Received: 16 November 2016 | Revised: 10 May 2017 | Accepted: 11 June 2017

DOI: 10.1002/sim.7398

RESEARCH ARTICLE

WILEY Statistics in Medicine

Meta-analysis of gene-environment interaction exploiting gene-environment independence across multiple case-control studies

Jason P. Estes[1] | John D. Rice[1] | Shi Li[2] | Heather M. Stringham[1] | Michael Boehnke[1] | Bhramar Mukherjee[1,3]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI48109, U.S.A.

[2]Genentech, 1 DNA Way South San Francisco, CA94080, U.S.A.

[3]Department of Epidemiology, University of Michigan, Ann Arbor, MI48109, U.S.A.

Correspondence
Bhramar Mukherjee, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.
Email: bhramar@umich.edu

Funding information
National Science Foundation, Grant/Award Number: 1406712 ; National Institute of Health, Grant/Award Number: 20811 ; DK062370

Multiple papers have studied the use of gene-environment (*G-E*) independence to enhance power for testing gene-environment interaction in case-control studies. However, studies that evaluate the role of *G-E* independence in a meta-analysis framework are limited. In this paper, we extend the single-study empirical Bayes type shrinkage estimators proposed by Mukherjee and Chatterjee (2008) to a meta-analysis setting that adjusts for uncertainty regarding the assumption of *G-E* independence across studies. We use the retrospective likelihood framework to derive an adaptive combination of estimators obtained under the constrained model (assuming *G-E* independence) and unconstrained model (without assumptions of *G-E* independence) with weights determined by measures of *G-E* association derived from multiple studies. Our simulation studies indicate that this newly proposed estimator has improved average performance across different simulation scenarios than the standard alternative of using inverse variance (covariance) weighted estimators that combines study-specific constrained, unconstrained, or empirical Bayes estimators. The results are illustrated by meta-analyzing 6 different studies of type 2 diabetes investigating interactions between genetic markers on the obesity related *FTO* gene and environmental factors body mass index and age.

KEYWORDS

case-control study, efficiency, empirical Bayes, individual patient data, meta-analysis, type 2 diabetes

1 | INTRODUCTION

Studies suggest that the risks of many complex diseases depend on the combined effects of genetic susceptibility factors *G* and environmental exposures *E*. Studies of *G-E* interactions (GEI), particularly for rare exposures, require large sample sizes and efficient designs. Exploiting independence between the genetic and environmental factors in case-control studies to gain efficiency has been noted by several authors.[1-3] In particular, Chatterjee and Carroll[3] studied the semiparametric maximum likelihood estimates of logistic regression parameters that exploit the *G-E* independence assumption in a general regression setting that may involve continuous exposures, nonrare diseases, and other stratification variables. While Chatterjee and Carroll[3] alleviates many of the limitations of prior work, retrospective methods that assume *G-E* independence have the potential to yield severely biased estimates and inflated type 1 errors when the assumption is violated. Several studies have addressed this issue and proposed more robust strategies for testing GEI.[4-7] For example, using the retrospective likelihood framework in Chatterjee and Carroll[3], Mukherjee and Chatterjee[4] proposed an adaptive estimator that does not impose the independence assumption exactly and allows for uncertainty in the assumption of gene-environment independence. The *G-E* log-odds ratio parameter is estimated in an empirical Bayes (EB) fashion to arrive at a final shrinkage estimator that "shrinks" the semiparametric retrospective maximum likelihood estimates under *G-E* dependence to those under *G-E* independence to trade off between bias and efficiency.

Detecting gene-environment interactions with small effect sizes will often require a meta-analytic approach. There are several methods of meta-analyzing a single scalar gene-environment interaction effect across studies. For example, one can use an inverse-variance weighted fixed-effect approach for the GEI parameter[8-10] when individual patient data (IPD) are not available. To meta-analyze a parameter vector, one can use an inverse variance-covariance weighted estimator.[11,12] Alternatively, when individual patient–level data from all studies are available, the data can be analyzed simultaneously, commonly called joint analysis or mega-analysis. Furthermore, Lin and Zeng[10,11] showed that meta-analysis based on summary statistics has the same asymptotic efficiency as the maximum likelihood estimates (MLE) resulting from the full data if the former analysis is performed jointly on all common parameters across studies. One can easily incorporate the work of Mukherjee and Chatterjee[4] into the meta-analysis framework by using the aforementioned inverse-variance or inverse variance-covariance approach with study specific EB estimators; however, such an approach does not directly borrow information across studies with respect to the uncertainty around the *G-E* independence assumption.

To date, there are no papers that study the role of *G-E* independence in a meta-analysis framework where uncertainty in the assumption can vary across studies. In this work, we consider several multiple-study EB (MSEB) type shrinkage estimators that extend the EB type shrinkage estimators proposed in Mukherjee and Chatterjee[4] to a multiple-study setting that can borrow information across studies. Furthermore, our MSEB estimators can be readily constructed using existing software such as CGEN[13], making our proposed estimators easily implementable. We propose MSEB estimators in cases where (1) IPD are available and (2) only study level summary statistics are available.

Our paper is organized as follows. We introduce the proposed MSEB estimators in Section 2.2, and simulation studies are performed in Section 3. In Section 4, we illustrate our methods by meta-analysis of *G-E* interactions of single nucleotide polymorphisms SNPs on *FTO* gene with body mass index (BMI) and age using data from 6 different studies of type 2 diabetes. Concluding remarks are presented in Section 5.

2 | PROPOSED MSEB TYPE SHRINKAGE ESTIMATORS

2.1 | Model specification

Let $D = 1(D = 0)$ denote the presence (absence) of a disease, G denote a genetic factor, E denote an environmental exposure, and $\mathbf{S}$ denote a vector of covariates. The subscript $k = 1, \dots, K$ is used to index K independent studies, and the subscript $i = 1, \dots, n_k$ is used to index individuals within the kth study of size n_k. Consider the following factorization of the retrospective likelihood akin to Chatterjee and Carroll[3],

$$\begin{aligned} L^R &= \prod_{k=1}^{K}\prod_{i=1}^{n_k} \text{pr}\,(G_{ki}, E_{ki}, S_{ki}|D_{ki}) \\ &= \prod_{k=1}^{K}\prod_{i=1}^{n_k} \frac{\text{pr}\,(D_{ki}|G_{ki}, E_{ki}, S_{ki})\ \text{pr}\,(G_{ki}|E_{ki}, S_{ki})\ \text{pr}\,(E_{ki}, S_{ki})}{\sum_{G,E,S} \text{pr}\,(D_{ki}|G, E, S)\ \text{pr}\,(G|E, S)\ \text{pr}\,(E, S)}. \end{aligned} \tag{1}$$

For continuous exposure E, the sum with respect to E in the denominator of 1 is replaced by an integral. The components of the retrospective likelihood are modeled as follows. Assume a logistic disease incidence model

$$\text{pr}\,(D_{ki}|G_{ki}, E_{ki}, S_{ki}) = H\{\gamma_{0k} + m(G_{ki}, E_{ki}, \mathbf{S}_{ki}; \boldsymbol{\gamma})\} \tag{2}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}, m(\cdot)$ is a known but arbitrary function, γ_{0k} are intercept parameters, and $\boldsymbol{\gamma}$ is a vector of parameters of interest. In this model specification, the intercept is allowed to vary with respect to the K studies, whereas the parameter vector, $\boldsymbol{\gamma}$, of log odds ratios associated with G, E and S is shared among the studies. For a dominant susceptibility model of G, we consider a logistic model

$$\text{pr}\,(G_{ki} = 1|E_{ki}, \mathbf{S}_{ki}) = H\{\eta_{0k} + \boldsymbol{\eta}_k\mathbf{S}_{ki} + \theta_k E_{ki}\}, \tag{3}$$

where θ_k are study level nuisance parameters that measure dependence between G and E within the kth study, η_{0k} are intercept parameters, and $\boldsymbol{\eta}_k$ are study-specific row vectors of parameters corresponding to individual covariates. Under the assumption of G-E independence (conditional on $\mathbf{S}$) within each study k, the parameters θ_k are all set to 0, and model 3 reduces to

$$\text{pr}\,(G_{ki} = 1|E_{ki}, \mathbf{S}_{ki}) = H\left\{\eta^0_{0k} + \boldsymbol{\eta}^0_k\mathbf{S}_{ki}\right\}. \tag{4}$$

For an additive susceptibility model of G, one might consider a proportional odds model for pr$(G|E, S)$. For a codominant susceptibility model of G, one might consider polychotomous logistic regression. In 3, one can alternatively model these probabilities under Hardy-Weinberg equilibrium[14] (see the Supporting Information). Finally, the joint distribution function for $(E, \mathbf{S})$ is allowed to remain completely nonparametric.[3]

The aforementioned model formulation is quite flexible in allowing parameters to depend on k. For example, one may assume a common $G - S$ association across studies in model 3, ie, $\boldsymbol{\eta}_k = \eta$ for $k = 1, \dots, K$ and some constant η. Similarly, one may require $\theta_k = \theta$ for $k = 1, \dots, K$. We proceed with the most general formulation $(\boldsymbol{\eta}_k, \theta_k)$ in model 3 but assume a shared common effect $\boldsymbol{\gamma}$ among the K studies in model 2. Different choices are investigated in our data application.

2.2 | MSEB shrinkage estimators

In this section, we extend the EB shrinkage estimator proposed in Mukherjee and Chatterjee[4] to an appropriate MSEB estimator. In Section 2.2.1, we detail our proposed estimators under the assumption that IPD are available for each study, and in Section 2.2.2, we detail our proposed estimators using summary (aggregate) data from each of the K studies.

When one is not certain about the G-E independence across the k studies, one may conceptually posit a stochastic framework for the underlying true parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K) \sim MVN(\mathbf{0}, \mathbf{A})$ where $\mathbf{0}$ is a $K \times 1$ vector of zeros and $\mathbf{A}$ is a $K \times K$ diagonal matrix whose nonzero elements are all equal to some nonnegative constant τ^2, which reflects a measure of uncertainty about the independence assumption. This is the stochastic framework governing the methods we present subsequently.

2.2.1 | IPD analysis

Let $\boldsymbol{\beta} = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\eta}_0, \boldsymbol{\eta})^{\mathrm{T}}$ denote the focus parameters of the unconstrained model 3 where $\boldsymbol{\gamma}_0 = (\gamma_{01}, \dots, \gamma_{0K})^{\mathrm{T}}, \boldsymbol{\eta}_0 = (\eta_{01}, \dots, \eta_{0K})^{\mathrm{T}}, \boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_K)^{\mathrm{T}}$ and $\boldsymbol{\gamma}$ represents a parameter vector shared among the K studies. A superscript of 0 will be used to denote the corresponding parameters under the constrained model 4, eg, $\boldsymbol{\beta}^0 = (\boldsymbol{\gamma}^0_0, \boldsymbol{\gamma}^0, \boldsymbol{\eta}^0_0, \boldsymbol{\eta}^0)^{\mathrm{T}}$. The MLEs $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{\beta}}^{\,0}$ for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\boldsymbol{\beta}^0$ are obtained, along with their estimated asymptotic variances $\widehat{\boldsymbol{V}}_{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}$ and $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}^{\,0}}$, using the profile-likelihood techniques of Chatterjee and Carroll[3], respectively. Intuitively, given $\boldsymbol{\theta}$ and in the absence of any prior information on $\boldsymbol{\beta}$, a natural way to estimate $\boldsymbol{\beta}$ is to use $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, the profile MLE of $\boldsymbol{\beta}$ for a fixed $\boldsymbol{\theta}$. Define $\boldsymbol{\beta}(\boldsymbol{\theta})$ to be the limiting value of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, which is a population parameter when $\boldsymbol{\theta}$ is fixed at the true value. The estimate $\widehat{\boldsymbol{\beta}}(\mathbf{0})$ denotes the profile MLE of $\boldsymbol{\beta}$ under the constrained model when $\boldsymbol{\theta} = \mathbf{0}$. The goal is to obtain an estimator of $\boldsymbol{\gamma}(\boldsymbol{\theta})$, the common set of parameters in the disease incidence model shared among the K studies, which takes into account the uncertainty about the G-E independence assumption on $\boldsymbol{\theta}$, which may vary across studies. Thus, we developed weighted estimators of $\boldsymbol{\gamma}(\boldsymbol{\theta})$ whose weights shrink the estimates of $\boldsymbol{\gamma}(\boldsymbol{\theta})$ towards $\widehat{\boldsymbol{\gamma}}(\mathbf{0})$ when there is less uncertainty regarding G-E independence.

To achieve the goal of developing our MSEB estimators of $\boldsymbol{\gamma}(\boldsymbol{\theta})$, we first approximate the distributions of $\boldsymbol{\beta}(\boldsymbol{\theta})$ and $\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})$ using the prior distribution $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \mathbf{A})$. A first-order Taylor's expansion of $\boldsymbol{\beta}(\boldsymbol{\theta})$ about $\boldsymbol{\theta} = \mathbf{0}$ gives

$$\boldsymbol{\beta}(\boldsymbol{\theta}) \approx \boldsymbol{\beta}(\mathbf{0}) + \Delta^{\mathrm{T}}\boldsymbol{\theta} \tag{5}$$

where $\Delta^{\mathrm{T}} \equiv \partial\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\mathbf{0}}$ is the gradient matrix evaluated at $\boldsymbol{\theta}$=$\mathbf{0}$. Thus, we can approximate the distribution of $\boldsymbol{\beta}(\boldsymbol{\theta})$ via $MVN(\boldsymbol{\beta}(\mathbf{0}), \Delta^{\mathrm{T}}\boldsymbol{A}\Delta)$. Finally, we approximate the distribution of $\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})$ via its asymptotic distribution $MVN\{\boldsymbol{\beta}(\boldsymbol{\theta}), \mathbf{V}_{\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})}\}$ leading to the Bayes estimate of $\boldsymbol{\beta}(\boldsymbol{\theta})$ as the posterior mean

$$\Delta^{\mathrm{T}}\boldsymbol{A}\Delta\{\boldsymbol{V}_{\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})} + \Delta^{\mathrm{T}}\boldsymbol{A}\Delta\}^{-1}\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) + \boldsymbol{V}_{\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})}\{\boldsymbol{V}_{\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})} + \Delta^{\mathrm{T}}\boldsymbol{A}\Delta\}^{-1}\boldsymbol{\beta}(\mathbf{0}) \tag{6}$$

of $\boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})$ under our Gaussian-Gaussian model. Our Bayes estimate of $\boldsymbol{\gamma}$ is taken to be the corresponding subvector of our Bayes estimate of $\boldsymbol{\beta}(\boldsymbol{\theta})$. The components of the weights in 6 are estimated as follows. Replace $\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})$ and $\boldsymbol{\beta}(\mathbf{0})$ with $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{\beta}}(\mathbf{0})$, respectively, and replace $\boldsymbol{V}_{\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})}$ with the corresponding submatrix of $\widehat{\boldsymbol{V}}_{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}$. From the Taylor's approximation in 5, we approximate $\{\boldsymbol{\beta}(\boldsymbol{\theta}) - \boldsymbol{\beta}(\mathbf{0})\}\boldsymbol{\theta}^{\mathrm{T}}$ via $\Delta^{\mathrm{T}}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}$. The matrix $\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}$ is not invertible when $K > 1$, so we use its Moore-Penrose inverse $(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}})^{+}$ leading to the use of $\Delta^{\mathrm{T}}(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}})(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}})^{+}$ as an approximation to $\{\boldsymbol{\beta}(\boldsymbol{\theta}) - \boldsymbol{\beta}(\mathbf{0})\}\boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}})^{+}$. In general, $(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}})(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}})^{+}$ is not equal to the identity matrix $\boldsymbol{I}_K$ of dimension $K\times K$, so we replace it with its expectation. The matrix $(\boldsymbol{\theta\theta}^{\mathrm{T}})(\boldsymbol{\theta\theta}^{\mathrm{T}})^{+}$ has expectation $K^{-1}\boldsymbol{I}_K$ and variance $(K^{-1}-K^{-2})I_K$, yielding our final approximation $K\{\boldsymbol{\beta}(\boldsymbol{\theta})-\boldsymbol{\beta}(\mathbf{0})\}\boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{\theta\theta}^{\mathrm{T}})^{+}$ of Δ^{T} (see Theorem 1 in the Supporting Information). Because $\boldsymbol{\beta}(\boldsymbol{\theta})$, $\boldsymbol{\beta}(\mathbf{0})$, and $\boldsymbol{\theta}$ are unknown, we replace them with their estimates. We consider 4 different estimates of $\boldsymbol{A}=\tau^2\boldsymbol{I}_K$. If estimators (ii) to (iv) result in a negative value, we take $\hat{\tau}^{\,2}$ to be 0 (the well-known positive part estimator).[15] The estimator presented in (i) below will be serve as the primary estimator of $\boldsymbol{A}$ throughout the paper, whereas estimators (ii) to (iv) are presented as alternative natural approaches that other practitioners may think of.

- Our first estimate $K^{-1}(\hat{\boldsymbol{\theta}}^{\mathrm{T}}\hat{\boldsymbol{\theta}})$ of τ^2 is conservative and motivated by the asymptotic distribution of $\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}$ marginalized over $\boldsymbol{\theta}$, to wit, $N(\mathbf{0},\boldsymbol{A}+\boldsymbol{V}_{\hat{\theta}})$.
- Our second estimate $K^{-1}\{\hat{\boldsymbol{\theta}}^{\mathrm{T}}\hat{\boldsymbol{\theta}}-tr(\hat{\boldsymbol{V}}_{\hat{\theta}})\}$ of τ^2 adjusts for the conservative nature of our first estimate.
- Our third estimate of τ^2 is motivated by maximizing the log marginal likelihood obtained from the multivariate density $N(\mathbf{0},\boldsymbol{A}+\boldsymbol{V}_{\hat{\theta}})$ with respect to τ^2 given $\boldsymbol{V}_{\hat{\theta}}=\hat{\boldsymbol{V}}_{\hat{\theta}}$. Let $\{\hat{v}_1,\ \dots\ ,\hat{v}_K\}$ denote the diagonal elements of $\hat{\boldsymbol{V}}_{\hat{\theta}}$. We maximize the marginal likelihood
$$\mathcal{L}(\tau^2|\hat{\boldsymbol{\theta}},\boldsymbol{V}_{\hat{\theta}}=\hat{\boldsymbol{V}}_{\hat{\theta}})=\prod_{k=1}^{K}\left\{2\pi(\tau^2+\hat{v}_k)\right\}^{-\frac{1}{2}}\exp\left\{-\frac{\hat{\theta}_k^{\,2}}{2(\tau^2+\hat{v}_k)}\right\} \tag{7}$$
with respect to τ^2 by setting the derivative
$$\frac{d}{d\tau^2}\left[\log\left\{\mathcal{L}(\tau^2|\hat{\boldsymbol{\theta}},\boldsymbol{V}_{\hat{\theta}}=\hat{\boldsymbol{V}}_{\hat{\theta}})\right\}\right]=-\frac{1}{2}\sum_{k=1}^{K}\left\{\frac{1}{\tau^2+\hat{v}_k}-\frac{\hat{\theta}_k^{\,2}}{(\tau^2+\hat{v}_k)^2}\right\}$$
equal to 0. We implement the uniroot function in R to numerically approximate $\hat{\tau}^{\,2}$.
- Our fourth estimate of τ^2 results from an iterative process proposed in Morris[16,17], extending our second estimate by considering weights (other than K^{-1}) that depend on variances as follows. The update of $\hat{\tau}^2_{(n)}$ at iteration $n+1$ is given by
$$\hat{\tau}^2_{(n+1)}=\left\{\sum_{k=1}^{K}w_{k,(n)}\right\}^{-1}\sum_{k=1}^{K}w_{k,(n)}\left\{\hat{\theta}_k^{\,2}-\hat{v}_k\right\}, \tag{8}$$
where $w_{k,(n)}=\left\{\hat{v}_k+\hat{\tau}^2_{(n)}\right\}^{-1}$ and the initial guess $\tau^2_{(0)}$ is the estimate resulting from the maximization of the marginal likelihood in 7. Our estimate of τ^2 is taken to be $\tau^2_{(m)}$ where $m\in\mathbb{N}$ is some iteration step (greater than 0) such that $|\tau^2_{(m+1)}-\tau^2_{(m)}|<\varepsilon$ where ε is some positive tolerance value, which we take to be 10^{-8}.

We denote our EB estimates of $\boldsymbol{\gamma}$ resulting from the 4 proposed estimators (i), (ii), (iii), and (iv) of $\boldsymbol{A}$ by $\hat{\boldsymbol{\gamma}}_{\mathrm{EB1}}$, $\hat{\boldsymbol{\gamma}}_{\mathrm{EB2}}$, $\hat{\boldsymbol{\gamma}}_{\mathrm{EB3}}$, and $\hat{\boldsymbol{\gamma}}_{\mathrm{EB4}}$, respectively. We note in the case that $\hat{\tau}^2$ is estimated to be 0; Equation 6 reduces to $\boldsymbol{\beta}(\mathbf{0})$, which is estimated via constrained maximum likelihood (CML). This property affects estimators $\hat{\boldsymbol{\gamma}}_{\mathrm{EB2}}-\hat{\boldsymbol{\gamma}}_{\mathrm{EB4}}$ but does not affect $\hat{\boldsymbol{\gamma}}_{\mathrm{EB1}}$ (with probability 1) because of the conservative nature of (i). We refer the reader to variance approximations of these estimators in the Supporting Information. The operating characteristics of our estimators are evaluated via simulation study in Section 3.

2.2.2 | Meta-analysis using summary measures

In the absence of individual level data, we consider a meta-analytic approach using effect and variance estimates. Within each study, we denote the MLEs of $\boldsymbol{\beta}_k=(\gamma_{0k},\boldsymbol{\gamma}_k,\eta_{0k},\boldsymbol{\eta}_k)^{\mathrm{T}}$ and θ_k under the unconstrained model by $\widetilde{\boldsymbol{\beta}}_k=(\tilde{\gamma}_{0k},\widetilde{\boldsymbol{\gamma}}_k,\tilde{\eta}_{0k},\widetilde{\boldsymbol{\eta}}_k)^{\mathrm{T}}$ and $\tilde{\theta}_k$, respectively, and the MLEs of $\boldsymbol{\beta}_k^0=\left(\gamma_{0k}^0,\boldsymbol{\gamma}_k^0,\eta_{0k}^0,\boldsymbol{\eta}_k^0\right)^{\mathrm{T}}$ under the constrained model by $\widetilde{\boldsymbol{\beta}}_k^{\ 0}=\left(\tilde{\gamma}_{0k}^{\ 0},\widetilde{\boldsymbol{\gamma}}_k^{\ 0},\tilde{\eta}_0^{\ 0},\widetilde{\boldsymbol{\eta}}_k^{\ 0}\right)^{\mathrm{T}}$. We use $\boldsymbol{V}_{\widetilde{\boldsymbol{\alpha}}}$ to denote the covariance matrix of a generic parameter estimate $\widetilde{\boldsymbol{\alpha}}$, and $\widetilde{\boldsymbol{V}}_{\tilde{\alpha}}$ will be used to denote its covariance estimate. Intuitively, given θ_k and in the absence of any prior information on $\boldsymbol{\beta}_k$, a natural way to estimate $\boldsymbol{\beta}_k$ is to use $\widetilde{\boldsymbol{\beta}}_k(\theta_k)$, the profile MLE of $\boldsymbol{\beta}_k$ for a fixed θ_k. To combine the information of the K studies, we consider the inverse variance-covariance meta-analysis estimates of the common focus parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^0$ given by
$$\widetilde{\boldsymbol{\gamma}}=\left\{\sum_k\widetilde{\boldsymbol{V}}^{-1}_{\widetilde{\boldsymbol{\gamma}}_k}\right\}^{-1}\sum_k\widetilde{\boldsymbol{V}}^{-1}_{\widetilde{\boldsymbol{\gamma}}_k}\widetilde{\boldsymbol{\gamma}}_k \text{ and } \widetilde{\boldsymbol{\gamma}}^{\ 0}=\left\{\sum_k\widetilde{\boldsymbol{V}}^{-1}_{\widetilde{\boldsymbol{\gamma}}_k^{\ 0}}\right\}^{-1}\sum_k\widetilde{\boldsymbol{V}}^{-1}_{\widetilde{\boldsymbol{\gamma}}_k^{\ 0}}\widetilde{\boldsymbol{\gamma}}_k^{\ 0}$$
with variance estimates $\left\{\sum_k\widetilde{\boldsymbol{V}}^{-1}_{\widetilde{\boldsymbol{\gamma}}_k}\right\}^{-1}$ and $\left\{\sum_k\widetilde{\boldsymbol{V}}^{-1}_{\widetilde{\boldsymbol{\gamma}}_k^{\ 0}}\right\}^{-1}$, respectively. These meta-analysis estimators can be viewed as functions of $\boldsymbol{\theta}$ because $\boldsymbol{\gamma}_{\underset{\sim}{k}}$ are estimated using the profile MLEs $\widetilde{\boldsymbol{\beta}}_k(\theta_k)$ for a fixed θ_k. Thus, $\widetilde{\boldsymbol{\gamma}}\equiv\widetilde{\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}})$ and $\widetilde{\boldsymbol{\gamma}}^{\ 0}\equiv\widetilde{\boldsymbol{\gamma}}^{\ 0}(0)$ can be viewed as estimates of $\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}})$ and $\boldsymbol{\gamma}(\mathbf{0})$, respectively.

Similar to Section 2.2.1, we use the prior distribution $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \mathbf{A})$ and a first-order Taylor's expansion $\boldsymbol{\gamma}(\boldsymbol{\theta}) \approx \boldsymbol{\gamma}(\mathbf{0}) + \Delta^{\mathrm{T}}\boldsymbol{\theta}$, where $\Delta^{\mathrm{T}} \equiv \partial\boldsymbol{\gamma}^{\mathrm{T}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\theta=0}$, to approximate the distributions of $\boldsymbol{\gamma}(\boldsymbol{\theta})$ and $\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}})$ via $MVN\{\boldsymbol{\gamma}(\mathbf{0}), \Delta^{\mathrm{T}}A\Delta\}$ and $MVN\{\boldsymbol{\gamma}(\boldsymbol{\theta}), V_{\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}})}\}$, respectively. Our EB estimate of $\boldsymbol{\gamma}(\boldsymbol{\theta})$ is taken to be the posterior mean

$$\Delta^{\mathrm{T}}\boldsymbol{A}\Delta\{\boldsymbol{V}_{\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}})} + \Delta^{\mathrm{T}}\boldsymbol{A}\Delta\}^{-1}\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}}) + \boldsymbol{V}_{\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}})}\{\boldsymbol{V}_{\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}})} + \Delta^{\mathrm{T}}\boldsymbol{A}\Delta\}^{-1}\boldsymbol{\gamma}(\mathbf{0}) \tag{9}$$

of $\boldsymbol{\gamma}(\boldsymbol{\theta})|\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}})$ under our Gaussian-Gaussian model. The components of the weights in 9 are estimated as follows. Replace $\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}}), \boldsymbol{\gamma}(\mathbf{0})$, $\boldsymbol{V}_{\boldsymbol{\gamma}(\widetilde{\boldsymbol{\theta}})}$, and Δ^{T} with $\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\gamma}}^0, \left\{\sum_k \widetilde{\boldsymbol{V}}_{\widetilde{\boldsymbol{\gamma}}_k}^{-1}\right\}^{-1}$ and $K\{\widetilde{\boldsymbol{\gamma}}(\theta) - \widetilde{\boldsymbol{\gamma}}(\mathbf{0})\}\theta^{\mathrm{T}}(\theta\theta^{\mathrm{T}})^{+}$ (see Section 2.2.1), respectively. Finally, we consider 4 different estimates of $\boldsymbol{A}$ as defined in Section 2.2.1 except that the estimates in the expressions are not the result of IPD but the result of meta-analysis of the K independent studies (hats are replaced with tildes). We denote these estimators by $\widetilde{\boldsymbol{\gamma}}_{\mathrm{EB1}}, \widetilde{\boldsymbol{\gamma}}_{\mathrm{EB2}}, \widetilde{\boldsymbol{\gamma}}_{\mathrm{EB3}}$, and $\widetilde{\boldsymbol{\gamma}}_{\mathrm{EB4}}$, respectively. For convenience, we may refer to these estimators via EB1, EB2, EB3, and EB4, respectively, (IPD joint analysis vs summary statistic meta-analysis will be clear from context). Alternatively, one can use the inverse variance-covariance meta-analysis estimates of the focus parameters resulting from single-study EB estimation, eg,

$$\widetilde{\gamma}_{EB} = \left\{\sum_k \widetilde{\boldsymbol{V}}_{\widetilde{\gamma}_k^{EB}}^{-1}\right\}^{-1} \sum_k \widetilde{\boldsymbol{V}}_{\widetilde{\gamma}_k^{EB}}^{-1} \widetilde{\boldsymbol{\gamma}}_k^{EB}$$

where $\widetilde{\boldsymbol{\gamma}}_k^{EB}$ and $\widetilde{\boldsymbol{V}}_{\widetilde{\gamma}_k^{EB}}$ are defined as in Mukherjee and Chatterjee.[4] For convenience, we may refer to this estimator as EB (IPD joint analysis vs summary statistic meta-analysis will be clear from context). The operating characteristics of our estimators are evaluated and compared with other sensible estimators via simulation study in Section 3.

3 | SIMULATION STUDY

We carry out simulation studies to study the performance of our proposed MSEB estimators (EB1-EB4) in the IPD and meta-analysis settings relative to the standard inverse variance-covariance weighted logistic regression (LOG), unconstrained maximum likelihood (UML), CML, and EB using the statistical R package CGEN. The operating characteristics of interest are bias, average estimated standard error, empirical standard error, and mean squared error (MSE) with estimates defined by $R^{-1}\sum_{r=1}^{R}(\hat{\xi}_r - \xi)$, $R^{-1}\sum_{r=1}^{R}\hat{V}_{\hat{\xi}_r}$, $\{(R-1)^{-1}\sum_{r=1}^{R}(\hat{\xi}_r - \bar{\xi}_r)^2\}^{1/2}$, and $R^{-1}\sum_{r=1}^{R}(\hat{\xi}_r - \xi)^2$ where ξ is the parameter of interest, $\hat{\xi}_r$ is its estimate, $\bar{\xi}_r = R^{-1}\sum_{r=1}^{R}\hat{\xi}_r$, and $\hat{V}_{\hat{\xi}_r}$ is the variance estimate of $\hat{\xi}_r$ in the $r = 1, \dots, R$ study replications. In our simulation studies, we set R equal to 1000.

The components of the model defined in 2 and 3 are as follows. The subscript $i = 1, \dots, n_k$ is used to index subjects, and $k = 1, \dots, K$ is used to index $K = 10$ subcohorts with sample sizes $n_k = 1000 + 100(k-1)$ yielding a total of $n = \sum_{k=1}^{10} n_k = 14500$ subjects. We considered the stratification vector $\boldsymbol{S}_{ki} = (S_{1ki}, S_{2ki})$ where S_{1ki} is a Bernoulli random variable with parameter .5 and S_{2ki} is a normal random variable with mean 0 and standard deviation .5. Environmental exposures were generated via $E_{ki} = \min\{5, \exp(X_{ki})\}$ where $X_{ki}|S_{1ki} = 0 \sim N(0, .5^2)$ and $X_{ki}|S_{1ki} = 1 \sim N(.1, .5^2)$. Conditional on $(E_{ki}, \boldsymbol{S}_{ki})$, a genetic factor (under Hardy-Weinberg equilibrium) was generated via a multinomial random variable with parameters $(1-q_{ki})^2, 2q_{ki}(1-q_{ki})$, and q_{ki}^2 defined by $q_{ki} = H\{\eta_{0k} + \boldsymbol{\eta}_k\boldsymbol{S}_{ki} + \theta_k E_{ki}\}$, where $\eta_{0k} \overset{iid}{\sim} \mathrm{Uniform}(-1.2, -1.0)$, $\boldsymbol{\eta}_k = (\eta_{1k}, \eta_{2k})$, $\eta_{1k} \overset{iid}{\sim} \mathrm{Uniform}(0.1, 0.2)$, $\eta_{2k} \overset{iid}{\sim} \mathrm{Uniform}(0, 0.1)$and θ_k is generated as follows: (1) $\theta_k = 0$ for all k, (2) $\theta_k = .1$ for all k, (3) $\theta_k = -.5$ for all k, (4) $\theta_k \overset{iid}{\sim} \mathrm{N}(.2, .1^2)$, and (5) $\theta_k \overset{iid}{\sim} \mathrm{Unif}(-.2, .2)$. Binary disease outcome D_{ki} for subject i belonging to the kth study was generated from a Bernoulli random variable with rate parameter defined by $H\{\gamma_{0k} + \gamma_G G_{ki} + \gamma_E E_{ki} + \gamma_{GE} G_{ki} E_{ki} + \boldsymbol{\gamma}_S \boldsymbol{S}_{ki}\}$. We constructed our case-control sample by randomly selecting $n_k/2$ cases and $n_k/2$ controls within the kth subpopulation from the generated population data $\{(D_{ki}, G_{ki}, E_{ki}, \boldsymbol{S}_{ki}) : i = 1, \dots, N_k; k = 1, \dots, K\}$ with $N_k = 200n_k$. In our simulation setup, the prevalence of disease was approximately 4% and the minor allele frequency was approximately 26% within our K subpopulations. In this section, IPD will refer to combining the generated data for a joint analysis, which uses the retrospective likelihood defined in 1.

3.1 | Simulation results

The simulation results are summarized in terms of bias, standard error, empirical standard error, and MSE under G-E independence (Table 1) and G-E dependence (Tables 2 and 3). For convenience, we multiplied all estimated MSE values by 100 and will refer to these scaled values as MSE. In addition, we restrict our attention only to the G×E interaction effect estimate. Under G-E independence, each of the estimators has an estimated bias close to 0 (within .000-.004) in both the IPD and meta-analysis

TABLE 1 Bias (BIAS), standard errors (SE1), empirical standard errors (SE2) and 100 × MSE (MSE) of $\hat{\gamma}_E$, $\hat{\gamma}_G$, and $\hat{\gamma}_{GE}$ resulting from standard logistic regression (LOG), unconstrained maximum likelihood (UML), constrained maximum likelihood (CML), empirical Bayes (EB), and our proposed multistudy empirical Bayes estimators EB1 to EB4 in both individual patient data (IPD) and summary statistic meta-analysis (META) simulation settings under *G-E* independence over 1000 Monte Carlo runs

	Main effect of *E*				**Main effect of *G***				***G*×*E* Interaction**			
IPD	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.011	.0235	.0236	.067	.007	.0263	.0269	.077	.004	.0271	.0279	.080
UML	.012	.0233	.0233	.068	.006	.0261	.0268	.075	.002	.0263	.0268	.072
CML	.015	.0207	.0207	.064	.006	.0260	.0267	.074	−.004	.0163	.0169	.030
EB	.013	.0214	.0213	.063	.006	.0260	.0267	.074	.001	.0218	.0221	.049
EB1	.012	.0231	.0229	.067	.006	.0261	.0268	.075	.002	.0256	.0257	.066
EB2	.013	.0231	.0222	.066	.006	.0261	.0267	.075	.000	.0237	.0227	.051
EB3	.013	.0216	.0223	.067	.006	.0261	.0267	.075	.000	.0200	.0228	.052
EB4	.013	.0216	.0222	.066	.006	.0261	.0267	.075	.000	.0200	.0227	.052
META	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.010	.0236	.0235	.065	.005	.0263	.0268	.074	.002	.0272	.0276	.076
UML	.011	.0233	.0232	.066	.005	.0261	.0267	.074	.001	.0263	.0267	.071
CML	.014	.0207	.0206	.062	.005	.0261	.0266	.074	−.003	.0164	.0168	.029
EB	.014	.0211	.0208	.062	.005	.0261	.0266	.073	−.002	.0203	.0197	.039
EB1	.012	.0231	.0228	.065	.005	.0261	.0267	.074	.000	.0254	.0253	.064
EB2	.013	.0219	.0217	.063	.006	.0261	.0267	.074	−.001	.0206	.0211	.045
EB3	.013	.0218	.0218	.064	.005	.0261	.0266	.074	−.001	.0203	.0212	.045
EB4	.013	.0218	.0217	.063	.005	.0261	.0267	.074	−.001	.0203	.0212	.045

Note. In the meta-analysis setting, we use the inverse variance-covariance weighted approach to obtain the standard logistic, unconstrained, constrained, and EB results

settings. Additionally, the CML estimator has the smallest MSE (.030 and .029) in comparison with the other estimators (.039-.080) in both the IPD and meta-analysis settings, respectively. Under *G-E* independence, the proposed estimators EB2, EB3, and EB4 perform similarly to the EB estimator with respect to MSE (IPD: .049 vs .051 or .052) under the IPD setting, whereas the proposed estimator EB1 does not. This is explained by the fact that τ^2 is overestimated in EB1 and the other estimators EB2, EB3, and EB4 reduce to the CML estimator when τ^2 is estimated to be 0, which happened in approximately 8% of the Monte Carlo runs. Similar findings are observed (Table 1) in the meta-analysis setting. A contributing factor to these findings, similar to IPD, is that EB2, EB3, and EB4 yielded exactly the same estimates as CML in various individual study analysis. Thus, in both IPD and meta-analysis, EB1 tends to have higher MSE than EB2 to EB4 under *G-E* independence because EB2 to EB4 have the ability to reduce to CML, the most efficient estimator of all estimators considered under *G-E* independence.

In Table 2, we consider modest and strong departure from *G-E* independence by fixing θ_k at .1 and −.5 for all k. In both cases, the CML estimator is severely biased (IPD: .065 and −.331; meta: .065 and −.329) and has the largest MSE (IPD: .449 and 11.034; meta: .450 and 10.858). In both scenarios, our estimators EB1 to EB4 outperform CML and EB with respect to bias (IPD: .002-.007 vs −.331 to .003; meta: .000-.009 vs −.329 to .065) and MSE (IPD: .065-.115 vs .101-11.034; meta: .063-.114 vs .216-10.858). Simulation studies fixing θ_k at a constant c for all k (not reported) suggest that even minor departures from 0, eg, $c = .05$, can produce inflated type I error and severe bias.

In Table 3, we consider modest and mild departure from *G-E* independence by specifying the distributions of θ_k to be independently normal with mean and standard deviation .2 and .1, respectively, or independently uniform with parameters −.2 and .2 for all k. In the case of mild and approximately symmetric departure from *G*×*E* independence (about 0), our MSEB estimators perform similarly to the EB estimator, which has the smallest MSE (.063-.064 vs .059), in the IPD setting. In the meta-analysis setting, our MSEB estimators have the smallest MSE (.061-.062) among all estimators considered. In the case of modest departure under a normal distribution on $\boldsymbol{\theta}$ with nonzero mean, our MSEB estimators perform almost identically to the UML estimator with respect to MSE (IPD: .060 vs .059; meta: .059 vs .059), which has the smallest MSE in both the IPD and meta-analysis settings. Most notably, our MSEB estimators outperform the EB estimators with respect to MSE in both meta-analysis settings considered (.059-.062 vs .077 and .218). A contributing factor to these results is that the inverse variance-covariance weighted EB estimator does not necessarily produce an estimator that lies between the inverse variance-covariance weighted

TABLE 2 Bias (BIAS), estimated standard errors (SE1), empirical standard errors (SE2), and 100 × MSE (MSE) of $\hat{\gamma}_E$, $\hat{\gamma}_G$, and $\hat{\gamma}_{GE}$ resulting from standard logistic regression (LOG), unconstrained maximum likelihood (UML), constrained maximum likelihood (CML), empirical Bayes (EB), and our proposed multistudy empirical Bayes estimators in both individual patient data (IPD) and summary statistic meta-analysis (META) simulation settings when *G-E* independence is violated

$\theta_k = .1$ for all k	**Main effect of *E***				**Main effect of *G***				***G*×*E* Interaction**			
IPD	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.012	.0251	.0249	.077	.007	.0257	.0253	.069	.006	.0262	.0260	.072
UML	.014	.0247	.0244	.078	.006	.0255	.0251	.066	.004	.0253	.0246	.062
CML	−.023	.0215	.0216	.098	.011	.0255	.0250	.074	.065	.0158	.0155	.449
EB	.001	.0253	.0251	.063	.011	.0255	.0250	.074	.016	.0274	.0274	.101
EB1	.013	.0248	.0245	.077	.006	.0255	.0251	.066	.005	.0256	.0249	.065
EB2	.012	.0257	.0252	.078	.006	.0256	.0251	.067	.007	.0272	.0271	.078
EB3	.012	.0247	.0252	.078	.006	.0255	.0252	.067	.007	.0253	.0269	.077
EB4	.012	.0247	.0252	.078	.006	.0255	.0252	.067	.007	.0253	.0269	.077
META	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.012	.0252	.0247	.075	.005	.0258	.0252	.066	.004	.0263	.0258	.068
UML	.013	.0247	.0243	.075	.005	.0255	.0251	.065	.003	.0253	.0246	.061
CML	−.023	.0215	.0215	.100	.011	.0255	.0249	.074	.065	.0158	.0155	.450
EB	−.017	.0226	.0223	.077	.010	.0255	.0249	.072	.041	.0211	.0216	.216
EB1	.012	.0248	.0244	.074	.005	.0255	.0251	.065	.004	.0256	.0249	.063
EB2	.009	.0246	.0263	.077	.005	.0255	.0252	.066	.009	.0249	.0304	.101
EB3	.010	.0245	.0257	.076	.005	.0255	.0252	.066	.008	.0249	.0288	.089
EB4	.010	.0245	.0259	.076	.005	.0255	.0252	.066	.008	.0247	.0294	.093
$\theta_k = -.5$ for all k	**Main effect of *E***				**Main effect of *G***				***G*×*E* Interaction**			
IPD	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.006	.0195	.0193	.041	.004	.0314	.0323	.106	.004	.0358	.0348	.123
UML	.006	.0194	.0193	.041	.003	.0313	.0321	.104	.002	.0352	.0339	.115
CML	.096	.0185	.0185	.960	−.049	.0304	.0312	.340	−.331	.0223	.0224	11.034
EB	.010	.0195	.0194	.048	−.010	.0317	.0325	.116	.003	.0360	.0350	.123
EB1	.006	.0194	.0193	.041	.003	.0313	.0321	.104	.002	.0352	.0339	.115
EB2	.006	.0194	.0193	.041	.003	.0313	.0321	.104	.002	.0352	.0339	.115
EB3	.006	.0194	.0193	.041	.003	.0313	.0321	.104	.002	.0352	.0339	.115
EB4	.006	.0194	.0193	.041	.003	.0313	.0321	.104	.002	.0352	.0339	.115
META	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.005	.0195	.0193	.040	.001	.0315	.0321	.103	.001	.0360	.0345	.119
UML	.005	.0195	.0192	.040	.001	.0313	.0319	.102	.001	.0353	.0338	.114
CML	.095	.0186	.0185	.943	−.048	.0305	.0310	.330	−.329	.0223	.0224	10.858
EB	.034	.0197	.0197	.153	−.040	.0310	.0315	.258	−.033	.0377	.0366	.241
EB1	.005	.0195	.0192	.040	.001	.0313	.0319	.102	.000	.0353	.0338	.114
EB2	.005	.0195	.0192	.040	.001	.0313	.0319	.102	.000	.0353	.0338	.114
EB3	.005	.0195	.0192	.040	.001	.0313	.0319	.102	.000	.0352	.0338	.114
EB4	.005	.0195	.0192	.040	.001	.0313	.0319	.102	.000	.0352	.0338	.114

Note. In the meta-analysis setting, we use the inverse variance-covariance weighted approach to obtain the standard logistic, unconstrained, constrained, and EB results.

CML estimator and the inverse variance-covariance weighted UML estimator. For example, in our simulation runs, the EB estimator fell in between the UML and CML estimators in only 473 of the 1000 Monte Carlo runs under *G-E* independence. Further simulation (not reported) considered normal distributions with location parameters further from 0, eg, −.5 and .5. In these settings, we found the MSEB estimators to perform nearly identical to the UML estimator with respect to MSE in both the IPD and meta-analysis setting.

These simulation results indicate the data-adaptive feature of MSEB estimators across varying simulation scenarios. Although they do not outperform other options considered in some of the simulation scenarios, they do offer protection against bias and

TABLE 3 Bias (BIAS), estimated standard errors (SE1), empirical standard errors (SE2), and 100 × MSE (MSE) of $\hat{\gamma}_E, \hat{\gamma}_G$, and $\hat{\gamma}_{GE}$ resulting from standard logistic regression (LOG), unconstrained maximum likelihood (UML), constrained maximum likelihood (CML), empirical Bayes (EB), and our proposed multistudy empirical Bayes estimators in both individual patient data (IPD) and summary statistic meta-analysis (META) simulation settings when *G-E* independence is violated

$\theta_k \overset{iid}{\sim}$ N(.2, .1^2)	Main effect of *E*				Main effect of *G*				*G*×*E* Interaction			
IPD	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.015	.0269	.0268	.094	.009	.0253	.0247	.069	.006	.0254	.0253	.067
UML	.017	.0264	.0265	.099	.008	.0251	.0244	.066	.003	.0245	.0241	.059
CML	−.074	.0226	.0293	.629	.020	.0250	.0245	.102	.137	.0156	.0272	1.949
EB	.007	.0275	.0273	.080	.018	.0251	.0245	.092	.011	.0260	.0260	.079
EB1	.017	.0264	.0265	.098	.008	.0251	.0244	.066	.004	.0246	.0241	.060
EB2	.017	.0264	.0265	.098	.008	.0251	.0244	.066	.004	.0246	.0242	.060
EB3	.017	.0264	.0265	.098	.008	.0251	.0244	.066	.004	.0245	.0242	.060
EB4	.017	.0264	.0265	.098	.008	.0251	.0244	.066	.004	.0245	.0242	.060
META	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.014	.0270	.0266	.091	.007	.0253	.0246	.066	.004	.0255	.0251	.064
UML	.016	.0264	.0264	.094	.007	.0251	.0243	.064	.002	.0246	.0241	.059
CML	−.072	.0226	.0290	.609	.020	.0250	.0245	.102	.136	.0157	.0269	1.928
EB	−.023	.0255	.0270	.124	.019	.0250	.0244	.095	.038	.0236	.0275	.218
EB1	.015	.0264	.0264	.093	.007	.0251	.0243	.064	.003	.0246	.0241	.059
EB2	.015	.0265	.0264	.093	.007	.0251	.0243	.064	.003	.0246	.0242	.059
EB3	.015	.0264	.0264	.093	.007	.0251	.0243	.064	.003	.0245	.0242	.059
EB4	.015	.0264	.0264	.093	.007	.0251	.0243	.064	.003	.0245	.0242	.059
$\theta_k \overset{iid}{\sim}$ **Unif(−.2, .2)**	**Main effect of *E***				**Main effect of *G***				***G*×*E* Interaction**			
IPD	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.011	.0236	.0234	.066	.007	.0263	.0258	.071	.006	.0269	.0262	.072
UML	.012	.0233	.0233	.068	.006	.0261	.0255	.068	.004	.0260	.0255	.066
CML	.011	.0207	.0248	.074	.006	.0261	.0255	.069	.003	.0163	.0300	.091
EB	.011	.0219	.0227	.063	.006	.0261	.0255	.069	.005	.0235	.0239	.059
EB1	.012	.0232	.0231	.068	.006	.0261	.0254	.068	.003	.0256	.0250	.064
EB2	.012	.0232	.0231	.067	.006	.0261	.0254	.068	.003	.0255	.0250	.063
EB3	.012	.0231	.0231	.067	.006	.0261	.0254	.068	.003	.0254	.0249	.063
EB4	.012	.0231	.0231	.067	.006	.0261	.0254	.068	.003	.0254	.0249	.063
META	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**	**BIAS**	**SE1**	**SE2**	**MSE**
LOG	.010	.0236	.0232	.063	.005	.0264	.0257	.068	.004	.0270	.0260	.069
UML	.011	.0233	.0231	.065	.005	.0262	.0254	.067	.003	.0261	.0254	.065
CML	.011	.0208	.0246	.073	.007	.0261	.0255	.070	.005	.0164	.0297	.090
EB	.012	.0216	.0237	.071	.005	.0261	.0254	.067	.002	.0222	.0276	.077
EB1	.011	.0232	.0230	.065	.005	.0262	.0254	.067	.003	.0256	.0248	.062
EB2	.011	.0232	.0229	.066	.006	.0262	.0254	.068	.003	.0252	.0247	.062
EB3	.012	.0231	.0229	.066	.005	.0262	.0254	.068	.002	.0250	.0247	.061
EB4	.012	.0231	.0229	.066	.005	.0262	.0254	.068	.002	.0250	.0247	.061

Note. In the meta-analysis setting, we use the inverse variance-covariance weighted approach to obtain the standard logistic, unconstrained, constrained and EB results.

inflated MSE across all simulation scenarios investigated should *G*×*E* independence be incorrectly assumed. Further simulation studies with individual study sample sizes n_k (100-300 subjects), number of individual studies *K* (2 and 5), minor allele frequency MAF (5% and 10%), marginal disease prevalence (10% and 20%), and case-control ratios (1:2 and 1:4) are presented in the Supporting Information (Tables S1-S16). The results of these additional simulation studies are consistent with the findings noted in our simulation setup described above. More specifically, adjustments in individual study sample sizes, number of individual studies, MAF, and case-control ratios resulted in different (from our main simulation setup) magnitudes of bias, standard error, and MSE, but relative performance of the methods did not notably change. Similar findings resulted when the marginal disease prevalence was increased to 10%, but when increased to 20%, CML was generally biased and inefficient; however, our MSEB estimators still offered excellent protection against bias and loss of efficiency.

4 | ANALYSIS OF TYPE 2 DIABETES DATA

A number of SNPs in the fat mass associated *FTO* gene (region 16q12.2) have previously been found to be associated with type 2 diabetes (T2D) and BMI.[18-20] Body mass index with a strong genetic heritability also has environmental contribution, so it is likely that the assumption of *G-E* independence is violated between SNPs on the *FTO* gene and BMI. In our data analysis, we apply our proposed methods to estimate the interaction effects between SNPs (rs11642841, rs6499640, and rs1121980) in the *FTO* gene and environmental factors (age and BMI) on T2D using case-control studies that are a part of FIN-D2D 2007 (D2D2007), The DIAbetes GENetic Study (DIAGEN), the Finland-United States Investigation of NIDDM Genetics Stage 2 (FUSION S2), The Nord-Trøndelag Health Study 2 (HUNT), the METabolic Syndrome In Men Study (METSIM), and the Tromsø Study (TROMSO).[18] There are a total of $N = 9616$ individuals, composed of 4418 cases and 5198 controls. Sample sizes within each study varied from 1058 to 2215, and the case to control ratio within each study varied from .37 to 1.75 with an overall case to control ratio of .85. Descriptive summary statistics from the 6 studies are shown in Table 4. Individuals in the case group were significantly older (62.2 vs 59.0, P value < 0.001), had significantly higher BMI (30.1 vs 26.4, P value < 0.001), and had a significantly lower percentage of females (34% vs 45%, P value < 0.001) than individuals in the control group. The minor allele frequency of SNPs rs11642841, rs6499640, and rs1121980 across the 6 studies range from .40 to .45, .37 to .42,

TABLE 4 Summary statistics of age, body mass index (BMI), sex, and MAF (SNPs rs11642841, rs6499640, and rs1121980) stratified by cohort and T2D disease status in 6 case-control data sets sampled from D2D2007 (1), DIAGEN (2), FUSION S2 (3), HUNT (4), METSIM (5), and TROMSO (6)

							rs11642841			rs6499640			rs1121980		
		Age		BMI		Prop.		Trend[a]			Trend[b]			Trend[c]	
Study	*N*	Mean	SD	Mean	SD	female	MAF	Age	*p*	MAF	BMI	*p*	MAF	BMI	*p*
1	1698	59.5	8.3	27.3	5.0	.57	.41	−.16	5.9×10^{-1}	.42	−.39	3.0×10^{-2}	.40	.38	3.0×10^{-2}
2	1058	61.0	14.1	28.1	5.5	.57	.42	.07	9.1×10^{-1}	.40	.08	7.3×10^{-1}	.47	.19	4.2×10^{-1}
3	2215	59.3	8.2	28.7	5.1	.42	.40	−.02	9.3×10^{-1}	.42	−.02	8.8×10^{-1}	.41	.42	1.0×10^{-2}
4	1330	67.1	13.1	28.0	4.4	.48	.45	−.81	1.1×10^{-1}	.37	−.16	3.6×10^{-1}	.47	.30	7.0×10^{-2}
5	1899	58.0	6.9	28.6	5.0	.00	.43	−.19	3.9×10^{-1}	.41	−.21	2.0×10^{-1}	.44	.56	6.8×10^{-4}
6	1416	59.9	12.5	27.6	4.7	.50	.44	.26	5.8×10^{-1}	.38	.01	9.7×10^{-1}	.49	.33	6.0×10^{-2}
Total	9616	60.5	10.6	28.1	5.0	.40	.43	−.12	4.3×10^{-1}	.40	−.13	7.0×10^{-2}	.44	.38	1.4×10^{-7}
1	458	63.4	7.5	30.5	5.5	.41	.43	.47	3.5×10^{-1}	.39	−.02	9.5×10^{-1}	.40	.76	4.0×10^{-2}
2	434	65.8	11.7	30.1	6.2	.50	.44	.07	9.3×10^{-1}	.40	.26	5.0×10^{-1}	.48	.19	6.3×10^{-1}
3	1033	59.7	8.7	30.9	5.4	.44	.42	−.48	1.9×10^{-1}	.42	.51	3.0×10^{-2}	.43	.24	3.1×10^{-1}
4	577	68.9	11.4	29.2	4.6	.48	.47	−1.51	2.0×10^{-2}	.38	−.09	7.4×10^{-1}	.50	.22	4.0×10^{-1}
5	1209	60.5	6.6	30.2	5.2	.00	.43	−.16	5.5×10^{-1}	.40	−.18	3.9×10^{-1}	.45	.64	2.5×10^{-3}
6	707	59.9	12.5	29.2	4.9	.50	.45	.59	3.7×10^{-1}	.37	.04	8.9×10^{-1}	.49	.55	3.0×10^{-2}
Cases	4418	62.2	10.1	30.1	5.3	.34	.44	−.17	3.8×10^{-1}	.40	.09	4.2×10^{-1}	.46	.45	3.8×10^{-5}
1	1240	58.1	8.2	26.1	4.3	.63	.41	−.37	2.6×10^{-1}	.43	−.38	3.0×10^{-2}	.40	.24	1.7×10^{-1}
2	624	57.6	14.6	26.7	4.4	.61	.41	−.28	7.4×10^{-1}	.39	−.15	5.4×10^{-1}	.46	.02	9.2×10^{-1}
3	1182	59.0	7.6	26.9	3.9	.40	.39	.32	3.2×10^{-1}	.43	−.37	2.0×10^{-2}	.39	.23	1.7×10^{-1}
4	753	65.8	14.2	27.1	4.0	.48	.44	−.54	4.6×10^{-1}	.37	−.21	3.1×10^{-1}	.45	.15	4.7×10^{-1}
5	690	53.7	5.0	25.9	3.1	.00	.41	−.36	1.9×10^{-1}	.43	.01	9.3×10^{-1}	.44	.20	2.5×10^{-1}
6	709	59.9	12.5	25.9	3.8	.50	.44	−.05	9.4×10^{-1}	.38	−.04	8.4×10^{-1}	.48	.03	8.6×10^{-1}
Controls	5198	59.0	10.9	26.4	4.0	.45	.41	−.19	3.6×10^{-1}	.41	−.23	3.3×10^{-3}	.43	.17	3.0×10^{-2}

Note. Age values are measured in years, and BMI values are measured in kg/m^2. SNP effect estimates and P values are reported for the following trend tests.

[a] Age is regressed on SNP rs11642841 (0, 1, or 2) controlling for BMI and sex.

[b] BMI is regressed on SNP rs6499640 (0, 1, or 2) controlling for age and sex.

[c] BMI is regressed on SNP rs1121980 (0, 1, or 2) controlling for age and sex.

and .40 to .49, respectively, and the coefficients of linkage disequilibrium are $r^2 = .24$ (rs11642841 and rs6499640), $r^2 = .48$ (rs11642841 and rs1121980), and $r^2 = .09$ (rs6499640 and rs1121980).

The 6 different studies were treated as independent contributors to the IPD/meta-analysis. We applied our proposed MSEB estimators using several variants of the model specified in Section 2.1 for the following (observed) 3 scenarios: (1) weak *G-E* (age and rs11642841) association (trend test *P* value in Table 4 among controls: 3.6×10^{-1}) supporting *G-E* independence; (2) strong *G-E* (BMI and rs6499640) association (trend test *P* value in Table 4 among controls: 3.3×10^{-3}); and (3) modest *G-E* (BMI and rs1121980) association (trend test *P* value in Table 4 among controls: 3.0×10^{-2}), which are reflected in the (conditional) *G-E* association in the control group across the 6 studies (Table 4). The outcome variable D in each scenario indicates the presence/absence of T2D, and the variable G (coded 0,1,2) represents the particular SNP specified in each scenario. In scenario (1), the exposure variable E denotes age and we used the stratification variable $\boldsymbol{S} = (S_1, S_2)$ where S_1 is BMI and S_2 is gender. In scenarios (2) and (3), the exposure variable E denotes BMI and we used the stratification variable $\boldsymbol{S} = (S_1, S_2)$ where S_1 is age and S_2 is gender. In particular, the standard logistic regression model was specified as

$$\log\left\{\frac{P(D_{ki} = 1|E_{ki}, G_{ki}, \boldsymbol{S}_{ki})}{P(D_i = 0|E_i, G_i, \boldsymbol{S}_{ki})}\right\} = \beta_{0k} + \beta_1 E_{ki} + \beta_2 G_{ki} + \beta_3 G_{ki} E_{ki} + \beta_4 S_{ki1} + \beta_5 S_{i2} \tag{10}$$

in the IPD joint-analysis and specified as

$$\log\left\{\frac{P(D_{ki} = 1|E_{ki}, G_{ki}, \boldsymbol{S}_{ki})}{P(D_i = 0|E_i, G_i, \boldsymbol{S}_{ki})}\right\} = \beta_{0k} + \beta_{1k} E_{ki} + \beta_{2k} G_{ki} + \beta_{3k} G_{ki} E_{ki} + \beta_{4k} S_{ki1} + \beta_{5k} S_{i2} \tag{11}$$

TABLE 5 Meta-analysis results of *G-E* interactions (SNP1 × age, SNP2 × body mass index (BMI), and SNP3 × BMI) for the 6 case-control data sets sampled from D2D2007, DIAGEN, FUSION S2, HUNT, METSIM, and TROMSO controlling for BMI, age, and sex resulting from a standard logistic regression models

		Univariate			Multivariate		
$G \times E$	Model	Estimate	SE	*P* value	Estimate	SE	*P* value
SNP1	LOG	−0.0009	0.0034	8.0×10^{-1}	−0.0029	0.0033	3.8×10^{-1}
× age	UML	−0.0009	0.0034	8.0×10^{-1}	−0.0023	0.0030	4.4×10^{-1}
	CML	−0.0011	0.0030	7.1×10^{-1}	−0.0030	0.0023	1.8×10^{-1}
	EB	−0.0011	0.0032	7.4×10^{-1}	−0.0028	0.0028	3.1×10^{-1}
	EB1				−0.0028	0.0026	2.7×10^{-1}
	EB2				−0.0030	0.0023	1.8×10^{-1}
	EB3				−0.0030	0.0023	1.9×10^{-1}
	EB4				−0.0030	0.0023	1.8×10^{-1}
SNP2	LOG	0.0241	0.0084	4.1×10^{-3}	0.0209	0.0083	1.1×10^{-2}
× BMI	UML	0.0241	0.0084	4.1×10^{-3}	0.0182	0.0067	6.5×10^{-3}
	CML	0.0174	0.0068	1.1×10^{-2}	0.0024	0.0041	5.6×10^{-1}
	EB	0.0179	0.0076	1.9×10^{-2}	0.0050	0.0050	3.2×10^{-1}
	EB1				0.0178	0.0068	8.9×10^{-3}
	EB2				0.0167	0.0070	1.7×10^{-2}
	EB3				0.0174	0.0069	1.1×10^{-2}
	EB4				0.0174	0.0069	1.2×10^{-2}
SNP3	LOG	0.0007	0.0082	9.3×10^{-1}	0.0014	0.0080	8.6×10^{-1}
× BMI	UML	0.0007	0.0082	9.3×10^{-1}	0.0035	0.0066	5.9×10^{-1}
	CML	0.0046	0.0067	5.0×10^{-1}	0.0152	0.0041	1.8×10^{-4}
	EB	0.0026	0.0071	7.1×10^{-1}	0.0065	0.0062	3.0×10^{-1}
	EB1				0.0041	0.0068	5.5×10^{-1}
	EB2				0.0152	0.0041	1.8×10^{-4}
	EB3				0.0152	0.0041	1.8×10^{-4}
	EB4				0.0152	0.0041	1.8×10^{-4}

Note. SNPs are abbreviated as SNP1 = rs11642841, SNP2 = rs6499640 and SNP3 = rs1121980. Estimates shown are inverse-variance weighted averages across all studies.

in the meta-analysis where D denotes T2D disease status, G (coded 0,1,2) corresponds to rs11642841 (SNP1), rs6499640 (SNP2), or rs1121980 (SNP3), E corresponds to age (BMI) when G corresponds to SNP1 (SNP2 or SNP3), S_1 corresponds to BMI (age) when E corresponds to age (BMI), S_2 indicates male sex, i indexes subjects, and k indexes studies. For the UML, CML, EB, and our proposed MSEB estimators, we further model the minor allele frequency using a logistic regression model with covariates E and $\boldsymbol{S}$ under HWE similar to Section 2.1.

4.1 | Meta-analysis

Table 5 displays the $G{\times}E$ interaction effect estimates resulting from a standard inverse variance weighted univariate meta-analysis and our proposed MSEB estimators in the meta-analysis setting. As expected, an interaction effect was not detected in scenarios (1) and (3) and was detected in scenario (2). In our data results, our MSEB estimators tend to have smaller standard error than the standard logistic regression approach and UML approach. In addition, we point out that our proposed estimators EB2, EB3, and EB4 reduce to the CML estimates in the case that the uncertainty parameter τ^2 is estimated to be 0 as was the case in scenarios (1) and (3).

To evaluate the G-E independence assumption, we estimate the association parameter of G (rs6499640) with BMI among controls across the 6 case-control data sets sampled from D2D2007 (1), DIAGEN (2), FUSION S2 (3), HUNT (4), METSIM (5), and TROMSO (6) adjusting for age and sex using a standard multivariate regression model. In Figure 1, we report these estimated association parameters −0.38 (−.72, −.04), −0.15 (−.65, .34), −0.37 (−.68, −.06), −0.21 (−.62, .20), 0.01 (−.32, .35), and −0.04 (−.46, .38) for the 6 individual studies, respectively. While the figure may suggest evidence against G-E independence, we draw attention to the differences in the confidence intervals with respect to the point estimates and widths which reflects varying uncertainty in the G-E independence assumption.

4.2 | Individual patient data

In Table 6, we display the results for scenarios (1) to (3) using our proposed MSEB estimators in the following variants of the model specified in Section 2.1: (1) $\theta_k = \theta$ for all k and $\boldsymbol{\eta}_k$ is allowed to vary across each study; (2) both θ_k and $\boldsymbol{\eta}_k$ are allowed to vary across studies; (3) $\theta_k = \theta$ and $\boldsymbol{\eta}_k = \boldsymbol{\eta}$ for all k; and (4) θ_k is allowed to vary across studies while $\boldsymbol{\eta}_k = \boldsymbol{\eta}$ for all k. The estimates and standard errors are similar in each scenario; however, there is generally precision gain in comparison with the standard logistic regression and comparable results to the MSEB estimators in the meta-analysis setting (Table 4). In particular, we make the following two key observations: (1) In the estimation of the age × rs11642841 (weak G-E association) interaction effect, all the methods provide very close estimates and improved precision over the standard logistic or unconditional approach,

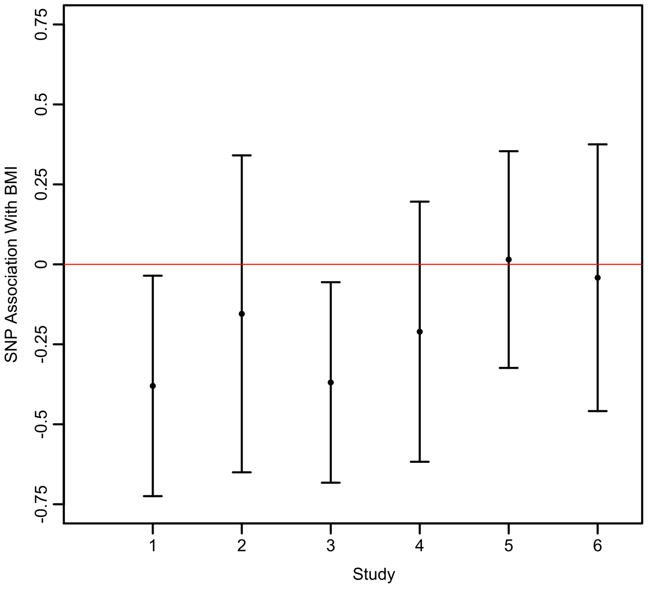

FIGURE 1 Estimated association parameter of G (rs6499640) with body mass index (BMI) among controls across the 6 case-control data sets sampled from D2D2007 (1), DIAGEN (2), FUSION S2 (3), HUNT (4), METSIM (5), and TROMSO (6) adjusting for age and sex using a standard multivariate regression model [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 6 Individual patient data results of *G-E* interactions (SNP1 × age, SNP2 × body mass index (BMI), and SNP3 × BMI) for the 6 case-control data sets sampled from D2D2007, DIAGEN, FUSION S2, HUNT, METSIM, and TROMSO controlling for BMI, age, sex, and study resulting from our proposed methods

		$\theta_k = \theta, \eta_k$			θ_k, η_k			$\theta_k = \theta, \eta_k = \eta$			$\theta_k, \eta_k = \eta$		
$G \times E$	**Model**	**Est.**	**SE**	p	**Est.**	**SE**	p	**Est.**	**SE**	p	**Est.**	**SE**	p
SNP1	LOG	−0.0009	0.0031	7.7×10^{-1}	−0.00090	0.00310	7.7×10^{-1}	−0.00090	0.00310	7.7×10^{-1}	−0.0009	0.0031	7.7×10^{-1}
× age	UML	−0.0013	0.0028	6.4×10^{-1}	−0.00140	0.00290	6.3×10^{-1}	−0.00110	0.00280	7.0×10^{-1}	−0.0011	0.0029	7.0×10^{-1}
	CML	−0.0025	0.0022	2.7×10^{-1}	−0.00250	0.00220	2.7×10^{-1}	−0.00230	0.00220	3.1×10^{-1}	−0.0023	0.0022	3.1×10^{-1}
	EB	−0.0022	0.0025	3.9×10^{-1}	−0.00220	0.00250	3.9×10^{-1}	−0.00210	0.00250	4.0×10^{-1}	−0.0021	0.0025	4.0×10^{-1}
	EB1	−0.0023	0.0025	3.5×10^{-1}	−0.00160	0.00290	5.7×10^{-1}	−0.00210	0.00250	4.1×10^{-1}	−0.0011	0.0029	7.0×10^{-1}
	EB2	−0.0025	0.0022	2.7×10^{-1}	−0.00250	0.00220	2.7×10^{-1}	−0.00230	0.00220	3.1×10^{-1}	−0.0023	0.0022	3.1×10^{-1}
	EB3	−0.0025	0.0022	2.7×10^{-1}	−0.00150	0.00290	6.1×10^{-1}	−0.00230	0.00220	3.1×10^{-1}	−0.0011	0.0029	7.0×10^{-1}
	EB4	−0.0025	0.0022	2.7×10^{-1}	−0.00250	0.00220	2.7×10^{-1}	−0.00230	0.00220	3.1×10^{-1}	−0.0023	0.0022	3.1×10^{-1}
SNP2	LOG	0.0231	0.0082	4.8×10^{-3}	0.02310	0.00820	4.8×10^{-3}	0.02310	0.00820	4.8×10^{-3}	0.0231	0.0082	4.8×10^{-3}
× BMI	UML	0.0188	0.0066	4.3×10^{-3}	0.01880	0.00670	5.1×10^{-3}	0.01910	0.00660	3.6×10^{-3}	0.0189	0.0067	4.8×10^{-3}
	CML	0.0029	0.0041	4.9×10^{-1}	0.00290	0.00410	4.9×10^{-1}	0.00270	0.00410	5.0×10^{-1}	0.0027	0.0041	5.0×10^{-1}
	EB	0.0203	0.0088	2.1×10^{-2}	0.02030	0.00880	2.1×10^{-2}	0.02030	0.00880	2.1×10^{-2}	0.0203	0.0088	2.1×10^{-2}
	EB1	0.0166	0.0070	1.7×10^{-2}	0.01840	0.00680	6.9×10^{-3}	0.01700	0.00690	1.4×10^{-2}	0.0189	0.0067	4.7×10^{-3}
	EB2	0.0164	0.0070	1.9×10^{-2}	0.01820	0.00690	7.9×10^{-3}	0.01680	0.00700	1.6×10^{-2}	0.0189	0.0067	4.7×10^{-3}
	EB3	0.0163	0.0070	1.9×10^{-2}	0.01830	0.00670	6.6×10^{-3}	0.01670	0.00700	1.7×10^{-2}	0.0189	0.0067	4.7×10^{-3}
	EB4	0.0164	0.0070	1.9×10^{-2}	0.01820	0.00670	6.8×10^{-3}	0.01680	0.00700	1.6×10^{-2}	0.0189	0.0067	4.7×10^{-3}
SNP3	LOG	0.0026	0.0080	7.5×10^{-1}	0.00259	0.00799	7.5×10^{-1}	0.00259	0.00799	7.5×10^{-1}	0.0026	0.0080	7.5×10^{-1}
× BMI	UML	0.0055	0.0065	3.9×10^{-1}	0.00484	0.00660	4.6×10^{-1}	0.00556	0.00646	3.9×10^{-1}	0.0049	0.0066	4.6×10^{-1}
	CML	0.0160	0.0040	8.0×10^{-5}	0.01598	0.00404	7.6×10^{-5}	0.01562	0.00403	1.1×10^{-4}	0.0156	0.0040	1.1×10^{-4}
	EB	0.0061	0.0087	4.8×10^{-1}	0.00610	0.00873	4.8×10^{-1}	0.00615	0.00872	4.8×10^{-1}	0.0061	0.0087	4.8×10^{-1}
	EB1	0.0082	0.0070	2.4×10^{-1}	0.00540	0.00678	4.3×10^{-1}	0.00831	0.00695	2.3×10^{-1}	0.0049	0.0066	4.6×10^{-1}
	EB2	0.0088	0.0069	2.0×10^{-1}	0.00599	0.00693	3.9×10^{-1}	0.00892	0.00690	2.0×10^{-1}	0.0049	0.0066	4.6×10^{-1}
	EB3	0.0090	0.0069	1.9×10^{-1}	0.00536	0.00677	4.3×10^{-1}	0.00923	0.00685	1.8×10^{-1}	0.0049	0.0066	1.9×10^{-1}
	EB4	0.0088	0.0069	2.0×10^{-1}	0.00598	0.00677	3.8×10^{-1}	0.00892	0.00685	1.9×10^{-1}	0.0049	0.0066	4.6×10^{-1}

Note. SNPs are abbreviated as SNP1 = rs11642841, SNP2 = rs6499640, and SNP3 = rs1121980.

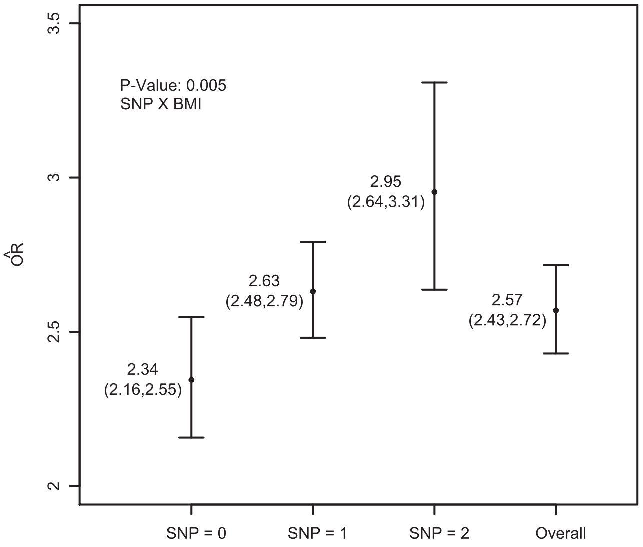

FIGURE 2 Estimated disease (T2D) odds ratios among subjects with genotype measured for the SNP rs6499640 (coded 0, 1, or 2) located on the *FTO* gene associated with a 5-unit increase (interquartile range) in exposure (body mass index (BMI)) adjusting for age, sex, and study cohort resulting from individual patient data joint analysis using standard logistic regression. Estimated marginal odds ratio for BMI without adjusting for genotype is also provided and labeled "Overall"

and (2) in the estimation of the BMI × rs1121980 and BMI × rs6499640 interactions effects, the constrained model estimates differed substantially from the estimates resulting from the unconstrained model. As expected, our MSEB estimates tend towards the constrained model estimates when the *G-E* association is weak and tends towards the unconstrained estimate when the *G-E* association is strong.

In Figure 2, we present stratified odds ratios associated with a 5-unit change (1 interquartile change) in BMI across the 3 genotype subgroups after adjusting for age, sex, and study cohort. The marginal OR is 2.57, 95% CI (2.43, 2.72). When stratified by genotype status the homozygous major allele $G = 0$ group had OR of 2.34, 95% CI (2.16, 2.55) while the homozygous minor allele subgroup had an OR of 2.95 (2.64, 3.31). The P value for testing interaction was 0.005. This finding needs follow up and replication in future studies. In Li et al[12], we noticed a significant BMI × rs1121980 interaction effect with high-density lipoproteins cholesterol level as outcome.

5 | DISCUSSION

In this paper, we extended the single-study EB type shrinkage estimators proposed by Mukherjee and Chatterjee (2008) to a meta-analysis setting that adjusts for uncertainty in the assumption of *G-E* independence. We used the retrospective likelihood framework to derive an adaptive combination of estimators obtained under the constrained model (assuming *G-E* independence) and unconstrained model (without any assumptions of independence) with weights determined by measures of *G-E* association derived from multiple studies. Our simulation studies indicate that these newly proposed MSEB estimators have smaller MSE than the standard alternative of using constrained, unconstrained, or EB estimators in the meta-analysis summary statistic setting when the *G-E* independence assumption is moderately violated. As previously noted, a contributor to these results is that the standard inverse variance-covariance weighted EB estimates do not necessarily lie between the standard inverse variance-covariance weighted UML and CML estimates (weight contamination). We also note that in other simulation settings the CML and EB can sometimes offer better performance than the newly proposed MSEB estimators. However, in terms of average performance across different scenarios of uncertainty regarding the *G-E* independence assumption, the newly proposed MSEB estimators offer protection in terms of bias without sacrificing much efficiency.

In the face of uncertainty, when historic data nor biology has established *G-E* independence, our recommendation is that the EB estimator be used in the IPD setting and our proposed MSEB estimators be used in the meta-analysis setting when IPD are not available. Specifically, we point to our MSEB estimator EB1 as the estimator of choice (in the meta-analysis setting) for the following various reasons. (1) EB1 is easy to implement and does not require an iterative process in the estimation of the covariance matrix $\boldsymbol{A}$ (required in EB3 and EB4); (2) EB1 does not have the inherent problem (with probability 1) of yielding estimates identical to CML as does EB2 to EB4 (when $\hat{\tau}^2$ is estimated to be 0), which deteriorates their ability to protect against bias and loss of efficiency when the assumption of *G*-*E* independence is false; (3) EB1 has smaller MSE relative to standard logistic regression (LOG) and the UML estimator when *G*-*E* independence holds; (4) EB1 offers protection against inflated bias and MSE when *G*-*E* independence does not hold; (5) the derived variance approximation formula (Supporting Information) for EB1 is simpler, easier to implement, and more stable than the derived variance approximation formulas for EB2 to EB4. While our simulation studies support the use of our variance approximations, it is important to note that the variance approximations for EB2, EB3, and EB4 depend on $(\hat{\theta}^{\mathrm{T}}\hat{\theta})^{-2}$, which can become very large if θ_k is estimated too close to 0 for all k. This can make these approximations unstable. A caveat of our proposed method is the additional modeling of the conditional mean of G given E. Various forms of model misspecification in the profile likelihood approach of Chatterjee and Carroll[3] have been considered in literature, eg, Tchetgen Tchetgen and Kraft.[21,22] Misspecification of either $P(D|G,E)$ or $P(G|E)$ can lead to bias. Model misspecification and measurement error are ubiquitous in the $G\times E$ literature and also remain a concern in our approach.

We applied our methods to 6 different case-control data sets sampled from D2D2007, DIAGEN, FUSION S2, HUNT, METSIM, and TROMSO. Our MSEB estimators reported sensible results relative to other considered estimates (eg, standard logistic regression, UML, and CML) and suggest that there is evidence to support SNP × BMI effects (rs6499640 and rs1121980) on T2D. We provide R codes for the simulation study at https://github.com/jpestes/mseb/blob/master/sim.txt.

Typically, IPD are not available to researchers for systematic review. Summary data provide a nice alternative avenue for analysis; however, it may be challenging to obtain the full covariance matrix from the researchers responsible for each individual study because only the diagonal elements are typically published. Nevertheless, increasing communication and advancements in technology among researchers can alleviate these issues.

ACKNOWLEDGEMENTS

The research of BM was supported by the National Science Foundation grant DMS 1406712 and the National Institute of Health grant ES 20811. The FUSION study was supported by DK062370. We thank the D2D2007, DIAGEN, FUSION S2, HUNT, METSIM, and TROMSO investigators for providing access to their data.

REFERENCES

1. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994;13(2):153-162.
2. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med*. 1997;16(15):1731-1743.
3. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005;92(2):399-418.
4. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*. 2008;64(3):685-694.
5. Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genet Epidemiol*. 2008;32(7):615-626.
6. Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. *Am J Epidemiol*. 2009;169(4):497-504.
7. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol*. 2009;169(2):219-226.
8. de Bakker P, Ferreira M, Jia X, Neale B, Raychaudhuri S, Voight B. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Mol Genet*. 2008;17:R122-R128.
9. Zeggini E, Ioannidis J. Meta-analysis in genome-wide association studies. *Pharmacogenomics*. 2010;97:321-332.
10. Lin DY, Zeng D. On the relative efficiency of using summary statistics versus individual level data in meta-analysis. *Biometrika*. 2010;10:191-201.
11. Lin DY, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet Epidemiol*. 2010;34(1):60-66.
12. Li S, Mukherjee B, Taylor JMG, et al. The role of environmental heterogeneity in meta-analysis of gene-environment interactions with quantitative traits. *Genet Epidemiol*. 2014;38(5):416-429.
13. Bhattacharjee S, Chatterjee N, Han S, Wheeler W. Cgen: An R package for analysis of case-control studies in genetic epidemiology. R package version 2.2.0; 2012.
14. Chen Y-H, Chatterjee N, Carroll RJ. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *J Am Stat Assoc*. 2009;104(485):220-233.
15. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. New York, New York : Springer Verlag; 1985.
16. Morris CN. Parametric empirical Bayes inference: theory and applications. *J Am Stat Assoc*. 1983;78(381):47-55.

17. Morris CN. Parametric empirical Bayes confidence intervals. *Sci Inference, Data Anal, Robustness*. 1983:25-50.
18. Scott L, Mohlke K, Bonnycastle L, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Sci*. 2007;316(5829):1341-1345.
19. Zeggini E, Weedon M, Lindgren C, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Sci*. 2007;316(5829):1336-1341.
20. Zeggini E, Scott L, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*. 2008;40(5):638-645.
21. Tchetgen Tchetgen EJ, Kraft P. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. *Epidemiol*. 2011;22(2):257-261.
22. Tchetgen Tchetgen EJ. Robust discovery of genetic associations incorporating gene-environment interaction and independence. *Epidemiol*. 2011;22(2):262-272.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Estes JP, Rice JD, Li S, Stringham HM, Boehnke M, Mukherjee B. Meta-analysis of gene-environment interaction exploiting gene-environment independence across multiple case-control studies. *Statistics in Medicine*. 2017;36:3895–3909. https://doi.org/10.1002/sim.7398