Biometrika (2019), **106**, 3, *pp.* 567–585
Printed in Great Britain
doi: 10.1093/biomet/asz030
Advance Access publication 13 July 2019

Generalized meta-analysis for multiple regression models across studies with disparate covariate information

BY PROSENJIT KUNDU, RUNLONG TANG AND NILANJAN CHATTERJEE

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 21205, U.S.A.

pkundu@jhu.edu rtang15@jhu.edu nchatte2@jhu.edu

SUMMARY

Meta-analysis is widely popular for synthesizing information on common parameters of interest across multiple studies because of its logistical convenience and statistical efficiency. We develop a generalized meta-analysis approach to combining information on multivariate regression parameters across multiple studies that have varying levels of covariate information. Using algebraic relationships among regression parameters in different dimensions, we specify a set of moment equations for estimating parameters of a maximal model through information available from sets of parameter estimates for a series of reduced models from the different studies. The specification of the equations requires a reference dataset for estimating the joint distribution of the covariates. We propose to solve these equations using the generalized method of moments approach, with the optimal weighting of the equations taking into account uncertainty associated with estimates of the parameters of the reduced models. We describe extensions of the iterated reweighted least-squares algorithm for fitting generalized linear regression models using the proposed framework. Based on the same moment equations, we also develop a diagnostic test for detecting violations of underlying model assumptions, such as those arising from heterogeneity in the underlying study populations. The proposed methods are illustrated with extensive simulation studies and a real-data example involving the development of a breast cancer risk prediction model using disparate risk factor information from multiple studies.

Some key words: Data integration; Empirical likelihood; Generalized method of moments; Meta-analysis; Missing data; Semiparametric inference.

1. INTRODUCTION

In many areas of applications, including observational epidemiological studies, clinical trials and modern genome-wide association studies, meta-analysis is widely used to synthesize information on underlying common parameters of interest across multiple studies (Dersimonian & Laird, 1986, 2015; Ioannidis, 2005; Kavvoura & Ioannidis, 2008). The popularity of meta-analysis stems from the fact that it can be performed based only on estimates of model parameters and standard errors, avoiding various logistical, ethical and privacy concerns associated with accessing the individual-level data required in pooled analysis. Moreover, in many common settings, it can be shown that under reasonable assumptions, meta-analysed estimates of model parameters are asymptotically as efficient as those obtained from pooled analysis (Olkin & Sampson, 1998; Mathew & Nordstrom, 1999; Lin & Zeng, 2010). In fact, meta-analysis methods are now being used in divide-and-conquer approaches to big data, even when individual-level data are

potentially available, because of the daunting computational task of model fitting with extremely large sample sizes (Jordan, 2013; Fan et al., 2014; Chun et al., 2015).

In this article, we study the problem of multivariate meta-analysis in the setting of parametric regression modelling of an outcome given a set of covariates. In standard settings, if estimates of multivariate parameters for an underlying common regression model and their associated covariances are available across all the studies, then meta-analysis can be performed by taking the inverse variance-covariance weighted average of the vector of regression coefficients (van Houwelingen et al., 2002; Ritz et al., 2008; Jackson et al., 2011). In many applications, a typical problem is that different studies include different, but possibly overlapping, sets of covariates. In a large consortium of epidemiological studies, for example, some key risk factors will be measured across all the studies, but inevitably there will be potentially important covariates that are measured only in some, but not all, of the studies. It is also possible that some covariates will be measured at a more detailed level or with a finer instrument in some studies than in others. Disparate sets of covariates across studies mean that standard meta-analysis is applicable only to the development of models limited to a core set of variables that are measured in the same way across all the studies.

We propose a generalized meta-analysis method, which we call GENMETA, for building rich models using information on model parameters across studies with disparate covariate information. Our approach is built upon a fundamental mathematical relationship, presented in our recent work (Chatterjee et al., 2016), between parameters of two regression models in different dimensions. In the present article, we use this mathematical relationship to develop a general framework for combining information on parameters of various models of different dimensions within the generalized method of moments framework (Hansen, 1982; Imbens, 2002). We develop an iterated reweighted least-squares algorithm that allows stable and speedy computation of estimates. The proposed method requires access to a reference dataset for estimating the joint distribution of the covariates in a nonparametric fashion. We show how the reference dataset can be used to derive an optimal estimator and the associated variances and covariances, even when entire variance-covariance matrices for model parameter estimates may not be obtainable from individual studies.

2. MODELS AND METHODS

2.1. *Model formulation*

Suppose that we have parameter estimates $\hat{\theta}_k$ and associated estimates of their covariance matrices S_k from K independent studies that have fitted reduced regression models, of the form $g_k(Y \mid X_{A_k}; \theta_k)$, where Y is a common underlying outcome of interest, but the vector of covariates X_{A_k} is potentially distinct across different studies. Let X be the set of covariates used in at least one study, and assume that the true distribution of Y given X can be specified by a maximal regression model $f(Y \mid X; \beta)$. Our goal is to estimate and make inference about β^*, the true value of β, based on summary-level information, $(\hat{\theta}_k, S_k)$, from the K studies.

In the proposed set-up it is possible, but not necessary, that some of the studies will have information on all covariates to fit the maximal model by themselves. Under certain study designs, such as multi-phase designs (Breslow & Cain, 1988; Breslow & Holubkov, 1997; Scott & Wild, 1997; Whittemore, 1997) and the partial questionnaire design (Wacholder & Carroll, 1994), data could be partitioned into independent sets such that the maximal model can be fitted on some sets and various reduced models fitted on others. The maximal model $f(Y \mid X; \beta)$ and the reduced models $g_k(Y \mid X_{A_k}; \theta_k)$ may have different parametric forms, such as logistic and probit models

when Y is a binary disease outcome. This set-up also allows incorporation of covariates which may be measured more accurately, or in a more refined manner in some studies than in others. For example, different studies may include two types of measurements, say Z_1 and Z_2, for the same covariate, with Z_2 being a more refined measurement. In this case the different reduced models may include Z_1 or Z_2, but we require that the reference dataset include both Z_1 and Z_2. In the maximal model, we can force Y to be independent of Z_1 given Z_2 by setting the regression parameters associated with Z_1 to zero.

If all of the reduced models are the same, i.e., all the studies have the same covariate information, then $X_k = X$, $\theta_k = \beta$ and $g_k = f$ for each k, and the common parameter of interest β^* can be efficiently estimated by the fixed-effect meta-analysis estimator $\hat{\beta}_{\text{meta}} = \sum_{k=1}^{K}(\sum_{k=1}^{K} S_k^{-1})^{-1} S_k^{-1}\hat{\theta}_k$, the variance of which can in turn be estimated by $\hat{\Sigma}_{\text{meta}} = (\sum_{k=1}^{K} S_k^{-1})^{-1}$ (van Houwelingen et al., 2002; Ritz et al., 2008; Jackson et al., 2011).

2.2. *A special case involving the linear regression model*

As readers may have difficulty comprehending how it is possible to estimate parameters of the maximal model when no single study may have ascertained Y and all components of X simultaneously, here we give a linear model example to help develop insight into the problem. Suppose that one is interested in developing a multiple linear regression model for Y based on a set of covariates X in the form

$$Y = \alpha + \sum_{k=1}^{K} \beta_k X_k + \epsilon,$$

where it is further assumed that $\epsilon \sim N(0, \sigma^2)$. Without loss of generality, we assume that all the variables $Y, X_1, \ldots, X_K$ are standardized to have mean 0 and variance 1. Under this model, the population parameter $\beta = (\beta_1, \ldots, \beta_K)^{\mathrm{T}}$ can be expressed as $\beta = E(X^{\mathrm{T}}X)^{-1}E(X^{\mathrm{T}}Y) = R^{-1}E(X^{\mathrm{T}}Y)$, where R is the population correlation matrix of X. Now, suppose we have no data available on Y and multivariate X on the same sample, but we do have estimates available for parameters θ_k $(k = 1, \ldots, K)$ for univariate linear regression models of the form

$$Y = \theta_k X_k + \psi_k.$$

From above, $\theta_k = E(X_k Y)$, and so $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)$ provides an estimate of the cross product terms $E(X^{\mathrm{T}}Y)$, which are required in estimating β. Further, if we have a reference dataset which contains information on multivariate X, but is not required to be linked to Y, it can be used to estimate R, as $\hat{R}$ say, and a consistent estimate of β can then be obtained simply as $\hat{\beta} = \hat{R}^{-1}\hat{\theta}$. Thus, it is possible to estimate parameters of a multiple regression model using information on parameters of a series of univariate regression models and a reference dataset. In fact, this observation that information on univariate regression parameters, known as summary-level statistics, can be used to reconstruct estimates of parameters of multivariate regression models has revolutionized the field of statistical genetics. Recently, a great variety of methods have been developed for inference on parameters underlying multivariate regression models that utilize widely available summary-level results from large genome-wide association studies and reference datasets to estimate linkage disequlibrium across genetic markers (Yang et al., 2012; Bulik-Sullivan et al., 2015; Zhu et al., 2016; Pasaniuc & Price, 2017). In the following, we describe a more general statistical formulation of the problem that allows consideration of nonlinear models and use of information from arbitrary types of reduced models, as opposed to simply univariate models.

2.3. *Generalized meta-analysis*

The key idea underlying the proposed generalized meta-analysis is that we convert information on parameters from reduced models into a set of equations that are informative about the parameters of the maximal model. We will make the following assumptions: (i) the same probability law for (Y, X) holds for all the underlying populations; (ii) $f(Y \mid X; \beta)$ is a correctly specified model for the conditional distribution of $(Y \mid X)$; and (iii) we have a reference dataset to estimate empirically the joint distribution of all the factors included in X.

Here we assume that all the studies employ a random sampling design and that the same probability law for (Y, X) holds for all of the underlying populations. Let $s_k(y \mid x_{A_k}; \theta_k) = \partial \log g_k(y \mid x_{A_k}; \theta_k)/\partial\theta_k$ be the score function of the kth reduced model, and write $u_k(x; \beta, \theta_k) = \int s_k(y \mid x_{A_k}; \theta_k) f(y \mid x; \beta)\, \mathrm{d}y$. Assume that $\hat{\theta}_k$ is the maximum likelihood estimator from the kth study, and denote by θ_k^* the asymptotic limit of $\hat{\theta}_k$. Irrespective of whether the reduced models are correct, $E_{\mathrm{pr}^*}\{s_k(Y \mid X_{A_k}; \theta_k^*)\} = 0$ holds, where pr* denotes the true probability law. Assuming that the maximal model is correctly specified, we can write $\mathrm{pr}^*(Y, X_{A_k}) = \int_{X \setminus X_{A_k}} f(Y \mid X; \beta^*)\, \mathrm{d}F^*(X)$. Hence, a general equation describing the relationship between β^* and θ_k^* is of the form (Chatterjee et al., 2016)

$$\int u_k(x; \beta^*, \theta_k^*)\, \mathrm{d}F^*(x) = 0.$$

As we may not have individual-level data from the studies, these equations cannot be evaluated directly. Instead, we assume that we have a reference sample of size n, independent of the study samples, on which measurements of X are available. The reference sample need not be linked with the outcome Y of interest, and its sample size can be fairly modest compared with the study sample sizes.

With $\hat{\theta}_k$ from the studies and the reference sample $\{X_i\}_{i=1}^n$, we can set up the estimating equations $U_n(\beta, \hat{\theta}) = (1/n)\sum_{i=1}^n U(X_i; \beta, \hat{\theta}) = 0$, where $U(x; \beta, \theta) = \{u_1^{\mathrm{T}}(x; \beta, \theta_1), \ldots, u_K^{\mathrm{T}}(x; \beta, \theta_K)\}^{\mathrm{T}}$, $\hat{\theta} = (\hat{\theta}_1^{\mathrm{T}}, \ldots, \hat{\theta}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\theta = (\theta_1^{\mathrm{T}}, \ldots, \theta_K^{\mathrm{T}})^{\mathrm{T}}$. Denote the dimensions of θ_k and β by d_k and p, respectively. Because the number of equations $d = \sum_{k=1}^K d_k$ can be larger than the number of unknown parameters p, it may be that the estimating equations cannot be solved exactly. Based on the generalized method of moments, we propose the following generalized meta-analysis estimator of $\beta^\star$: $\hat{\beta} = \arg\min_\beta Q_{\hat{C}}(\beta)$ where $Q_{\hat{C}}(\beta) = U_n(\beta, \hat{\theta})^{\mathrm{T}} \hat{C} U_n(\beta, \hat{\theta})$, with $\hat{C}$ being a positive-semidefinite weighting matrix. Using the well-established theory of generalized method of moments (Hansen, 1982; Engle & McFadden, 1994), we derive the asymptotic properties of our estimator. Assume that the study summary statistics $\hat{\theta}_k$ are independent across studies, that $n_k^{1/2}(\hat{\theta}_k - \theta_k^*) \to N(0, \Sigma_k)$ in distribution, that $\lim_{n\to\infty} n_k/n = c_k > 0$ for each k, and that the reference sample is independent of the study samples. Let $\Gamma = E\{\partial U(X; \beta, \theta^\star)/\partial\beta|_{\beta=\beta^*}\}$, $\Delta = E\{U(X; \beta^*, \theta^\star) U^{\mathrm{T}}(X; \beta^*, \theta^\star)\}$ and $\Lambda = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_K)$, where $\Lambda_k = (1/c_k) W_k \Sigma_k W_k^{\mathrm{T}}$ with $W_k = E\{\partial u_k(X; \beta^*, \theta_k)/\partial\theta_k\}|_{\theta_k=\theta_k^*}$ for $k = 1, \ldots, K$.

THEOREM 1 (Consistency and asymptotic normality of $\hat{\beta}$). *Suppose that the positive-semidefinite weighting matrix $\hat{C}$ tends to C in probability. Then, under Assumptions A1–A4 in the Appendix, $\hat{\beta} \to \beta^*$ in probability. Further, if β^* is an interior point, then under the additional Assumptions A5–A9 in the Appendix, $n^{1/2}(\hat{\beta} - \beta^*)$ converges in distribution to the normal distribution $N\{0, (\Gamma^{\mathrm{T}} C \Gamma)^{-1} \Gamma^{\mathrm{T}} C (\Delta + \Lambda) C \Gamma (\Gamma^{\mathrm{T}} C \Gamma)^{-1}\}$.*

The optimal C that minimizes the above asymptotic covariance matrix is $C_{\text{opt}} = (\Delta + \Lambda)^{-1}$, and the corresponding optimal asymptotic covariance matrix is $\{\Gamma^{\mathrm{T}}(\Delta + \Lambda)^{-1}\Gamma\}^{-1}$. Because C_{opt} itself depends on unknown underlying parameters, it requires iterative evaluation. In our applications, we first evaluate an initial estimator with a simple choice of $\hat{C}$, such as the identity matrix. We then obtain the iterated estimator by continuing to set $\hat{C} = \hat{C}_{\text{opt}}$ based on the latest parameter estimate until convergence. By Theorem 1, $\hat{\beta}$ with C_{opt} approximately follows a Gaussian distribution with mean β^* and covariance matrix

$$\left[\Gamma^{\mathrm{T}}\left\{\frac{1}{n}\Delta + \operatorname{diag}\left(\frac{1}{n_1}W_1\Sigma_1 W_1^{\mathrm{T}}, \ldots, \frac{1}{n_K}W_K\Sigma_K W_K^{\mathrm{T}}\right)^{-1}\right\}\Gamma\right]^{-1}, \tag{1}$$

which indicates that the precision of our estimator depends on the size of the reference sample, n, as well as on the sample sizes of the studies, n_k. However, as we will see in § 3, the study sample sizes are the dominant factor controlling the precision of our estimator, and with the n_k fixed the precision quickly reaches a plateau as a function of n.

For the implementation of the optimal generalized meta-analysis and the variance estimation of any of the generalized meta-analysis estimators, one needs to have valid estimates of Λ_k, which depend on Σ_k, the asymptotic covariance matrices of the estimates of the reduced model parameters. Ideally, the studies should provide robust estimates of the covariance matrices, such as the sandwich covariance estimators, so that they are valid irrespective of whether the underlying reduced models are correctly specified or not. In practice, however, while we expect some kind of estimate of standard errors of the individual parameters to be available from a study, obtaining the desired robust estimate of the entire covariance matrix can be difficult. When no estimate of Σ_k is available from the kth study, one can take advantage of the reference sample to estimate it by $\hat{\Sigma}_k^{\text{ref}} = \hat{J}^{-1}\hat{V}\hat{J}^{-1}$, where $\hat{J} = P_n[E_{Y|X}\{\nabla_{\theta_k} s_k(\theta_k)\}]|_{\theta_k=\hat{\theta}_k}$ and $\hat{V} = P_n[E_{Y|X}\{s_k(\theta_k)s_k(\theta_k)^{\mathrm{T}}\}]|_{\theta_k=\hat{\theta}_k}$ with $s_k(\hat{\theta}_k) = s_k(Y \mid X_{A_k}; \theta_k)|_{\theta_k=\hat{\theta}_k}$; here $\hat{\theta}_k$ is a consistent estimator of $\theta_k^\star$, $\hat{E}_{Y|X}$ is the expectation with respect to the distribution of $Y \mid X$ with $\beta^\star$ replaced by a consistent estimator $\hat{\beta}$, and P_n is the empirical measure with respect to the reference sample. Further, assuming $E_{Y|X}\{\nabla_{\theta_k} s_k(\theta_k)\}|_{\theta_k=\theta_k^*} = \nabla_{\theta_k} E_{Y|X}\{s_k(\theta_k)\}|_{\theta_k=\theta_k^*}$, it follows that $\Lambda_k = (1/c_k)E_{(Y,X)}\{s_k(\theta_k)s_k(\theta_k)^{\mathrm{T}}\}|_{\theta_k=\theta_k^*}$, which can be estimated by $\hat{\Lambda}_k^{\text{ref}} = (1/c_k)P_n[E_{Y|X}\{s_k(\theta_k)s_k(\theta_k)^{\mathrm{T}}\}]|_{\theta_k=\hat{\theta}_k}$. For example, suppose that $Y \mid X$ and $Y \mid X_{A_k}$ follow logistic distributions with parameters $\beta^\star$ and θ_k, respectively. Write $X = (1, X^{\mathrm{T}})^{\mathrm{T}}$ and $X_{A_k} = (1, X_{A_k}^{\mathrm{T}})^{\mathrm{T}}$. Then

$$\begin{aligned}\hat{\Lambda}_k^{\text{ref}} = \frac{1}{c_k}P_n\Big(\big[\{1+\exp(X_{A_k}^{\mathrm{T}}\hat{\theta}_k)\}^{-2}\{1+\exp(-X^{\mathrm{T}}\beta)\}^{-1} \\ + \{1+\exp(-X_{A_k}^{\mathrm{T}}\hat{\theta}_k)\}^{-2}\{1+\exp(X^{\mathrm{T}}\hat{\beta})\}^{-1}\big]X_{A_k}X_{A_k}^{\mathrm{T}}\Big).\end{aligned} \tag{2}$$

In § 3 we will study the properties of our generalized meta-analysis estimators using either covariance matrices estimated from studies or the reference sample.

It is illuminating to explore the connection between our proposed approach and standard meta-analysis when all of the reduced models are identical to the maximal model, that is, when $\theta_k^* = \beta^*$, $X_{A_k} = X$ and $g_k = f$ for each k. In this set-up, the moment vector evaluated at the true parameters becomes zero for each study, i.e., $u_k(X; \beta^*, \theta_k^*) = u_k(X; \beta^*, \beta^*) = 0$. This simplification implies $\Delta = 0$, and hence the optimal weighting matrix is $C_{\text{opt}} = \Lambda^{-1} = \operatorname{diag}(c_1\Sigma, \ldots, c_K\Sigma)$, where Σ

is the inverse of the Fisher's information matrix of f. Denote by $\hat{\beta}_{\text{opt}}$ the GENMETA estimator with a consistent estimator of C_{opt}. Then, by arguments similar to those in the proof of Theorem 1, $\hat{\beta}_{\text{opt}}$ can be expressed as

$$\hat{\beta}_{\text{opt}} = \hat{\beta}_{\text{meta}} + o_{\text{p}}(1/n^{1/2}),$$

which implies that $\hat{\beta}_{\text{opt}}$ and $\hat{\beta}_{\text{meta}}$ are asymptotically equivalent in terms of limiting distributions.

2.4. *Generalized linear model and iterated reweighted least-squares algorithm*

Our generalized meta-analysis computation involves minimization of a quadratic form, $Q_C(\beta) = U_n^{\text{T}}(\beta, \hat{\theta}) C U_n(\beta, \hat{\theta})$, with a known weighting matrix C. In this subsection we derive the iterated reweighted least-squares algorithm for minimizing the quadratic form, assuming that the maximal and reduced models belong to the class of generalized linear models (McCullagh & Nelder, 1989). Specifically, the densities of $Y \mid X$ and $Y \mid X_{A_k}$ are of the forms $\exp(\{1/a(\phi)\}[yh(x^{\text{T}}\beta^\star) - b\{h(x^{\text{T}}\beta^\star)\}] + c(y;\phi))$ and $\exp(\{1/a(\phi_k)\}[yh(x_{A_k}^{\text{T}}\theta_k) - b\{h(x_{A_k}^{\text{T}}\theta_k)\}] + c(y;\phi_k))$, respectively, where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, $h(\cdot) = b'^{-1}\{g^{-1}(\cdot)\}$ with g a monotone and differentiable link function, and ϕ and ϕ_k are the dispersion parameters of the maximal and the kth reduced models, respectively.

First we assume that the dispersion parameters, ϕ and the ϕ_k, are known; later we will relax this assumption. In this case it follows that for each k,

$$u_k(x;\beta,\theta_k) = r_k(x;\beta,\theta_k,\phi_k)x_{A_k}, \tag{3}$$

where $r_k(x;\beta,\theta_k,\phi_k) = \{1/a(\phi_k)\}\{g^{-1}(x^{\text{T}}\beta) - g^{-1}(x_{A_k}^{\text{T}}\theta_k)\}h'(x_{A_k}^{\text{T}}\theta_k)$. Then the empirical moment vector is $U_n(\beta,\hat{\theta}) = P_n\{u_1(X;\beta,\hat{\theta}_1)^{\text{T}},\ldots,u_K(X;\beta,\hat{\theta}_K)^{\text{T}}\}^{\text{T}}$. The Newton–Raphson method for seeking the minimizer of $Q_C(\beta)$ can be written as

$$\beta^{(t+1)} = \beta^{(t)} - (X_{\text{rbind}}^{\text{T}} W^* X_{\text{rbind}})^{-1} X_{\text{rbind}}^{\text{T}} W X_{A_{\text{diag}}} C X_{A_{\text{diag}}}^{\text{T}} r. \tag{4}$$

In (4), $X_{\text{rbind}} = 1 \otimes X$ where $X_{(n\times p)}$ is the reference data matrix; $X_{A_{\text{diag}}} = \text{diag}(X_{A_1},\ldots,X_{A_K})$ where $X_{A_k(n\times d_k)}$ is the reference data matrix for the kth study; $W = \text{diag}(W_1,\ldots,W_K)$ with $W_k = \text{diag}(w_{k1},\ldots,w_{kn})$ and $w_{ki} = (1/[a(\phi_k)g'\{g^{-1}(X_i^T\beta^{(t)})\}])h'(X_{A_k,i}^{\text{T}}\hat{\theta}_k)$ $(k = 1,\ldots,K;\ i = 1,\ldots,n)$; W^* is the sum of $WX_{A_{\text{diag}}}CX_{A_{\text{diag}}}^{\text{T}}W$ and $\text{diag}(r^{\text{T}}X_{A_{\text{diag}}}CX_{A_{\text{diag}}}^{\text{T}}L)$, a diagonalized matrix from a vector; $r = (r_1^{\text{T}},\ldots,r_K^{\text{T}})^{\text{T}}$ with $r_k = (r_{k1},\ldots,r_{kn})^{\text{T}}$ and $r_{ki} = r_k(X_i;\beta^{(t)},\hat{\theta}_k,\phi_k)$; and $L = \text{diag}(L_1,\ldots,L_K)$ with $L_k = \text{diag}(l_{k1},\ldots,l_{kn})$ and $l_{ki} = -g''\{g^{-1}(X_i^{\text{T}}\beta^{(t)})\}/(a(\phi_k)[g'\{g^{-1}(X_i^{\text{T}}\beta^{(t)})\}]^3 h'(X_{A_k,i}^{\text{T}}\hat{\theta}_k))$. Equation (4) implies that the Newton–Raphson method is an iterated reweighted least-squares algorithm.

When ϕ and the ϕ_k are unknown, we propose to first obtain the estimator $\hat{\beta}$ of $\beta^\star$ as above with ϕ_k replaced by $\hat{\phi}_k$. Next, we consider the estimation of $\phi^\star$, the true value of ϕ. For the kth reduced model, we have an additional score function with respect to ϕ_k, from which we can obtain, similar to equation (3),

$$u_k(X;\beta,\phi,\theta_k,\phi_k) = -\frac{a'(\phi_k)}{a^2(\phi_k)}\big[g^{-1}(X^{\text{T}}\beta)h(X_{A_k}^{\text{T}}\theta_k) - b\{h(X_{A_k}^{\text{T}}\theta_k)\}\big] + q_k(X;\beta,\phi,\phi_k),$$

with $q_k = E_{Y|X}\{c'(Y;\phi_k)\}$ where $c'(Y;\phi_k)$ is the derivative of $c(Y;\phi_k)$ with respect to ϕ_k. Then the empirical moment vector for ϕ is $U_n(\phi) = P_n\{u_1(X;\hat{\beta},\phi,\hat{\theta}_1,\hat{\phi}_1)^{\text{T}},\ldots,u_K(X;\hat{\beta},\phi,\hat{\theta}_K,\hat{\phi}_K)^{\text{T}}\}^{\text{T}}$.

To estimate $\phi^\star$, we need to compute the minimizer of $U_n(\phi)^{\mathrm{T}} C U_n(\phi)$, where C is a known weighting matrix. The Newton–Raphson steps can be written as

$$\phi^{(t+1)} = \phi^{(t)} - J_n^{-1}(\phi^{(t)}) D_n(\phi^{(t)}), \tag{5}$$

where $J_n(\phi) = U_n^{\mathrm{T}}(\phi) C \mathrm{d}^2 q_n(\phi)/\mathrm{d}\phi^2 + \{\mathrm{d}q_n(\phi)/\mathrm{d}\phi\}^{\mathrm{T}} C \mathrm{d}q_n(\phi)/\mathrm{d}\phi$, $D_n(\phi) = U_n^{\mathrm{T}}(\phi^{(t)}) C \mathrm{d}q_n(\phi)/\mathrm{d}\phi$ and $q_n(\phi) = P_n\{q_1(X;\hat\beta,\phi,\hat\phi_1),\ldots,q_K(X;\hat\beta,\phi,\hat\phi_K)\}^{\mathrm{T}}$. In brief, when ϕ and ϕ_k ($k = 1,\ldots,K$) are unknown, we first choose initial estimates $\beta^{(0)}$ and $\phi^{(0)}$. Then we obtain the estimator $\hat\beta$ by iterating (4) until a stopping rule is reached. Subsequently $\phi^{(0)}$, $\hat\beta$ and the study estimates are inserted into (5), and the process is repeated until a stopping rule is reached, giving the GENMETA estimator of ϕ^*. In each Newton–Raphson step, the weighting matrix C is estimated by the estimates from the previous step.

2.5. *Diagnostic test for model violation*

Our generalized meta-analysis approach relies on several modelling assumptions, including homogeneity of the underlying populations with respect to the distribution of covariates and regression parameters, and correct specification of the maximal model. In the absence of individual-level data from the different studies, these assumptions cannot be tested in the usual manner using traditional diagnostic tests. However, even with summary-level data, some diagnostic testing is possible. In particular, from an intuitive perspective, departure of the GENMETA estimating equations, when evaluated at estimated parameter values, from their expected null value will be indicative of disagreement between the model and the observed data, i.e., the estimates of the parameters for the reduced models from different studies. For example, if the regression parameters underlying the maximal model are highly heterogeneous across studies, then the assumption of a common β in GENMETA will not be able to explain the heterogeneity that is expected to be present in overlapping reduced model parameters across the studies. Specifically, we propose to use the score test based on the statistic $T_{\mathrm{GENMETA}} = nQ_{\hat C_{\mathrm{opt}}}(\hat\beta)$, where $\hat\beta$ is the GENMETA estimate. When all the underlying assumptions are correct, by the standard generalized method of moments theory, T_{GENMETA} converges in distribution to a χ^2 distribution with $d - p$ degrees of freedom, where d is the total number of GENMETA equations and p is the total number of underlying parameters that are being estimated. The test is applicable only when $d > p$, which is the case when different studies have overlapping covariates.

3. SIMULATIONS

3.1. *Set-up*

We study the performance of our estimators through simulation studies in both idealized and non-idealized settings. In all simulations, we assume that the relationship between a binary outcome variable Y and three covariates (X_1, X_2, X_3) can be described with a logistic regression model of the form

$$Y \mid (X_1,X_2,X_3) \sim \mathrm{Ber}([1 + \exp\{-(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2 + \beta_3^* X_3)\}]^{-1}), \tag{6}$$

where (X_1,X_2,X_3) follows a multivariate normal distribution with mean $\mu = (\mu_1,\mu_2,\mu_3)$, variance $\sigma^2 = (\sigma_1^2,\sigma_2^2,\sigma_3^2)$ and underlying correlations $\rho = (\rho_{12},\rho_{13},\rho_{23})$. We choose $\beta_1^* = \beta_2^* = \beta_3^* = \log 1.3$ to reflect a moderate degree of association of the outcome with each covariate after adjusting for the others. We assume that there are three separate studies,

Table 1. *Simulation results for our generalized meta-analysis estimators in the logistic regression setting; estimated standard deviations were obtained by taking averages over simulated datasets and were used to construct* 95% *confidence intervals, whose coverage rates and average lengths are reported*

$n = 50$	β_i^*	Bias	SD (ESD_1, ESD_2)	RMSE	CR	AL
	β_1^*	0.010	0.161 (0.161, 0.162)	0.161	0.968, 0.964	0.642, 0.636
GENMETA.0	β_2^*	0.005	0.110 (0.111, 0.108)	0.110	0.958, 0.960	0.434, 0.423
	β_3^*	−0.001	0.138 (0.143, 0.142)	0.138	0.963, 0.964	0.559, 0.556
	β_1^*	0.005	0.117 (0.116, 0.110)	0.117	0.976, 0.966	0.455, 0.433
GENMETA.1	β_2^*	−0.003	0.101 (0.105, 0.099)	0.101	0.964, 0.955	0.411, 0.386
	β_3^*	0.001	0.099 (0.102, 0.097)	0.099	0.973, 0.961	0.402, 0.381
	β_1^*	0.007	0.115 (0.116, 0.111)	0.115	0.971, 0.964	0.455, 0.435
GENMETA.2	β_2^*	−0.003	0.102 (0.105, 0.099)	0.102	0.960, 0.959	0.413, 0.388
	β_3^*	0.003	0.098 (0.103, 0.098)	0.098	0.957, 0.957	0.403, 0.383

SD, standard deviation; ESD_1, estimated standard deviation using the reference sample; ESD_2, estimated standard deviation using the covariance estimates of reduced model parameters from the studies; RMSE, square root of mean square error; CR, coverage rate of 95% confidence intervals; AL, average length of 95% confidence intervals.

where each study fits a reduced logistic model for the outcome Y on two of the covariates in the form

$$Y \mid (X_i, X_j) \sim \text{Ber}([1 + \exp\{-(\theta^*_{0,ij} + \theta^*_{i,ij}X_i + \theta^*_{j,ij}X_j)\}]^{-1}), \tag{7}$$

with X_1 and X_2 included in study I, X_2 and X_3 in study II, and X_1 and X_3 in study III. Here, as the data for each study are generated using the maximal model, the reduced models are by definition incompatible due to the non-collapsibility of the logistic model. We fix the sample size of the studies at $n_1 = 300$, $n_2 = 500$ and $n_3 = 1000$, and vary the sample size of the reference dataset.

3.2. *Homogeneous population*

We assume that the studies are conducted in the same underlying population from which the reference sample is drawn. In this setting, there exists a common mean vector $\mu_b = (0, 0, 0)$, a common variance vector $\sigma_b^2 = (1, 1, 1)$ and a common correlation vector $\rho_b = (0.3, 0.6, 0.1)$, which describes the joint distribution of the three covariates across all the underlying populations. In the first set of simulations, we assume a fixed sample size $n = 50$ for the reference dataset. In all settings, we simulate data (Y, X_1, X_2, X_3) for the underlying studies based on the data-generating models as described above, and we fit the respective reduced models to obtain estimates of the reduced model parameters. For each set of simulated data, we obtain estimates of covariance matrices of the reduced model parameters using robust sandwich estimators based on either the study datasets themselves or the reference dataset; see (2). We consider three estimators: GENMETA.0, which is the initial GENMETA estimator with identity weighting matrix, and GENMETA.1 and GENMETA.2, which use covariance estimates from the reference dataset and from the studies, respectively.

From the results shown in Table 1, we see that all three estimators are nearly unbiased. The standard error estimates, irrespective of whether Σ_k $(k = 1, 2, 3)$ were estimated using the study datasets or the reference sample, accurately reflect the true standard errors of the GENMETA parameter estimates across different simulations. As a result, the 95% confidence intervals maintain the coverage probability at the nominal level. Among the three estimators considered, clearly

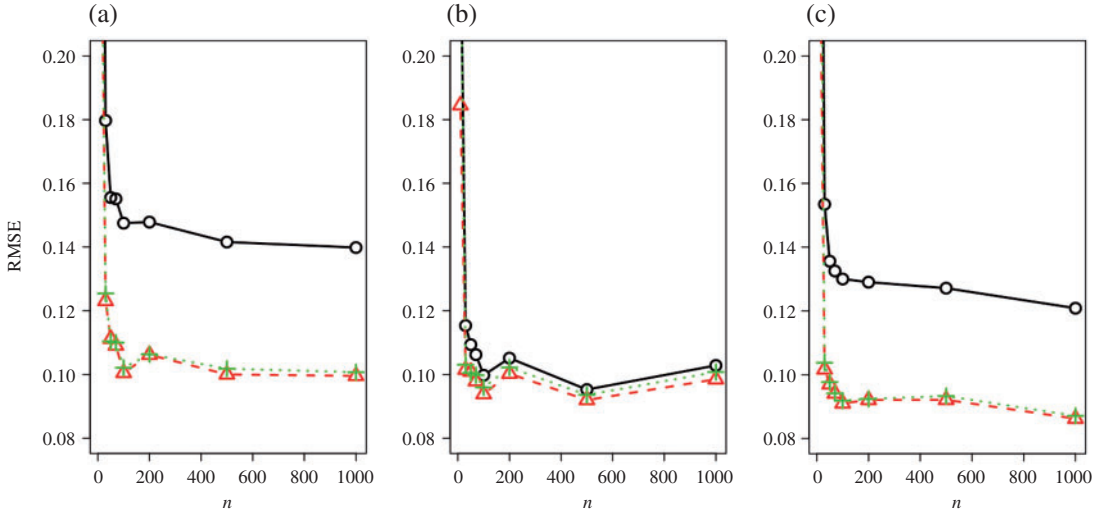

Fig. 1. Root mean square errors, RMSE, of the of GENMETA estimators for (a) β_1^*, (b) β_2^* and (c) β_3^* with fixed study sample sizes $n_1 = 300, n_2 = 500$ and $n_3 = 1000$ and varying reference sample size $n \in \{10, 30, 50, 70, 100, 200, 1000\}$: GENMETA.0, circles and solid line; GENMETA.1, triangles and dashed line; GENMETA.2, plus signs and dotted line.

GENMETA.0, which uses the non-optimal choice of $C = I$, is less efficient than GENMETA.1 and GENMETA.2, which had comparable efficiency.

In the same setting as above, when we vary n from 10 up to a maximum of 1000, we observe that the precision of the GENMETA estimates does not increase with n once it reaches a threshold of around 100, which is one-third of the minimum of the study sample sizes ($n_1 = 300$); see Fig. 1. The thresholds were even lower for estimation of coefficients associated with X_2, which had weak to moderate correlation with the other covariates in the model. That the reference dataset can be substantially smaller than the study datasets without having much impact on the precision of our estimator is encouraging, given that accessing a reference dataset with a large sample size may be difficult in practice.

Finally, we conduct additional simulation studies to gain more insight into results from the real-data analysis. The settings are similar to those described above, except that we assume there are only two studies: study I fits the maximal logistic regression model involving all three covariates, while study II involves only two covariates, X_1 and X_2. We assume $\rho_{\mathrm{I}} = \rho_{\mathrm{II}} = \rho_{\mathrm{b}}$. In our estimation, we further incorporated an added complexity to account for study-specific intercept terms for the maximal logistic regression model,

$$Y \mid (X_1, X_2, X_3, \text{study}) \sim \text{Ber}([1 + \exp\{-(\beta^*_{0,\text{study}} + \beta_1^* X_1 + \beta_2^* X_2 + \beta_3^* X_3)\}]^{-1}),$$

so that the prevalence of the outcome, $\text{pr}(Y = 1)$, could be different across the two studies. In this setting, the maximal set of parameters that are to be estimated through GENMETA can be defined as $\beta^* = (\beta_{0,\text{ study I}}, \beta_{0,\text{ study II}}, \beta_1, \beta_2, \beta_3)$. We simulated data using values of intercept parameters that are identical for the two models, but for estimation we allowed the intercept parameters to be different. For the sake of comparison, we also fitted a reduced model for study I and conducted a standard multivariate meta-analysis of the underlying common parameters associated with X_1 and X_2 across the two studies. We took the sample sizes for the two studies to be $n_1 = 500$ and $n_2 = 5000$, and that for the reference dataset to be $n = 300$.

Table 2 shows that in this simulation setting the reduced models produce biased estimates for β_1^*, but not for β_2^*. The result is intuitive given that the omitted covariate X_3 is primarily

Table 2. *A simulation for understanding the real-data analysis: point estimates and standard deviations from logistic regression with reduced and maximal models, meta-analysis, and* GENMETA *estimation with* $\beta_1^* = \beta_2^* = \beta_3^* = \log(1.3) \approx 0.262$

	Study I		Study II	Meta-analysis	GENMETA	
	Maximal	Reduced	Reduced	Reduced	Reduced	Maximal
β_i^*	PE (SD)	PE (SD)	PE (SD)	PE (SD)	PE (SD)	PE (SD)
β_1^*	0.270 (0.149)	0.429 (0.116)	0.424 (0.037)	0.424 (0.035)	0.425 (0.035)	0.268 (0.088)
β_2^*	0.263 (0.111)	0.243 (0.112)	0.236 (0.035)	0.236 (0.034)	0.237 (0.034)	0.263 (0.039)
β_3^*	0.258 (0.136)	NA	NA	NA	NA	0.255 (0.135)

PE, point estimate; SD, standard deviation; NA, no corresponding estimator.

correlated with X_1. As a result, standard meta-analysis was nearly unbiased for β_2^*, but not for β_1^*. Parameter estimates from the maximal model in study I are unbiased for all parameters, but have much larger standard errors compared to those obtained from meta-analysis for estimation of β_2^*. Our generalized meta-anlysis estimator produced unbiased estimates for all the parameters and, at the same time, has efficiency comparable to standard meta-analysis for estimation of β_2^*. These results highlight a desirable feature of our estimator, namely that it can effectively combine information across studies to minimize bias due to omitted covariates, and yet utilize all the information available across the partially informative studies.

3.3. *Heterogeneous population*

We now conduct simulation studies where the underlying assumption of homogeneity of the covariate distribution across populations may be violated in various ways. As a benchmark for comparison, setting (I) will be the same as the one we simulated under the homogeneous population. In setting (II), we allow the means and/or variances to vary across the populations, underlying the studies and the reference sample, while keeping the correlations constant. Specifically, the mean vector for the three covariates can take one of three possible values: $\mu_{\mathrm{h}} = (1, 1, 1)$, $\mu_{\mathrm{m}} = (0.5, 0.5, 0.5)$ and $\mu_{\mathrm{b}} = (0, 0, 0)$. Similarly, the variance vector is allowed to vary across three possible sets of values: $\sigma_{\mathrm{h}}^2 = (2, 2, 2)$, $\sigma_{\mathrm{l}}^2 = (0.5, 0.5, 0.5)$ and $\sigma_{\mathrm{b}}^2 = (1, 1, 1)$. In setting (III), we allow the correlations among the covariates to vary across populations; here we also consider three possible sets of correlation vectors, namely $\rho_{\mathrm{l}} = (0.2, 0.4, 0.0)$, $\rho_{\mathrm{h}} = (0.4, 0.8, 0.2)$ and $\rho_{\mathrm{b}} = (0.3, 0.6, 0.1)$. In simulation setting (IV), we allow for potentially different inclusion criteria across studies, leading to possible violations of the assumption of homogeneity of the covariate distribution. Specifically, we first simulate an underlying study base using the set-up described in simulation setting (I), and then for study I we keep only individuals with $X_1 > -0.5$ and $X_2 < 0.5$, while in study II we keep individuals with $X_1 > 0$. Finally, we consider an alternative simulation scenario where we assume that the covariates are log-normally distributed by defining $X = \exp(W)$, where W is generated from a multivariate normal distribution following the same settings as in (I)–(IV) above.

When the covariates are normally distributed, we observe that the proposed method is not very sensitive to the underlying assumption of homogeneity of the covariate distribution; see Table 3. In setting (II), where the means and/or variances of the covariates vary across the populations, but the correlations are fixed, there is virtually no bias. In setting (III), where the correlations are varied, we observe more noticeable, but still small, biases in the parameter estimates. In setting (IV), where the inclusion criteria vary across studies, there is also very minimal bias. When the covariates are log-normally distributed, however, the method can be more sensitive to violation

Table 3. *Robustness of generalized meta-analysis estimation: results for the* GENMETA *estimates using the study covariance estimators in the setting of logistic regression. In setting* (I)*, data are simulated in the ideal setting where the covariate distribution, characterized by the mean, standard deviation and correlation of normal variates, is the same across all populations. In settings* (II)–(IV)*, the assumption is violated by creating variations in means and/or standard deviations, correlations, and selection criteria across the studies and the reference sample. The vectors of covariate means, variances and correlations are denoted by* $\mu_* = (\mu_1, \mu_2, \mu_3)$*,* $\sigma_*^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2)$ *and* $\rho_* = (\rho_{12}, \rho_{23}, \rho_{13})$ *for* $* \in \{\mathrm{b}, \mathrm{l}, \mathrm{m}, \mathrm{h}\}$*, where* $\mu_\mathrm{b} = (0,0,0)$*,* $\mu_\mathrm{m} = (0.5, 0.5, 0.5)$*,* $\mu_\mathrm{h} = (1,1,1)$*,* $\sigma_\mathrm{b}^2 = (1,1,1)$*,* $\sigma_\mathrm{l}^2 = (0.5, 0.5, 0.5)$*,* $\sigma_\mathrm{h}^2 = (2,2,2)$*,* $\rho_\mathrm{b} = (0.3, 0.6, 0.1)$*,* $\rho_\mathrm{h} = (0.4, 0.8, 0.2)$ *and* $\rho_\mathrm{l} = (0.2, 0.4, 0)$*. Estimated standard deviation is obtained from the asymptotic formula* (1) *and used to construct* 95% *confidence intervals*

Setting	Study I	Study II	Study III	Reference	β_i^*	Bias	SD (ESD)	RMSE	CR	AL
	μ_b	μ_b	μ_b	μ_b	β_1^*	0.001	0.111 (0.112)	0.111	0.947	0.437
I	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	−0.002	0.098 (0.099)	0.098	0.956	0.389
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	0.005	0.096 (0.098)	0.096	0.954	0.382
	μ_b	μ_h	μ_m	μ_b	β_1^*	0.010	0.103 (0.104)	0.103	0.952	0.405
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	−0.006	0.083 (0.083)	0.083	0.954	0.324
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	0.005	0.085 (0.088)	0.085	0.956	0.343
	μ_b	μ_b	μ_b	μ_b	β_1^*	0.003	0.139 (0.136)	0.139	0.939	0.529
II	σ_b^2	σ_h^2	σ_l^2	σ_b^2	β_2^*	−0.003	0.084 (0.086)	0.084	0.956	0.335
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	0.003	0.112 (0.111)	0.112	0.949	0.431
	μ_b	μ_h	μ_m	μ_b	β_1^*	0.013	0.124 (0.126)	0.125	0.946	0.493
	σ_b^2	σ_h^2	σ_l^2	σ_b^2	β_2^*	−0.006	0.073 (0.075)	0.073	0.958	0.291
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	0.005	0.097 (0.100)	0.097	0.949	0.391
	μ_b	μ_b	μ_b	μ_b	β_1^*	−0.092	0.142 (0.151)	0.169	0.958	0.579
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	0.019	0.105 (0.109)	0.107	0.963	0.423
	ρ_b	ρ_b	ρ_b	ρ_h	β_3^*	0.053	0.120 (0.129)	0.131	0.971	0.495
	μ_b	μ_b	μ_b	μ_b	β_1^*	0.035	0.099 (0.099)	0.106	0.917	0.385
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	0.002	0.096 (0.096)	0.096	0.954	0.377
	ρ_b	ρ_b	ρ_b	ρ_l	β_3^*	0.012	0.087 (0.087)	0.088	0.944	0.343
	μ_b	μ_b	μ_b	μ_b	β_1^*	0.060	0.113 (0.113)	0.128	0.916	0.443
III	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	−0.001	0.096 (0.097)	0.096	0.955	0.379
	ρ_l	ρ_b	ρ_h	ρ_l	β_3^*	−0.006	0.103 (0.102)	0.104	0.944	0.398
	μ_b	μ_b	μ_b	μ_b	β_1^*	0.039	0.130 (0.132)	0.135	0.939	0.515
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	−0.006	0.097 (0.100)	0.097	0.958	0.392
	ρ_l	ρ_b	ρ_h	ρ_b	β_3^*	−0.027	0.116 (0.118)	0.119	0.944	0.461
	μ_b	μ_b	μ_b	μ_b	β_1^*	−0.036	0.165 (0.173)	0.169	0.957	0.671
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	0.013	0.103 (0.109)	0.104	0.962	0.424
	ρ_l	ρ_b	ρ_h	ρ_h	β_3^*	0.003	0.143 (0.153)	0.143	0.959	0.591
			μ_b	μ_b	β_1^*	0.014	0.123 (0.127)	0.124	0.961	0.494
IV	$X_1 > -0.5$,	$X_2 > 0$	σ_b^2	σ_b^2	β_2^*	−0.008	0.105 (0.109)	0.105	0.965	0.428
	$X_2 < 0.5$		ρ_b	ρ_b	β_3^*	−0.001	0.094 (0.093)	0.093	0.958	0.366

SD, standard deviation; ESD, estimated standard deviation; RMSE, square root of mean square error; CR, coverage rate of 95% confidence intervals; AL, average length of 95% confidence intervals.

of the underlying homogeneity assumption; see the Supplementary Material. In particular, when the inclusion criteria varied across studies in setting (IV), large bias in point estimates and low

coverage probability were observed for estimation of the coefficient associated with X_2, the covariate which is used to define fairly non-overlapping inclusion criteria across two studies. Notably, even in this scenario, minimal bias is observed for estimation of the other covariates in the model.

3.4. *Power evaluation of the diagnostic test*

We assess the power of the proposed test statistic, T_{GENMETA}, in the presence of heterogeneity in the regression parameters, β, across the studies. In the context of standard multivariate meta-analysis, where it is assumed that all the studies ascertain the same set of covariates, the test for heterogeneity is performed using the standard multivariate Cochran's test-statistic

$$Q = \sum_{k=1}^{K} (\hat{\beta}_k - \hat{\beta}_{\text{meta}})^{\text{T}} S_k^{-1} (\hat{\beta}_k - \hat{\beta}_{\text{meta}}),$$

where $\hat{\beta}_{\text{meta}}$ is the usual multivariate meta-analysis estimate and S_k is the standard error of $\hat{\beta}_k$ for $k = 1, \ldots, K$. We will use Q as a benchmark to evaluate the power of T_{GENMETA}.

In all simulations, as before, we assume that there are three separate studies, and that the relationship between a binary outcome variable Y and three covariates (X_1, X_2, X_3) in each study follows the same logistic regression model of the form (6). However, instead of assuming a fixed set of β across all studies, we simulate different values of β from a normal distribution with mean $(\beta_1^*, \beta_2^*, \beta_3^*) = (\log 1.3, \log 1.3, \log 1.3)$ and variance $\sigma^2 I$, where the parameter $\sigma^2 > 0$ is varied to control the degree of heterogeneity across studies. As before, we assume that (X_1, X_2, X_3) follows a multivariate normal distribution with zero mean, unit variance and underlying correlations $\rho_{12} = 0.3$, $\rho_{13} = 0.6$ and $\rho_{23} = 0.1$ for the three studies. We simulate data for the different studies from the above random-effects logistic regression model and then fit reduced models of the form (7) to the three studies. In particular, we assume that X_1 and X_2 are included in study I, X_2 and X_3 in study II, and X_1 and X_3 in study III. We fix the sample sizes of the studies at $n_1 = 3000$, $n_2 = 5000$ and $n_3 = 10\,000$, and vary the sample size of the reference dataset. The level of the test is set to 5%. For comparison, we also fit the maximal model to each study involving all three covariates and apply the standard Q-statistic for testing heterogeneity.

Comparison of the power of T_{GENMETA} and of the Q-statistic shows that, as expected, the power for both tests increases as a function of the degree of heterogeneity, σ^2; see Fig. 2. Clearly, T_{GENMETA} suffers some loss of power as it deals with the missing covariates, but it retains enough power, even with a small reference dataset ($n = 100$), to remain practically useful.

4. REAL-DATA ANALYSIS

In this section we illustrate application of the proposed method by developing a model for predicting the risk of breast cancer using a combination of different risk factors based on data from multiple studies. The first study, the Breast Prostate Colorectal Cancer Cohort, BPC3, study, includes a total of 7448 cases and 8812 controls, drawn from eight different underlying cohorts. Details of the study, including its recent application to the development of a breast cancer risk prediction model, can be found in Mass et al. (2016). Here we focus on the analysis of breast cancer risk associated with a selected set of factors, including family history, age at menarche, age at first birth, and body weight. The second study is the Breast Cancer Detection and Demonstration Project, BCDDP, with a dataset containing 1217 cases and 1616 controls. The study has previously been used to develop an updated version of the widely popular Breast

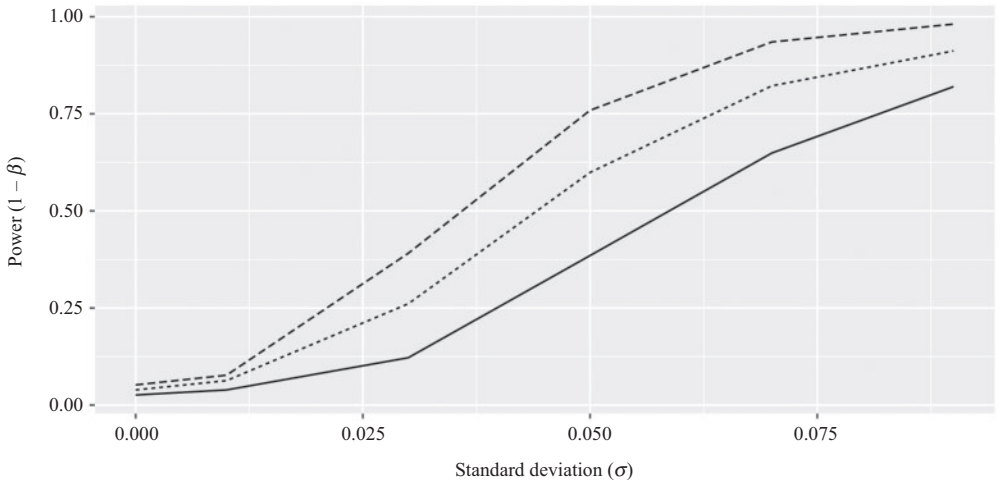

Fig. 2. Power curves of the simple multivariate meta-analysis test statistic, Q, and T_{GENMETA} for simulated datasets: the simple meta-analysis estimator (dashed) and GENMETA estimators with reference data sample sizes of 100 (solid) and 500 (dotted). The level of the test is set to 0.05.

Cancer Risk Assessment tool (Chen et al., 2006) that incorporates mammographic density, the areal proportion of breast tissue that is radiographically dense, which is known to be a strong risk factor for breast cancer. The dataset from the BCDDP study contains mammographic density and the number of previous breast biopsies, in addition to all the factors considered in the BPC3 data analysis. Let X denote the common set of covariates measured across both studies, and let Z represent the factors available only in BCDDP. The goal is to estimate parameters associated with an underlying logistic regression model that includes all of the different factors. While the BPC3 study is large in size and represents multiple populations, it has information on a more limited number of risk factors. The BCDDP study, on the other hand, has information on an extended set of risk factors, but is much smaller in size. A combined analysis of these two studies can potentially yield more generalizable and precise estimates of risk parameters.

Throughout the analysis, we use a sample of 137 cases and 163 controls from the BCDDP study as the reference sample, based on which the distribution of covariates is estimated. To maintain independence of the reference and study samples, we exclude the reference sample from the primary analysis of the BCDDP study, which involves estimation of the log-odds-ratio parameters. Further, both of the studies involve case-control sampling with similar case-control proportions. In general, if nonrandom sampling is used for selection of subjects in any of the studies, then the covariate distribution underlying the GENMETA estimating equation needs to be adjusted to account for the study design. In this application, because we had access to the BCDDP study, we could adjust for the design effect by simply selecting a reference sample that includes cases and controls in a similar ratio to that in the main studies. In general, however, the effect of nonrandom sampling design for the main studies may need to be accounted for through careful weighting of subjects in the reference sample.

For each of the eight cohorts in the BPC3 study and for the BCCDP study, we first fit a reduced logistic regression model including X. All models include age as an additional cofactor as well as study-specific intercept parameters and age effects. Specifically, we consider underlying models of the form

$$(Y \mid X, \text{Age}, \text{study} = k) \sim \text{Ber}([1 + \exp\{-(\theta_{0k} + \theta_{A_k}\text{Age} + \theta_X^{\mathrm{T}} X)\}]^{-1}). \tag{8}$$

We applied the diagnostic test for model violation to these datasets. We found the value of the test-statistic, $\hat{T}_{\text{GENMETA}}$, to be 59.01 and the corresponding p-value to be 0.366 under a $\chi^2_{(56)}$ distribution. Hence, it appears that the underlying model assumptions are unlikely to be grossly violated in this application.

First, to illustrate how our proposed estimator compares with the standard meta-analysis method, we estimate the common underlying parameters of interest θ_X using these two methods. We fitted model (8) separately for each study and obtained estimates of the parameters and covariance matrices. Then, for the underlying common parameter of interest θ_X, we conducted a standard multivariate meta-analysis using the corresponding subset of parameter estimates and covariance matrices. Alternatively, using the parameter estimates and variance-covariance matrices from the individual studies, and using the BCDDP sample that was set aside as the reference dataset to estimate the joint distribution of X and Age, we estimated all the parameters of model (8) using our procedure. From the results reported in Table 4, it can be seen that in this setting the meta-analysis and our estimators produce similar estimates and corresponding standard errors across all the different risk factors of interest. In one of the results stated earlier, we saw theoretically that in an idealized setting, where all the models and underlying populations are identical, the two estimators are asymptotically equivalent. It is encouraging to observe the close correspondence between the estimators in the data analysis, which involves a diverse set of studies that are likely to have significant heterogeneity across the underlying populations. In particular, for a number of the risk factors, such as family history, coefficient estimates were noticeably different for the two studies. When significant heterogeneity existed, the meta-analysed estimates were pooled closer to those from the BPC3 study because of its large sample size.

Next, we turn our attention to the analysis of data from the BCDDP study using a maximal model that includes X and the additional covariates, mammographic density and number of previous breast biopsies. Comparison of the parameter estimates associated with X across the maximal and reduced models within the BCDDP study indicates major differences in the estimates of the coefficients associated with weight. In the maximal model, higher weight is found to be much more strongly associated with increased risk of breast cancer. The unmasking of the effect of weight in the maximal model is intuitive, given that body weight and mammographic density are known to have a strong negative correlation. Although not as dramatic, there are some differences in the effects of age at menarche and age at first birth between the maximal and reduced models, also possibly due to the modest correlation of these factors with mammographic density and the number of previous breast biopsies. The effect of family history, however, is almost identical across the two models.

Finally, we used our generalized meta-anlysis method to combine estimates of the parameters of the maximal model from the BCDDP study and those from the reduced models for the eight BPC3 cohorts. We assumed an underlying maximal model of interest across the nine studies:

$$(Y \mid X, Z, \text{Age}, \text{study} = k) \sim \text{Ber}([1 + \exp\{-(\theta_{0k} + \theta_{A_k}\text{Age} + \beta_X^{\mathrm{T}} X + \beta_Z^{\mathrm{T}} Z)\}]^{-1}).$$

We observe that our generalized meta-analysis approach produces estimates of the effect of family history and associated standard error that are very similar to those based on standard meta-analysis of the reduced models across the nine cohorts. The estimate is pooled heavily towards the BPC3 study due to its large sample size. In contrast, the GENMETA estimates for weight are very similar to those obtained from the maximal model only within the BCDDP study. These results are consistent with the simulation studies, in which GENMETA behaves similarly to reduced-model meta-analysis when omitted covariates do not cause notable bias. In contrast, when omitted covariates cause considerable bias, our estimator is pooled towards estimates from

Table 4. *Real-data analysis results comparing meta-analysis and our generalized meta-analysis method: combined analysis of the BCDDP and BPC3 studies to develop a multivariate logistic regression model for breast cancer risk. For each cohort within BPC3 and for BCDDP, the standard logistic regression model is used to fit reduced models; parameter estimates of the reduced models across studies are then combined using standard meta-analysis or* GENMETA. *For the BCDDP study, a maximal logistic model is fitted including two additional covariates. These estimates are then combined with estimates of reduced model parameters from BPC3 to obtain* GENMETA *estimates of the maximal model*

	BPC3							
	CPS2	EPIC	MCCS	MEC	NHS	PLCO	WHI	WHS
Risk	cohort	cohort	cohort	cohort	cohort	cohort	cohort	cohort
factors	PE (SE)	PE (SE)	PE (SE)	PE (SE)	PE (SE)	PE (SE)	PE (SE)	PE (SE)
FH	0.47 (0.13)	0.29 (0.15)	0.56 (0.19)	0.41 (0.28)	0.48 (0.08)	0.39 (0.13)	0.30 (0.06)	0.28 (0.19)
AMEN1	−0.03 (0.14)	0.02 (0.09)	−0.19 (0.17)	−0.09 (0.24)	0.06 (0.09)	−0.05 (0.12)	0.13 (0.08)	0.03 (0.17)
AMEN2	−0.09 (0.17)	0.04 (0.12)	−0.44 (0.23)	0.35 (0.35)	0.19 (0.10)	0.03 (0.15)	0.19 (0.09)	0.14 (0.19)
AFB1	0.28 (0.17)	0.12 (0.14)	−0.08 (0.25)	0.06 (0.17)	0.39 (0.20)	0.16 (0.14)	0.19 (0.09)	0.92 (0.23)
AFB2	0.73 (0.24)	0.24 (0.17)	0.35 (0.30)	0.05 (0.26)	0.36 (0.22)	0.52 (0.22)	0.44 (0.13)	0.96 (0.28)
WT1	0.09 (0.14)	−0.01 (0.09)	0.22 (0.18)	0.09 (0.17)	0.21 (0.08)	0.09 (0.13)	−0.03 (0.08)	−0.01 (0.14)
WT2	0.16 (0.14)	0.24 (0.11)	0.45 (0.19)	−0.08 (0.18)	0.10 (0.08)	0.09 (0.13)	0.18 (0.08)	−0.16 (0.15)

	BCDDP		Meta-analysis	GENMETA	
	Maximal	Reduced	Reduced	Reduced	Maximal
Risk	model	model	model	model	model
factors	PE (SE)	PE (SE)	PE (SE)	PE (SE)	PE (SE)
FH	0.80 (0.14)	0.80 (0.14)	0.40 (0.04)	0.42 (0.04)	0.37 (0.08)
AMEN1	0.11 (0.10)	0.07 (0.10)	0.04 (0.04)	0.03 (0.04)	0.04 (0.06)
AMEN2	0.55 (0.15)	0.45 (0.15)	0.13 (0.05)	0.13 (0.05)	0.32 (0.08)
AFB1	0.06 (0.14)	0.18 (0.15)	0.21 (0.05)	0.20 (0.05)	0.05 (0.09)
AFB2	0.29 (0.20)	0.46 (0.20)	0.38 (0.06)	0.38 (0.07)	0.21 (0.12)
WT1	0.29 (0.11)	0.09 (0.11)	0.08 (0.04)	0.08 (0.04)	0.31 (0.07)
WT2	0.52 (0.13)	0.10 (0.13)	0.14 (0.04)	0.14 (0.04)	0.63 (0.09)
NBIOPS	0.13 (0.09)	NA	NA	NA	0.13 (0.10)
MD	0.46 (0.05)	NA	NA	NA	0.43 (0.06)

FH, binary indicator of family history; AMEN, age at menarche; AMEN1 and AMEN2, dummy variables associated with age-at-menarche categories $\geqslant$14, 12–13 and $\leqslant$11; AFB, age at first live birth; AFB1 and AFB2, dummy variables associated with age-at-first-live-birth categories $\leqslant$20, 21–29 and $\geqslant$30; WT, weight; WT1 and WT2, dummy variables associated with weight categories $\leqslant$62.6, 62.6–73.1 and $\geqslant$73.1 in kilograms; NBIOPS, number of previous biopsies coded as a continuous variable; MD, standardized mammographic density coded as a continuous variable; PE, point estimate; SE, standard error; NA, no corresponding estimator. CPS2, EPIC, MCCS, MEC, NHS, PLCO, WHI and WHS, abbreviated names of the eight cohorts of BPC3.

maximal or more complete models that may be available from a restricted set of studies. The behaviour of GENMETA for the other two covariates, age at menarche and age at first birth, was in between, which is also intuitive given that we observed their coefficients to have changed notably, but less dramatically, in the maximal model as compared to the reduced model within the BCDDP study. The GENMETA parameter estimates and standard errors for the additional variables of mammographic density and number of previous breast biopsies were similar to those observed for the maximal model in the BCDDP study, the only study for which information was available on these two factors. Thus, overall the data analysis demonstrates that our estimator behaves in a similar manner to meta-analysis for combining information across multiple possibly heterogeneous studies, but it has added flexibility to effectively combine information from disparate models.

5. DISCUSSION

The proposed method can be viewed as a natural extension of the traditional fixed-effect meta-analysis method that is widely used in practice. Our simulation studies and data analysis demonstrate that the method not only provides theoretically valid and efficient inference in idealized conditions, but also can perform robustly in non-idealized settings. A critical element of the proposed method is access to a reference dataset. While the ideal choice of reference dataset will vary by application, publicly available survey data, which contain information on a wide variety of factors, can be useful broadly. In fact, in large-scale genetic association studies, reference samples such as the 1000 Genomes Project are commonly used for estimating correlation parameters across genetic markers in the genome (The 1000 Genomes Project Consortium, 2012, 2015; Lee et al., 2013). For epidemiological studies, good sources of a reference dataset for the U.S. population include the National Health Interview Survey (Adams et al., 1999; Botman & Moriarity, 2000; Bloom et al., 2010) and the National Health and Nutrional Examination Survey (Fang & Alderman, 2000; He et al., 2001; de Ferranti et al., 2006; Idler & Angel, 2011; LaKind et al., 2012), which routinely collect data on a wide range of health- and lifestyle-related factors. If multiple studies coordinate through a consortium effort, which is becoming increasingly common in biomedical applications, then studies that have the most complete information, at least on some subsamples, can provide a reference sample.

When information on all covariates is not available in a single reference sample, one may have to consider using simulation to generate such data by combining information from multiple studies under some modelling assumptions. As access to large reference datasets can be difficult, researchers may find two aspects of our approach appealing. First, the sample size for the reference dataset can be small relative to the study datasets, and yet our generalized meta-analysis approach can have reasonable efficiency. In fact, increasing the sample size for the reference dataset beyond a certain threshold does not have an impact on the efficiency of our method. Secondly, although technically our method requires all the populations underlying the studies and the reference dataset to be the same, in practice the method can be robust against a reasonable degree of heterogeneity in the distribution of covariates. However, it is possible to have a large bias when estimating coefficients associated with covariates that have been used to define widely varying inclusion criteria. When different studies follow very different designs, it is best to obtain study-specific reference samples for estimating the underlying moment equations. Alternatively, it may be possible to modify a large reference sample by using study-specific sampling weights or inclusion criteria when estimating the moment equations. Dealing with study-specific covariates, such as centres within a study, can also pose challenges, as information on such variables is not expected to be available from a common reference sample. We have illustrated in our data example that it is possible to deal with such variables by imposing additional independence assumptions from other factors. In general, such complications need to be dealt with on a case-by-case basis, and some study-specific reference samples may be needed to avoid making strong assumptions. Further research is merited to explore these and other practical challenges in implementation of the proposed method.

In general, we believe that caution is needed for interpretations and applications of models developed by combining information from disparate models across multiple studies. A model developed from a single study with complete information may be inefficient and lack generalizability, but it is more likely to be internally consistent and thus can provide valid etiologic inference even if it is not representative of the general population. On the other hand, etiologic interpretation of parameters can be difficult when the underlying model is developed using information across multiple studies that are potentially heterogeneous. For predictive models, where the focus is not so much on parameter interpretation, development of rich models by combining

information across multiple studies and then validating such models in independent studies can be an appealing strategy. These and other practical issues related to model development using multiple data sources have also been discussed in several recent articles (Wang et al., 2015; Han & Lawless, 2017; Cheng et al., 2019; Estes et al., 2018).

In this article we have used generalized method of moments as the underlying inferential framework. Alternatively, inference could be performed using empirical likelihood theory (Qin & Lawless, 1994; Qin, 2000; Chatterjee et al., 2016), exploiting the same set of moment equations as we propose. While in small samples empirical likelihood estimators may perform better, their implementation can be substantially more complex. Recently, a simulation-based method has also been proposed for combining information on model parameters across disparate studies (Rahmandad et al., 2017). Computationally, our method may enjoy substantial advantages in dealing with complex models, such as those in high-dimensional settings, where repeated model fitting on simulated data is extensive. Further research is needed in multiple directions to increase the practical utility of GENMETA. It is possible that in some applications we may have information only on subsets of parameters underlying the fitted reduced models. It is an open question as to how such partial information can be used to set up the underlying moment equations in the GENMETA procedure. Ideally, to increase robustness of inference, the procedure should use study-specific reference samples for setting up the moment equations. For this purpose, it may be useful to develop strategies to combine information on a common reference sample with complete covariate information and data from individual studies that have partial covariate information.

Acknowledgement

This research was funded through a Patient-Centered Outcomes Research Institute Award, and the National Institutes of Health. The statements and opinions in this article are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute, its Board of Governors or its Methodology Committee. Chatterjee is also affiliated with the Department of Oncology at Johns Hopkins University.

Supplementary material

Supplementary material available at *Biometrika* online includes all the derivations, the proof of Theorem 1, and a table containing the simulation results for log-normally distributed covariates. The R (R Development Core Team, 2019) package `GENMETA` is available on CRAN at `https://cran.r-project.org/package=GENMETA`.

Appendix

Assumptions of Theorem 1

Assumptions A1–A4 are for consistency, and Assumptions A5–A9 are for asymptotic normality:

Assumption A1. C is positive semidefinite and $CE\{U(X;\beta,\theta^*)\} = 0$ if and only if $\beta = \beta^*$;

Assumption A2. $\beta^* \in D_\beta$, which is compact;

Assumption A3. $u_k(X;\beta,\theta_k)$ is continuous for each $(\beta,\theta_k) \in D_\beta \times \mathcal{N}(\theta_k^*)$ with probability 1, where $\mathcal{N}(\theta_k^*)$ is a neighbourhood of θ_k^* for $k = 1,\ldots,K$;

Assumption A4. $E\{\sup_{(\beta,\theta_k)\in D_\beta\times\mathcal{N}(\theta_k^*)} \|u_k(X;\beta,\theta_k)\|\} < \infty$ for $k = 1,\ldots,K$;

Assumption A5. $\partial u_k(X;\beta,\theta_k)/\partial\beta$ is continuous at each $(\beta,\theta_k) \in \mathcal{N}(\beta^*) \times \mathcal{N}(\theta_k^*)$ with probability 1, where $\mathcal{N}(\beta^*)$ is a neighbourhood of β^*;

Assumption A6. $E\{\sup_{(\beta,\theta_k)\in\mathcal{N}(\beta^*)\times\mathcal{N}(\theta_k^*)} \|\partial u_k(X,\beta,\theta_k)/\partial\beta\|\} < \infty$;

Assumption A7. $\partial u_k(X;\beta^*,\theta_k)/\partial\theta_k$ is continuous at each $\theta_k \in \mathcal{N}(\theta_k^*)$ with probability 1;

Assumption A8. $E\{\sup_{\theta_k\in\mathcal{N}(\theta_k^*)} \|\partial u_k(X,\beta^*,\theta_k)\|/\partial\theta_k\} < \infty$;

Assumption A9. $\Delta(\beta^*,\theta^*)$ exists and is finite, and $\Gamma(\beta^*,\theta^*)$ is of full rank.

Details of Assumption A1

In practice it is sometimes difficult to check the global identification condition. This motivates us to investigate conditions for local identifiability or, equivalently, the invertibility of the matrix of second derivatives at the true parameter, i.e., $\partial^2 Q(\beta)/\partial\beta^2|_{\beta=\beta^*} = [E\{\partial U(X;\beta)/\partial\beta\}^{\mathrm{T}} CE\{\partial U(X;\beta)/\partial\beta\}]|_{\beta=\beta^*}$ (Rothenberg, 1971; Engle & McFadden, 1994), assuming C is a positive-definite matrix. The condition can be stated in terms of the equivalent sample version of the matrix, given by $X_{\mathrm{rbind}}^{\mathrm{T}} W X_{A_{\mathrm{diag}}} C X_{A_{\mathrm{diag}}}^{\mathrm{T}} W X_{\mathrm{rbind}}$. As C is a positive-definite matrix, the entire local identifiability condition for the sample version then boils down to $X_{A_{\mathrm{diag}}}^{\mathrm{T}} W X_{\mathrm{rbind}}$ being a matrix of full column rank. A sufficient condition for this is the matrix $X_{A_{\mathrm{diag}}}$ to have information on all the covariates of the maximal model. In other words, the individual covariates in the maximal model have to be part of at least one of the reduced models.

REFERENCES

ADAMS, P. F., HENDERSHOT, G. E. & MARANO, M. A. (1999). Current estimates from the National Health Interview Survey, 1996. *Vital Health Statist.* **10**, 1–203.

BLOOM, B., COHEN, R. & FREEMAN, G. (2010). Summary health statistics for U.S. children: National Health Interview Survey, 2009. *Vital Health Statist.* **10**, 1–82.

BOTMAN, S. & MORIARITY, C. L. (2000). Design and estimation for the National Health Interview Survey, 1995–2004. *Vital Health Statist.* **2**, 1–31.

BRESLOW, N. E. & CAIN, K. C. (1988). Logistic regression for two-stage case control data. *Biometrika* **75**, 11–20.

BRESLOW, N. E. & HOLUBKOV, R. (1997). Maximum likelihood estimation for logistic regression parameters under two-phase, outcome-dependent sampling. *J. R. Statist. Soc.* B **59**, 447–61.

BULIK-SULLIVAN, B. K., LOH, P.-R., FINUCANE, H., RIPKE, S., YANG, J., PATTERSON, N., DALY, M. J., PRICE, A. L. & NEALE, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genet.* **47**, 291–5.

CHATTERJEE, N., CHEN, Y. H., MASS, P. & CARROLL, R. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Am. Statist. Assoc.* **111**, 891–921.

CHEN, J., PEE, D., AYYAGARI, R., GRAUBARD, B., SCHAIRER, C., BYRNE, C., BENICHOU, J. & GAIL, M. H. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *J. Nat. Cancer Inst.* **98**, 1215–26.

CHENG, W., TAYLOR, J. M. G., GU, T., TOMLINS, S. A. & MUKHERJEE, B. (2019). Informing a risk prediction model for binary outcomes with external coefficient information. *Appl. Statist.* **68**, 121–39.

CHUN, W., CHEN, M. & SCHIFANO, E. (2015). Statistical methods and computing for big data. arXiv: 1502.07989v2.

DE FERRANTI, S. D., GAUVREAU, K., LUDWIG, D. S., NEWBURGER, J. W. & RIFAI, N. (2006). Inflammation and changes in metabolic syndrome abnormalities in US adolescents: Findings from the 1988–1994 and 1999–2000 National Health and Nutrition Examination Surveys. *Clin. Chem.* **52**, 1325–30.

DERSIMONIAN, R. & LAIRD, N. (1986). Meta-analysis in clinical-trials. *Contr. Clin. Trials* **7**, 177–88.

DERSIMONIAN, R. & LAIRD, N. (2015). Meta-analysis in clinical trials revisited. *Contemp. Clin. Trials* **45**, 139–45.

ENGLE, R. & MCFADDEN, D. (1994). *Handbook of Econometrics*. Amsterdam: North Holland.

ESTES, J. P., MUKHERJEE, B. & TAYLOR, J. M. G. (2018). Empirical Bayes estimation and prediction using summary-level information from external big data sources adjusting for violations of transportability. *Statist. Biosci.* **10**, 568–86.

FAN, J., HAN, F. & LIU, H. (2014). Challenges of big data analysis. *Nat. Sci. Rev.* **1**, 293–314.

FANG, J. & ALDERMAN, M. (2000). Serum uric acid and cardiovascular mortality: The NHANES I epidemiologic follow-up study, 1971–1992. *J. Am. Med. Assoc.* **283**, 2404–10.

Han, P. & Lawless, J. F. (2017). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statist. Sinica*, DOI: 10.5705/ss.202017.0308.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–54.

He, J., Ogden, L., Bazzano, L., Vupputuri, S., Loria, C. & Whelton, P. (2001). Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study. *Arch. Intern. Med.* **161**, 996–1002.

Idler, E. L. & Angel, R. J. (2011). Self-rated health and mortality in the NHANES-I epidemiologic follow-up study. *Am. J. Public Health* **80**, 446–52.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *J. Bus. Econ. Statist.* **20**, 493–506.

Ioannidis, J. P. A. (2005). Meta-analysis in public health: Potentials and problems. *Eur. J. Public Health* **15**, 60–1.

Jackson, D., Riley, R. & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statist. Med.* **30**, 2481–98.

Jordan, M. I. (2013). On statistics, computation and scalability. *Bernoulli* **19**, 1378–90.

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

LaKind, J. S., Goodman, M. & Naiman, D. Q. (2012). Use of NHANES data to link chemical exposures to chronic diseases: A cautionary tale. *PLoS One* **8**, 1295–302.

Lee, S. H., Yang, J., Chen, G. B., Ripke, S., Stahl, E. A., Hultman, C. M., Sklar, P., Visscher, P. M., Sullivan, P. F., Goddard, M. E. et al. (2013). Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151–5.

Lin, D. Y. & Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97**, 321–32.

Mass, P., Barrdahl, M., Joshi, A. D., Auer, P. L., Gaudet, M. M., Milne, R. L., Schumacher, F. R., Anderson, W. F., Check, D., Chattopadhyay, S. et al. (2016). Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2**, 1295–302.

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

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Boca Raton, Florida: Chapman & Hall/CRC, 2nd ed.

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

Pasaniuc, B. & Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Rev. Genet.* **18**, 117–27.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika* **87**, 484–90.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–25.

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Rahmandad, H., Jalali, M. S. & Paynabar, K. (2017). A flexible method for aggregation of prior statistical findings. *PloS One* **12**, e0175111.

Ritz, J., Demidenko, E. & Spiegelman, D. (2008). Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects. *J. Statist. Plan. Infer.* **138**, 1919–33.

Rothenberg, T. (1971). Identification in parametric models. *Econometrica* **39**, 577–91.

Scott, A. J. & Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 705–17.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.

van Houwelingen, H. C., Arends, L. R. & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statist. Med.* **21**, 589–624.

Wacholder, S. & Carroll, R. J. (1994). The partial questionnaire design for case-control studies. *Statist. Med.* **13**, 623–34.

Wang, F., Song, P. X.-K. & Wang, L. (2015). Merging multiple longitudinal studies with study-specific missing covariates: A joint estimating function approach. *Biometrics* **71**, 929–40.

Whittemore, A. (1997). Multistage sampling designs and estimating equations. *J. R. Statist. Soc.* B **59**, 589–602.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N. & Loos, R. J. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genet.* **44**, 369–75.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M. et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genet.* **48**, 481–7.

[*Received on* 22 *October* 2017. *Editorial decision on* 5 *December* 2018]