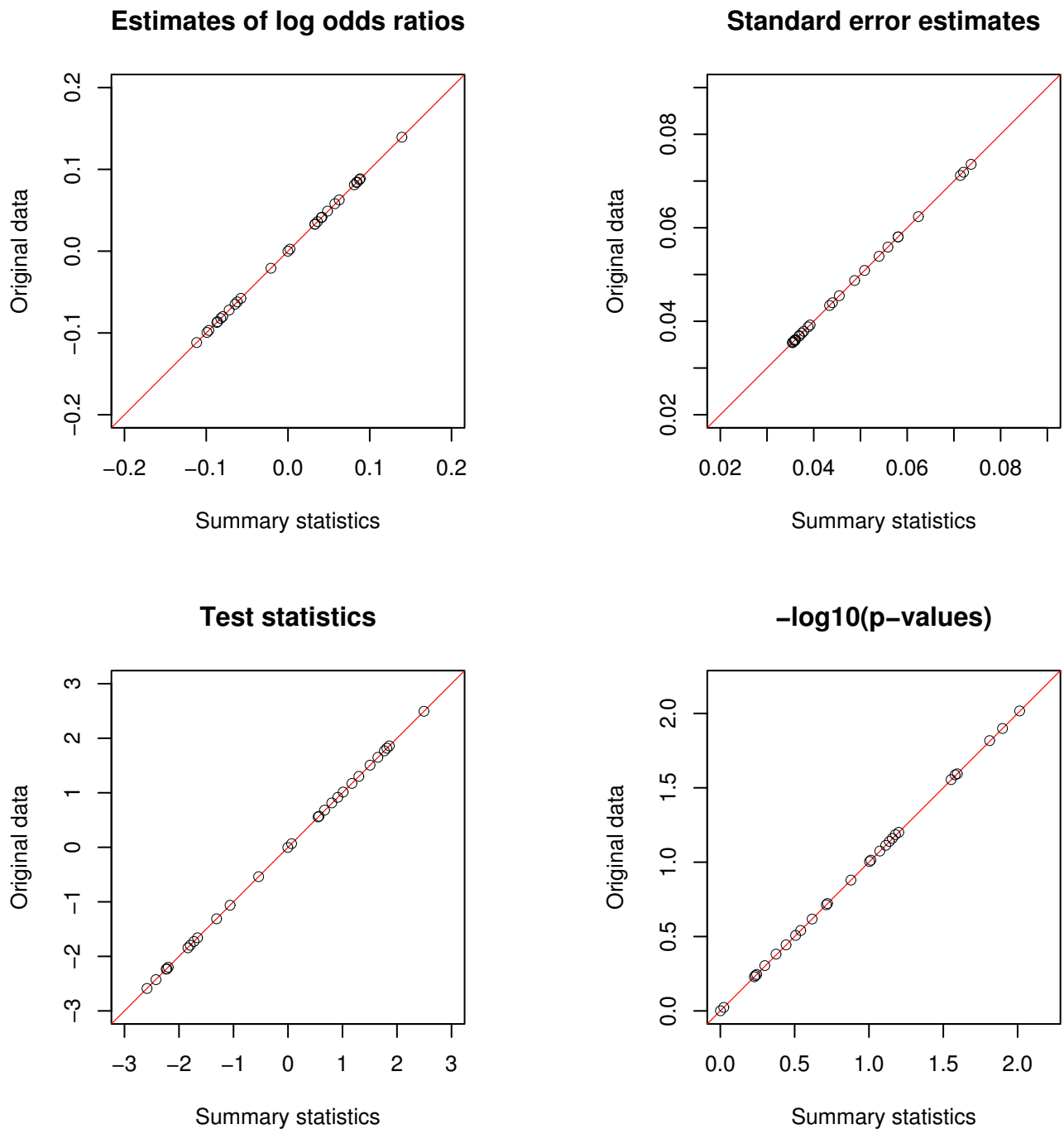
Estimates of log odds ratios
Original data
Summary statistics
−0.2
−0.1
0.0
0.1
0.2
Standard error estimates
Original data
Summary statistics
0.02
0.04
0.06
0.08
Test statistics
Original data
Summary statistics
−3
−2
−1
0
1
2
3
−log10(p−values)
Original data
Summary statistics
0.0
0.5
1.0
1.5
2.0

Figure 1:

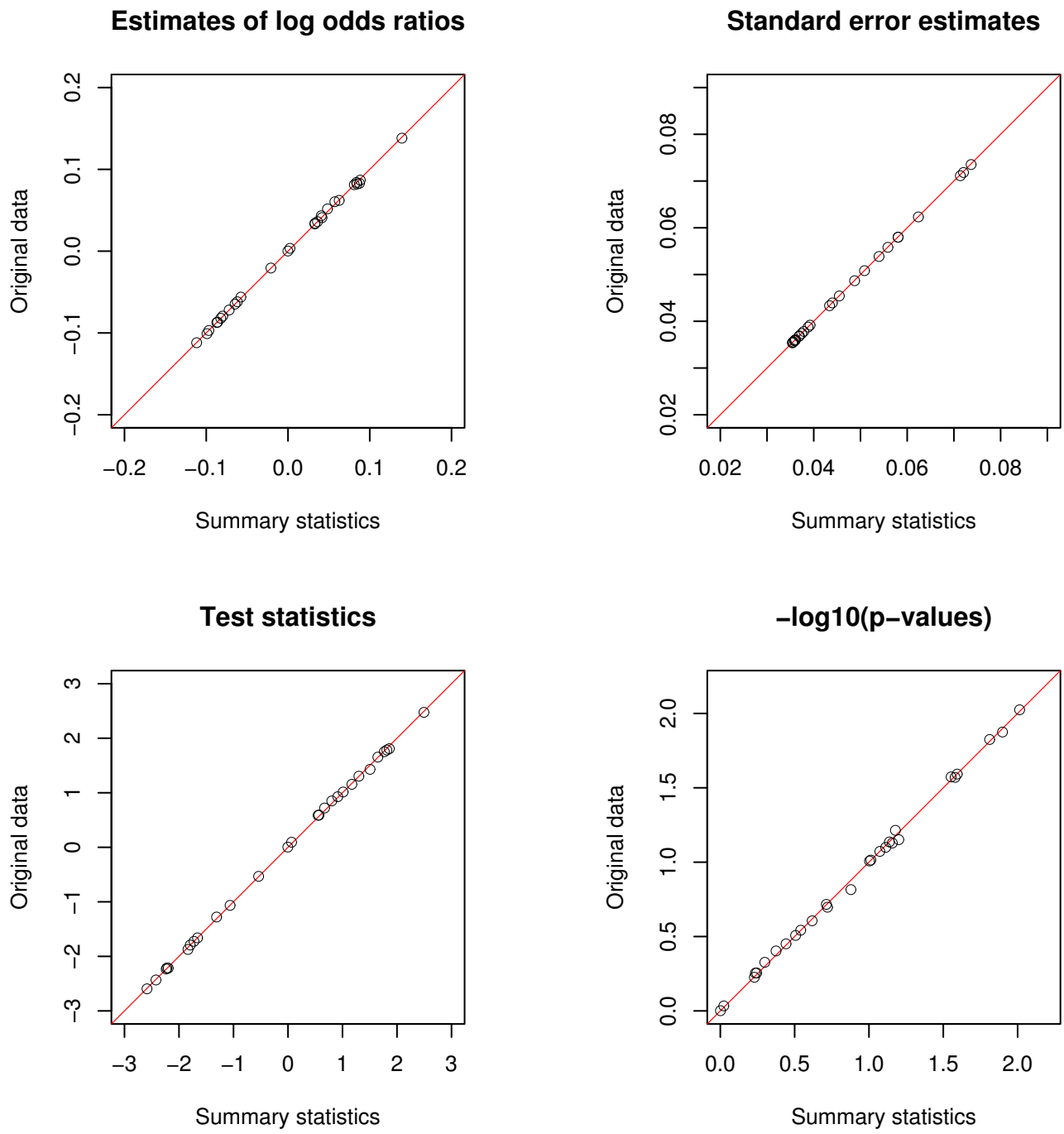
Estimates of log odds ratios
Original data
Summary statistics
−0.2
−0.1
0.0
0.1
0.2
Standard error estimates
0.02
0.04
0.06
0.08
Test statistics
−3
−2
−1
0
1
2
3
−log10(p−values)
0.0
0.5
1.0
1.5
2.0

Figure 2: