JUSTC

Article

http://justc.ustc.edu.cn

Received: January 08, 2022; Accepted: March 30, 2022

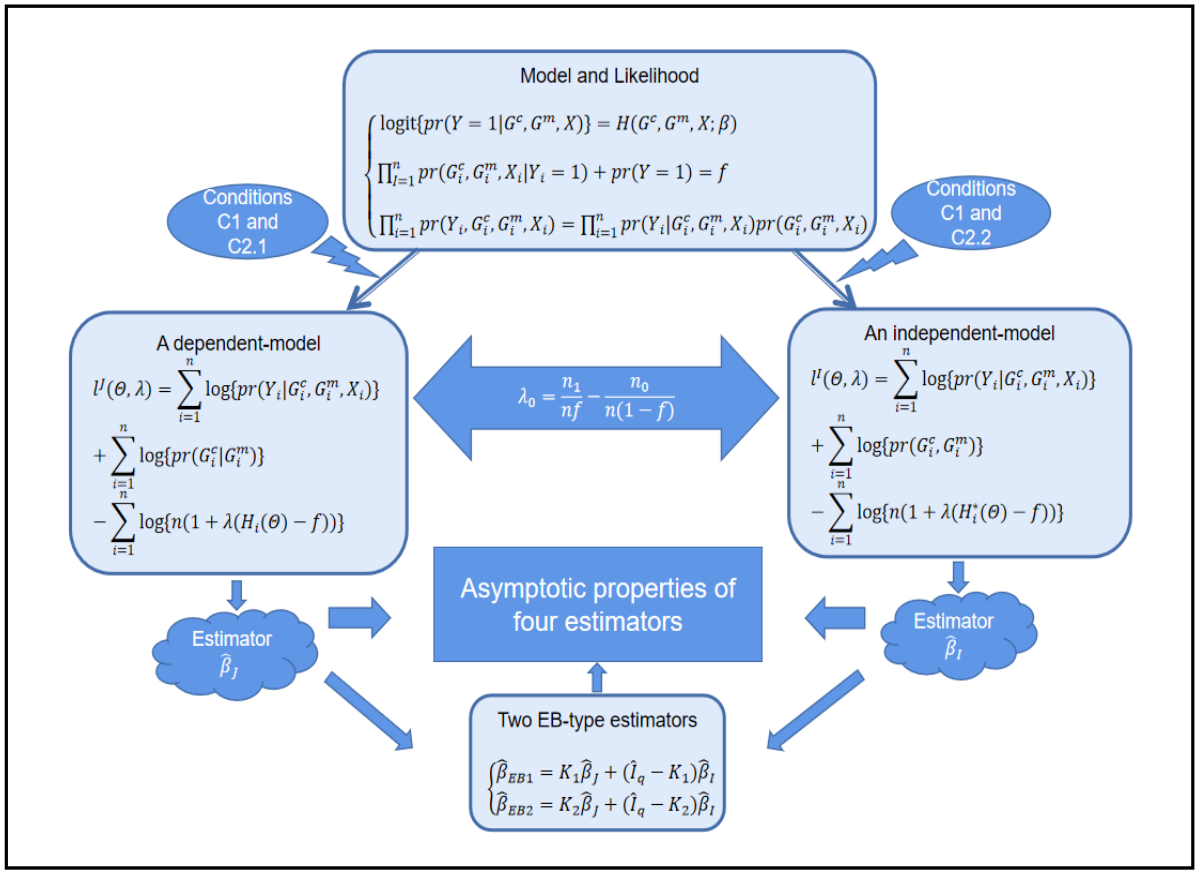
An empirical Bayes method for genetic association analysis using case-control mother-child pair data

Yanan Zhao, Weiqi Yang, and Hong Zhang ✉

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Hong Zhang, E-mail: zhangh@ustc.edu.cn

Graphical abstract



Based on a dependent-model estimator and an independent-model estimator, we obtain two Bayes-type estimators that balance robustness and efficiency.

Public summary

- Retrospective likelihood method is employed to improve statistical efficiency by fully utilizing available information.
- An efficient estimator and a robust estimator are combined to construct two novel empirical Bayes-type estimators using empirical the Bayes method.
- We establish the asymptotic properties of the proposed estimators.

Citation: Zhao Y N, Yang W Q, Zhang H. An empirical Bayes method for genetic association analysis using case-control mother-child pair data. *JUSTC*, **2022**, 52(5): 3. DOI: 10.52396/JUSTC-2022-0007

JUSTC

Article

http://justc.ustc.edu.cn

Received: January 08, 2022; Accepted: March 30, 2022

An empirical Bayes method for genetic association analysis using case-control mother-child pair data

Yanan Zhao, Weiqi Yang, and Hong Zhang ✉

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Hong Zhang, E-mail: zhangh@ustc.edu.cn

Cite This: *JUSTC*, **2022**, 52(5): 3 (13pp)

Read Online

SI Supporting Information

Abstract: Case-control mother-child pair data are often used to investigate the effects of maternal and child genetic variants and environmental risk factors on obstetric and early life phenotypes. Retrospective likelihood can fully utilize available information such as Mendelian inheritance and conditional independence between maternal environmental risk factors (covariates) and children's genotype given maternal genotype, thus effectively improving statistical inference. Such a method is robust to some extent if no relationship assumption is imposed between the maternal genotype and covariates. Statistical efficiency can be considerably improved by assuming independence between maternal genotype and covariates, but false-positive findings would be inflated if the independence assumption was violated. In this study, two empirical Bayes (EB) estimators are derived by appropriately weighting the above retrospective-likelihood-based estimators, which intuitively balance the statistical efficiency and robustness. The asymptotic normality of the two EB estimators is established, which can be used to construct confidence intervals and association tests of genetic effects and gene-environment interactions. Simulations and real-data analyses are conducted to demonstrate the performance of our new method.

Keywords: mother-child pair design; genetic association analysis; retrospective likelihood; profile likelihood; empirical Bayes

CLC number: O212.1 **Document code:** A

2020 Mathematics Subject Classification: 62P10

1 Introduction

It is well known that most obstetric and early-life diseases or phenotypes have a multifactorial etiology involving genetic factors, environmental exposures, and interactions between them[1–4]. For example, low maternal body mass index and genes of the mother and child were found to be associated with the risk of preterm delivery[5, 6]. Moreover, maternal COL24A1 variants were shown to have significant genome-wide interactions with maternal pre-pregnancy overweight/obesity on preterm birth risk using genetic data of 1733 African-American women from the Boston Birth Cohort[7]. Identifying potential genetic risk factors is important for understanding their biological effects and for developing public health strategies for prevention. A popular design for the integrative investigation of the effects affecting obstetrical and early life phenotypes is to collect risk factor information from children and their mothers[1, 8, 9]. In this study, the outcome of interest was a dichotomous phenotype of the mother or child, and cases and controls were ascertained based on this phenotype. This is the so-called mother-child case-control design[10, 11].

The standard prospective logistic regression method[12] is adopted widely in case-control studies. However, this method is not efficient because it does not utilize information such as Mendelian inheritance between parental and child genotypes. Alternative methods have been developed to assess the genetic effects on both children and mothers. Shi et al.[13] and Chen et al.[8] proposed the log-linear method and retrospective method for fitting logistic regression models, respectively, and improved the statistical inference efficiency by incorporating constraints on the genotype distribution, such as Mendelian inheritance, random mating, and Hardy-Weinberg equilibrium (HWE) in the population of interest. Chen et al.[10] extended the retrospective likelihood method to allow assessment of environmental effects and gene-environment interactions through a semiparametric maximum likelihood estimation (MLE) method, and improved statistical inference efficiency by incorporating an additional constraint that children's genotype is conditionally independent of maternal environmental risk factors given maternal genotype.

Using the retrospective likelihood method to analyze case-control mother-child pair data can result in a noticeable efficiency gain in estimating odds ratio (OR) parameters under the model assumptions of HWE and maternal gene-environment independence in the population of interest; however, it would produce serious bias if the independence assumption or HWE is violated. Chen et al.[10] alleviated the bias concern by allowing the conditional distribution of environmental risk factors for a given maternal genotype to be nonparametric, but this strategy is not sufficiently efficient. Alternatively, Zhang et al.[14] proposed a modified profile likelihood estima-

DOI: 10.52396/JUSTC-2022-0007

JUSTC, **2022**, 52(5): 3

tion method by relating the maternal genotype to environmental risk factors through a novel "double-additive" logistic regression (daLOG) model, which is computationally robust but not sufficiently robust because estimation bias may be introduced when the daLOG model is seriously violated.

Two novel estimators are proposed in this study: the weighted averages of an efficient estimator and a robust estimator. The second estimator is robust in that it is nearly unbiased, and the corresponding significance test has Type I error controlled around the nominal level. Both estimators are closely related to the modified profile MLE of Zhang et al.[14]. The efficient one imposes the assumptions of HWE and maternal gene-environment independence (referred to as "independent-model" estimator hereafter), which is biased if at least one of the assumptions is violated. In contrast, the robust one does not require the two assumptions, so it is relatively robust but might not be efficient (refer to as "dependent-model" estimator hereafter). Our weighting method was motivated by an empirical Bayes-type (EB-type) shrinkage estimator for estimating the effects of haplotypes and haplotype-environmental interactions with standard case-control data[15]. Here, such shrinkage technique is extended to assess OR parameters associated with maternal genotype and environmental risk factors using case-control mother-child genotype data. Briefly, the robust "dependent-model" estimator is shrunk toward the more efficient but possibly biased "independent-model" estimator. The magnitude of "shrinkage" is data-dependent, which tends to be large if and only if HWE and maternal gene-environmental independence are satisfied.

The remainder of this paper is organized as follows. In Section 2, two MLEs are derived based on dependent and independent models. Subsequently, two EB-type estimators are constructed based on the two MLEs, and their asymptotic properties are established. In Sections 3 and 4, the desired finite-sample performance of the proposed EB-type estimators is demonstrated through extensive simulations and real-data applications. Concluding remarks are provided in Section 5.

2 Materials and methods

2.1 Model, likelihood, and assumptions

Let Y denote the dichotomous disease status of either mother (e.g., pre-eclampsia) or child (e.g., low birth weight), that is, $Y = 0$ for a control pair and $Y = 1$ for a case pair, (G^m, G^c) denote the genotypes of a mother-child pair at a target genetic locus, and X denotes a vector of p maternal environmental risk factors. Y and (G^c, G^m, X) are related using the following logistic penetrance model:

$$\text{logit}\{\text{pr}(Y = 1|G^c, G^m, X)\} = H(G^c, G^m, X; \boldsymbol{\beta}) \tag{1}$$

where $\text{logit}(t) = \log(t/(1-t))$ is the logit function, $H(\cdot)$ is an arbitrarily specified function, and $\boldsymbol{\beta}$ is the q-vector of the regression parameters. G^c and G^m can be coded according to the mode of inheritance. The code set for a specific mode of inheritance is denoted as $\mathcal{M}$. For the additive mode of inheritance, the genotype is coded as the number of minor alleles ($\mathcal{M} = \{0, 1, 2\}$); for the dominant mode of inheritance, the genotype is coded as 1 if at least one of the minor alleles is present and 0 otherwise ($\mathcal{M} = \{0, 1\}$); for the recessive mode of inheritance, the genotype is coded as 1 if both alleles are minor alleles and 0 otherwise ($\mathcal{M} = \{0, 1\}$).

Suppose that (G^c, G^m, X) is collected from the n_0 control pairs ($Y = 0$) and n_1 case pairs ($Y = 1$). Let $n = n_0 + n_1$. The retrospective likelihood for case-control mother-child data is

$$\prod_{i=1}^{n} \text{pr}(G_i^c, G_i^m, X_i|Y_i).$$

Assume that the phenotype prevalence $\text{pr}(Y = 1)$ is known *a priori* to be f or that it can be estimated using extra data. According to the Bayesian theorem, under constraint

$$\text{pr}(Y = 1) = f \tag{2}$$

maximizing the likelihood function above is equivalent to maximizing:

$$\prod_{i=1}^{n} \text{pr}(Y_i, G_i^c, G_i^m, X_i) = \prod_{i=1}^{n} \text{pr}(Y_i|G_i^c, G_i^m, X_i)\text{pr}(G_i^c, G_i^m, X_i) \tag{3}$$

In this study, all the statistical inference procedures were based on the likelihood function (3). Under Mendelian inheritance and random mating, the joint probability $\text{pr}(G_i^c, G_i^m, X_i)$ in (3) can be further decomposed according to the two distributional assumptions on (G^c, G^m, X) (C1 and C2.1/C2.2).

C1. The child genotype is independent of maternal environmental risk factors conditional on the maternal genotype, i.e., $\text{pr}(G_i^c|G_i^m, X_i) = \text{pr}(G_i^c|G_i^m)$.

C2.1. The joint distribution (G_i^m, X_i) is completely unspecified (dependent model).

C2.2. G_i^m is independent of X_i (independent model).

Maternal environmental risk factors are generally maternal-related characteristics; therefore, they should be independent of the child genotype given the maternal genotype. Thus, C1 is a reasonable assumption in studies of gene-environment associations with obstetrical/maternal phenotypes using case control mother-child pair data. In contrast, maternal genotype and environmental factors can be correlated or uncorrelated, corresponding to Conditions C2.1 and C2.2, respectively.

In the following subsections, we modify the estimator proposed by Zhang et al.[14] to obtain two estimators of the regression parameter vector $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ under Conditions C1 and C2.1/C2.2. The first estimator based on C2.1 is robust but inefficient, and the second estimator is efficient but can be biased if C2.2 is violated. We then extend the EB theory of Chen et al.[15] to derive a shrinkage estimator that can adaptively balance bias and efficiency.

2.2 A dependent-model based estimator

Under Condition C1, the joint probability $\text{pr}(G^c, G^m, X)$ can be written as

$$\text{pr}(G^c, G^m, X) = \text{pr}(G^c|G^m)\text{pr}(G^m, X).$$

If Mendelian inheritance and random mating hold in the population of interest, the conditional distribution $\text{pr}(G^c|G^m)$ is a function of MAF θ and fixation index ρ (a parameter measuring deviation from HWE); refer to Ref. [14, Table 1] for the conditional distributions under various modes of inheritance.

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

These distributional constraints are incorporated into the likelihood function (3): Under C2.1, the empirical likelihood method[16] can be adopted by introducing the probability mass π_i for $\mathrm{pr}(G_i^m, X_i)$ that satisfies the constraint

$$\sum_{i=1}^{n}\pi_i = 1 \tag{4}$$

Denote $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$. The empirical log-likelihood function can be written as:

$$l_J(\boldsymbol{\Theta}, \boldsymbol{\pi}) = \sum_{i=1}^{n}[\log\{\mathrm{pr}(Y_i|G_i^c, G_i^m, X_i)\} + \log\{\mathrm{pr}(G_i^c|G_i^m)\}] + \sum_{i=1}^{n}\log\pi_i \tag{5}$$

As shown in Supporting Information Appendix S1, under constraints (2) and (4), the profile likelihood function of $\boldsymbol{\Theta} = (\theta, \rho, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ can be obtained using the Lagrange multiplier method, which takes the form:

$$l^J(\boldsymbol{\Theta}, \lambda) = \sum_{i=1}^{n}\log\{\mathrm{pr}(Y_i|G_i^c, G_i^m, X_i)\} + \sum_{i=1}^{n}\log\{\mathrm{pr}(G_i^c|G_i^m)\} - \sum_{i=1}^{n}\log\{n(1+\lambda(H_i(\boldsymbol{\Theta}) - f))\} \tag{6}$$

where

$$H_i(\boldsymbol{\Theta}) = \sum_{j\in\mathcal{M}}\mathrm{pr}(Y=1|G_i^c = j, G_i^m, X_i)\mathrm{pr}(G_i^c = j|G_i^m),$$

and λ satisfies the equation:

$$\sum_{i=1}^{n}\frac{H_i(\boldsymbol{\Theta}) - f}{1+\lambda(H_i(\boldsymbol{\Theta}) - f)} = 0 \tag{7}$$

which is exactly the "score" equation derived from profile likelihood l^J with respect to λ. Then, $\boldsymbol{\Theta}$ and λ can be estimated by solving Eq. (7) and $\partial l^J(\boldsymbol{\Theta}, \lambda)/\partial\boldsymbol{\Theta} = 0$,

As show in Ref. [17], it is computationally unstable to estimate $\boldsymbol{\Theta}$ and λ using the above method as the solution is a saddle point of l^J. The profile-likelihood function (6) can be modified to resolve this numerical problem. Specifically, λ in Eq. (6) can be replaced by its limiting value:[17]

$$\lambda_0 = \frac{n_1}{nf} - \frac{n_0}{n(1-f)} \tag{8}$$

The validity of this modification is as follows. Let $\boldsymbol{\Theta}_0$ be the true value of $\boldsymbol{\Theta}$. The "true" value λ_0 is the solution to the equations (refer to Supporting Information Appendix S2 for the detailed proof):

$$E\left[\frac{\partial l^J(\boldsymbol{\Theta}, \lambda)}{\partial\boldsymbol{\Theta}}\right]\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_0, \lambda=\lambda_0} = 0 \text{ and } E\left[\frac{\partial l^J(\boldsymbol{\Theta}, \lambda)}{\partial\lambda}\right]\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_0, \lambda=\lambda_0} = 0 \tag{9}$$

As a result, the so-called dependent-model estimator of $\boldsymbol{\Theta}$ can be obtained by solving

$$\frac{\partial l^J(\boldsymbol{\Theta}, \lambda_0)}{\partial\boldsymbol{\Theta}} = 0 \tag{10}$$

Asymptotic properties of the dependent-model estimator can be established under some regularity conditions[18]. First, with a probability tending to one as $n \to \infty$, there exists a solution $\hat{\boldsymbol{\Theta}}_J$ to Eq. (10), which is consistent with $\boldsymbol{\Theta}_0$. Second, $\hat{\boldsymbol{\Theta}}_J$ is asymptotically normally distributed (refer to Supporting Information Appendix S3 for a detailed derivation).

$$\sqrt{n}(\hat{\boldsymbol{\Theta}}_J - \boldsymbol{\Theta}_0) \xrightarrow{D} N\{\mathbf{0}, I_J^{-1}(\boldsymbol{\Theta}_0)\Sigma_J(\boldsymbol{\Theta}_0)I_J^{-1}(\boldsymbol{\Theta}_0)\} \tag{11}$$

where

$$I_J(\boldsymbol{\Theta}_0) = -\frac{1}{n}E\left[\frac{\partial^2 l^J(\boldsymbol{\Theta}, \lambda_0)}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^{\mathrm{T}}}\right]\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_0},$$
$$\Sigma_J(\boldsymbol{\Theta}_0) = \frac{1}{n}\mathrm{cov}\left(\frac{\partial l^J(\boldsymbol{\Theta}, \lambda_0)}{\partial\boldsymbol{\Theta}}\right)\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_0}.$$

Obviously, $I_J(\boldsymbol{\Theta}_0)$ can be consistently estimated by the empirical Fisher information matrix, that is,

$$\hat{I}_J(\hat{\boldsymbol{\Theta}}_J) := -\frac{1}{n}\frac{\partial^2 l^J(\boldsymbol{\Theta}, \lambda_0)}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^{\mathrm{T}}}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}_J} \tag{12}$$

Among n selected subjects, without loss of generality, assume that the first n_0 subjects are controls (i.e., $Y_i = 0$ for $i = 1, \cdots, n_0$) and the remaining $n_1 = n - n_0$ subjects are cases (i.e., $Y_i = 1$ for $i = n_0 + 1, \cdots, n = n_0 + n_1$). Then, $\Sigma_J(\boldsymbol{\Theta}_0)$ can be estimated consistently using

$$\hat{\Sigma}_J(\hat{\boldsymbol{\Theta}}_J) = \frac{n_0}{n(n_0-1)}\sum_{i=1}^{n_0}\left(\frac{\partial l_i^J(\boldsymbol{\Theta}, \lambda_0)}{\partial\boldsymbol{\Theta}} - \frac{1}{n_0}\sum_{j=1}^{n_0}\frac{\partial l_j^J(\boldsymbol{\Theta}, \lambda_0)}{\partial\boldsymbol{\Theta}}\right)^{\otimes 2}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}_J} + \frac{n_1}{n(n_1-1)}\sum_{i=n_0+1}^{n}\left(\frac{\partial l_i^J(\boldsymbol{\Theta}, \lambda_0)}{\partial\boldsymbol{\Theta}} - \frac{1}{n_1}\sum_{j=n_0+1}^{n}\frac{\partial l_j^J(\boldsymbol{\Theta}, \lambda_0)}{\partial\boldsymbol{\Theta}}\right)^{\otimes 2}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}_J} \tag{13}$$

where $a^{\otimes 2} = aa^{\mathrm{T}}$ for any vector a. Therefore, the limiting variance-covariance matrix of $\sqrt{n}\hat{\boldsymbol{\Theta}}_J$ can be consistently estimated using $\hat{I}_J^{-1}(\hat{\boldsymbol{\Theta}}_J)\hat{\Sigma}_J(\hat{\boldsymbol{\Theta}}_J)\hat{I}_J^{-1}(\hat{\boldsymbol{\Theta}}_J)$. Consequently, significance tests of the genetic effects can be constructed.

2.3 An independent-model based estimator

Now, we consider obtaining an independent-model estimator of $\boldsymbol{\Theta}$ under Conditions C1 and C2.2. In this case, the joint probability $\mathrm{pr}(G^c, G^m, X)$ can be written as $\mathrm{pr}(G^c, G^m)\mathrm{pr}(X)$. Under Mendelian inheritance and random mating, $\mathrm{pr}(G^c, G^m)$ is a function of θ or (θ, ρ) depending on whether HWE holds or not. The empirical likelihood method was adopted, as in the dependent model. Here, the distribution of X, instead of the joint distribution of (G^m, X), is allowed to be nonparametric. Let μ_i denote the probability mass for $\mathrm{pr}(X_i)$ that satisfies

$$\sum_{i=1}^{n}\mu_i = 1 \tag{14}$$

Denote $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$. Then, the corresponding empirical log-likelihood function is

$$l_I(\boldsymbol{\Theta}, \boldsymbol{\mu}) = \sum_{i=1}^{n}\log\{\mathrm{pr}(Y_i|G_i^c, G_i^m, X_i)\} + \sum_{i=1}^{n}\log\{\mathrm{pr}(G_i^c, G_i^m)\} + \sum_{i=1}^{n}\log\mu_i \tag{15}$$

Similar to the derivation of $l^J(\boldsymbol{\Theta}, \lambda)$, under constraints (2) and (14), the profile likelihood function corresponding to Eq. (15) takes the form (refer to Supporting Information Ap-

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

pendix S1 for a detailed derivation):

$$l^{\mathrm{I}}(\boldsymbol{\Theta},\lambda^*)=\sum_{i=1}^{n}\log\{\mathrm{pr}(Y_i|G_i^c,G_i^m,X_i)\}+\sum_{i=1}^{n}\log\{\mathrm{pr}(G_i^c,G_i^m)\}-\sum_{i=1}^{n}\log\{n(1+\lambda^*(H_i^*(\boldsymbol{\Theta})-f))\} \tag{16}$$

where

$$H_i^*(\boldsymbol{\Theta})=\sum_{j,k\in\mathcal{M}}\mathrm{pr}(Y=1|G_i^c=j,G_i^m=k,X_i)\mathrm{pr}(G_i^c=j,G_i^m=k)$$

and λ^* satisfies the equation:

$$\sum_{i=1}^{n}\frac{H_i^*(\boldsymbol{\Theta})-f}{1+\lambda^*(H_i^*(\boldsymbol{\Theta})-f)}=0.$$

The MLEs of $\boldsymbol{\Theta}$ and λ^* can then be obtained by solving the "score" equations

$$\frac{\partial l^{\mathrm{I}}(\boldsymbol{\Theta},\lambda^*)}{\partial\lambda^*}=0\quad\text{and}\quad\frac{\partial l^{\mathrm{I}}(\boldsymbol{\Theta},\lambda^*)}{\partial\boldsymbol{\Theta}}=0 \tag{17}$$

Again, λ^* can be replaced by λ_0 defined by Eq. (8). Similarly, λ_0 is the exact solution to the following equations:

$$E\left[\frac{\partial l^{\mathrm{I}}(\boldsymbol{\Theta},\lambda)}{\partial\boldsymbol{\Theta}}\right]\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_0,\lambda=\lambda_0}=0\text{ and }E\left[\frac{\partial l^{\mathrm{I}}(\boldsymbol{\Theta},\lambda)}{\partial\lambda}\right]\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_0,\lambda=\lambda_0}=0 \tag{18}$$

As a result, the so-called independent-model estimator of $\boldsymbol{\Theta}$ can be obtained by solving

$$\frac{\partial l^{\mathrm{I}}(\boldsymbol{\Theta},\lambda_0)}{\partial\boldsymbol{\Theta}}=0 \tag{19}$$

As Supporting Information Appendix S3 shows, the independent model estimator, denoted by $\hat{\boldsymbol{\Theta}}_{\mathrm{I}}$, is asymptotically normal under some regularity conditions:

$$\sqrt{n}(\hat{\boldsymbol{\Theta}}_{\mathrm{I}}-\boldsymbol{\Theta}_0)\xrightarrow{D}N\{\mathbf{0},I_{\mathrm{I}}^{-1}(\boldsymbol{\Theta}_0)\Sigma_{\mathrm{I}}(\boldsymbol{\Theta}_0)I_{\mathrm{I}}^{-1}(\boldsymbol{\Theta}_0)\} \tag{20}$$

where $I_{\mathrm{I}}(\boldsymbol{\Theta}_0)$ and $\Sigma_{\mathrm{I}}(\boldsymbol{\Theta}_0)$ are the analogs of $I_{\mathrm{J}}(\boldsymbol{\Theta}_0)$ and $\Sigma_{\mathrm{J}}(\boldsymbol{\Theta}_0)$, respectively. As in Eqs. (12) and (13), $I_{\mathrm{I}}(\boldsymbol{\Theta}_0)$ and $\Sigma_{\mathrm{I}}(\boldsymbol{\Theta}_0)$ can be estimated consistently.

2.4 Two empirical Bayes-type shrinkage estimators

Compared with the dependent-model estimator, the independent-model estimator is more efficient, but serious bias could be produced if environmental risk factors are correlated with maternal genotype. Chen et al.[15] were fully aware of the two possibly conflicting goals of improving efficiency and maintaining robustness when estimating the effects of haplotypes and haplotype-environmental interactions with standard case-control data and thus proposed an EB-type shrinkage estimator to balance efficiency and robustness in a data-adaptive fashion. Here, we adopt this "shrinkage" estimation technique to balance the estimation robustness and efficiency of the effects of maternal and child genetic variants and environmental risk factors in case-control mother-child studies.

Let $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{J}}$ denote all genetic and environmental related effect estimates in the independent-model and dependent-model estimators, respectively. Intuitively, a weighted estimator has the form

$$\hat{\boldsymbol{\beta}}_{\mathrm{W}}=K\hat{\boldsymbol{\beta}}_{\mathrm{J}}+(I_q-K)\hat{\boldsymbol{\beta}}_{\mathrm{I}}=\hat{\boldsymbol{\beta}}_{\mathrm{I}}+K(\hat{\boldsymbol{\beta}}_{\mathrm{J}}-\hat{\boldsymbol{\beta}}_{\mathrm{I}}) \tag{21}$$

where I_q denotes an identity matrix of size $q=\dim(\boldsymbol{\beta})$, K is a weight factor that favors $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ or $\hat{\boldsymbol{\beta}}_{\mathrm{J}}$ depending on the bias of independent-model estimator. Obviously, K should have the property that it favors the robust dependent-model estimator if the independent-model estimator is biased, and vice versa. In the following, we derive the optimal weight matrix K based on the EB theory.

Throughout this study, we assume that the dependent model described in Section 2.2 is true. We denote the true value of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_{\mathrm{J}}$. Let $\boldsymbol{\beta}_{\mathrm{I}}$ be the value of $\boldsymbol{\beta}$ that minimizes the Kullback-Leibler discrepancy between dependent and independent models. Then, $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ converges to $\boldsymbol{\beta}_{\mathrm{I}}$ under certain regularity conditions. The prior distribution of $\boldsymbol{\beta}_{\mathrm{J}}$ is assumed to be a multivariate normal distribution with expectation $\boldsymbol{\beta}_{\mathrm{I}}$ and variance-covariance matrix $\boldsymbol{A}$, where $\boldsymbol{A}$ is a $q\times q$ positive definite matrix. In view of the asymptotic normality of $\hat{\boldsymbol{\beta}}_{\mathrm{J}}$, conditional on the value of $\boldsymbol{\beta}_{\mathrm{J}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{J}}$ approximately follows a multivariate normal distribution with mean vector $\boldsymbol{\beta}_{\mathrm{J}}$ and variance-covariance matrix $\boldsymbol{V}$, where $\boldsymbol{V}$ is a $q\times q$ positive definite matrix. Consequently, the posterior expectation of $\boldsymbol{\beta}_{\mathrm{J}}$ is given by

$$\boldsymbol{A}(\boldsymbol{A}+\boldsymbol{V})^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{J}}+\boldsymbol{V}(\boldsymbol{A}+\boldsymbol{V})^{-1}\boldsymbol{\beta}_{\mathrm{I}} \tag{22}$$

where $\boldsymbol{\beta}_{\mathrm{I}}$ is estimated using $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$. The appropriate estimators of $\boldsymbol{A}$ and $\boldsymbol{V}$ are discussed as follows. Obviously, $\boldsymbol{\psi}:=\boldsymbol{\beta}_{\mathrm{J}}-\boldsymbol{\beta}_{\mathrm{I}}$ follows a multivariate normal distribution with a zero mean vector and variance-covariance matrix $\boldsymbol{A}$. The prior hyperparameter $\boldsymbol{A}$ can be conservatively estimated using $\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^{\mathrm{T}}$, where $\hat{\boldsymbol{\psi}}=\hat{\boldsymbol{\beta}}_{\mathrm{J}}-\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ [15, 19]. An estimator of $\boldsymbol{V}$, denoted by $\hat{\boldsymbol{V}}$, can be derived from the variance-covariance matrix of $\hat{\boldsymbol{\Theta}}_{\mathrm{J}}$ described in Section 2.2. The resulting EB-type shrinkage estimator can be expressed as

$$\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}=K_1\hat{\boldsymbol{\beta}}_{\mathrm{J}}+(I_q-K_1)\hat{\boldsymbol{\beta}}_{\mathrm{I}} \tag{23}$$

where the shrinkage factor, K_1, is

$$K_1=\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^{\mathrm{T}}(\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^{\mathrm{T}}+\hat{\boldsymbol{V}})^{-1}.$$

As in Ref. [15], another shrinkage factor K_2 alternative to K_1 is considered, which is a diagonal matrix with the ith diagonal element $k_i=\hat{\psi}_i^2/(\hat{v}_i+\hat{\psi}_i^2)$:

$$K_2=\mathrm{diag}(\hat{\psi}_1^2/(\hat{\psi}_1^2+\hat{v}_1),\cdots,\hat{\psi}_q^2/(\hat{\psi}_q^2+\hat{v}_q)),$$

where $\hat{v}_i$ and $\hat{\psi}_i$ are the ith diagonal elements of $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{\psi}}$, respectively. This results in an alternative EB estimator:

$$\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}=K_2\hat{\boldsymbol{\beta}}_{\mathrm{J}}+(I_q-K_2)\hat{\boldsymbol{\beta}}_{\mathrm{I}} \tag{24}$$

If the observed data support the independence of environmental risk factors and maternal genotype, then $\hat{\boldsymbol{\psi}}\approx\mathbf{0}_q$ such that both $\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}$ approximates $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$. If the observed data support the correlation between environmental risk factors and maternal genotype, then $\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^{\mathrm{T}}$ tends to be a non-zero q-dimensional matrix, which increases with the degree of deviation of the independent model; thus, $\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}$ would tend to $\hat{\boldsymbol{\beta}}_{\mathrm{J}}$.

In the following, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}$ is derived by applying the delta method. Note that they are functions of $(\hat{\boldsymbol{\beta}}_{\mathrm{I}},\hat{\boldsymbol{\beta}}_{\mathrm{J}})$. We first derive the asymptotic distribution of $(\hat{\boldsymbol{\beta}}_{\mathrm{I}},\hat{\boldsymbol{\beta}}_{\mathrm{J}})$. Based on the asymptotic expressions of $\hat{\boldsymbol{\Theta}}_{\mathrm{J}}$ and $\hat{\boldsymbol{\Theta}}_{\mathrm{I}}$ derived in Sections 2.2 and 2.3, we obtain the follow-

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

ing asymptotic normality:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1,\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_J)\xrightarrow{D}N(\mathbf{0},\Sigma_\beta),$$

where the variance-covariance matrix

$$\Sigma_\beta=\mathrm{cov}\begin{pmatrix}(\mathbf{0},I_q)I_1^{-1}(\boldsymbol{\Theta}_0)\dfrac{1}{\sqrt{n}}S_1^0(\boldsymbol{\Theta}_0)\\(\mathbf{0},I_q)I_J^{-1}(\boldsymbol{\Theta}_0)\dfrac{1}{\sqrt{n}}S_J^0(\boldsymbol{\Theta}_0)\end{pmatrix}$$

can be consistently estimated by

$$\hat{\Sigma}_\beta=\mathrm{cov}\begin{pmatrix}(\mathbf{0},I_q)\hat{I}_1^{-1}(\hat{\boldsymbol{\Theta}}_1)\dfrac{1}{\sqrt{n}}S_1^0(\hat{\boldsymbol{\Theta}}_1)\\(\mathbf{0},I_q)\hat{I}_J^{-1}(\hat{\boldsymbol{\Theta}}_J)\dfrac{1}{\sqrt{n}}S_J^0(\hat{\boldsymbol{\Theta}}_J)\end{pmatrix}.$$

Here, $\mathbf{0}$ is a $q\times 2$ matrix of zeros, and $I_J(\boldsymbol{\Theta}_0)$, $I_1(\boldsymbol{\Theta}_0)$, $\hat{I}_J(\hat{\boldsymbol{\Theta}}_J)$, and $\hat{I}_1(\hat{\boldsymbol{\Theta}}_1)$ are given in Sections 2.2 and 2.3. Furthermore, $S_1^0(\boldsymbol{\Theta})$ and $S_J^0(\boldsymbol{\Theta})$ are the partial derivative vectors of $l^1(\boldsymbol{\Theta},\lambda_0)$ and $l^J(\boldsymbol{\Theta},\lambda_0)$, respectively, with respect to $\boldsymbol{\Theta}$.

According to the delta method, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}$ can be consistently estimated by (See Supporting Information Appendix S4)

$$\hat{\Sigma}_{\mathrm{EB1}}=\frac{1}{n}\hat{G}_1\hat{\Sigma}_\beta\hat{G}_1^{\mathsf{T}}\ \text{ and }\ \hat{\Sigma}_{\mathrm{EB2}}=\frac{1}{n}\hat{G}_2\hat{\Sigma}_\beta\hat{G}_2^{\mathsf{T}},$$

respectively, where

$$\hat{G}_1=\left(\frac{1}{1+\hat{\boldsymbol{\psi}}^{\mathsf{T}}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\psi}}}I_q-\frac{2\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^{\mathsf{T}}\hat{\boldsymbol{V}}^{-1}}{(1+\hat{\boldsymbol{\psi}}^{\mathsf{T}}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\psi}})^2},\right.$$
$$\left.\frac{\hat{\boldsymbol{\psi}}^{\mathsf{T}}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\psi}}}{1+\hat{\boldsymbol{\psi}}^{\mathsf{T}}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\psi}}}I_q+\frac{2\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^{\mathsf{T}}\hat{\boldsymbol{V}}^{-1}}{(1+\hat{\boldsymbol{\psi}}^{\mathsf{T}}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\psi}})^2}\right),$$

$$\hat{G}_2=(D,I_q-D),$$

and

$$D=\mathrm{diag}\left(\frac{\hat{\nu}_1(\hat{\nu}_1-\hat{\psi}_1^2)}{(\hat{\nu}_1+\hat{\psi}_1^2)^2},\ldots,\frac{\hat{\nu}_q(\hat{\nu}_q-\hat{\psi}_q^2)}{(\hat{\nu}_q+\hat{\psi}_q^2)^2}\right).$$

Note that $\hat{G}_1$ and $\hat{G}_2$ are the partial derivative matrices of $\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}$ with respect to $(\hat{\boldsymbol{\beta}}_1^{\mathsf{T}},\hat{\boldsymbol{\beta}}_J^{\mathsf{T}})^{\mathsf{T}}$.

3 Simulation studies

The finite sample performance of the proposed method was evaluated through extensive simulations. First, we evaluated the estimation performance for the two EB-type estimators (23) and (24). The powers and Type I error rates of the corresponding significance tests were then examined.

3.1 Simulation setups

Several parameter combinations were considered, as described in the following subsections. For each parameter combination, genotype data (G^m,G^c) were generated for a population of 1×10^7 mother-child pairs, and two covariates were considered: a continuous covariate X_1 and a discrete covariate X_2, where X_1 was generated according to a linear model and could be either independent of or correlated with maternal genotype, and X_2 was sampled from the binomial distribution $B(1,0.5)$ and was independent of both maternal genotype and X_1. Given each vector (G^m,G^c,X_1,X_2), a binary disease outcome Y was generated according to the logistic regression model:

$$\begin{aligned}\mathrm{logit}\{\mathrm{pr}(Y=1)\}=&\beta_0+G^c\beta_c+G^m\beta_m+X_1\beta_{x_1}+X_2\beta_{x_2}+\\&G^cX_1\beta_{cx_1}+G^cX_2\beta_{cx_2}+G^mX_1\beta_{mx_1}+G^mX_2\beta_{mx_2}.\end{aligned}$$

For each parameter combination, 1000 datasets, each consisting of n_1 case pairs and n_0 control pairs, were sampled from the population.

In all simulations, both G^m and G^c were coded as minor allele counts, so that the additive mode of inheritance was adopted in the penetrance model (1). Note that $\mathrm{pr}(G^c|G^m)$ does not depend on the fixation index in this situation (refer to Ref. [14, Table 1]). Hereafter, let Dep represent the dependent-model-based method that does not depend on the HWE assumption. Let IndHWE and IndHWD represent the independent-model-based methods with and without the HWE assumption, respectively. Here, the HWD denotes Hardy-Weinberg disequilibrium ($\rho>0$). Let θ be fixed at 0.2, and let ρ be fixed at 0 for HWE or 0.1 HWD. The covariate X_1 was generated according to the following linear model:

$$X_1=\eta\times[G^m-E(G^m)]+e,$$

where e was independent of G^m, and follows a standard normal distribution. Note that η characterizes the correlation strength between X_1 and G^m, and a zero value of η indicates independence between X_1 and G^m and vice versa. In addition, the prevalence of Y was fixed at 0.03 and assumed to be known in the relevant methods.

The value of η was set to 0, log(1.5), or log(2.5) for various correlation strengths between X_1 and G^m. The odds ratio parameters were set to $\beta_c=\log(1.8)$, $\beta_m=\log(2)$, $\beta_{x_1}=\beta_{x_2}=-\log(1.2)$, $\beta_{cx_1}=\beta_{cx_2}=-\log(1.5)$ and $\beta_{mx_1}=\beta_{mx_2}=\log(1.5)$. We considered various sample-size combinations with equal or unequal numbers of case and control pairs. In the equal situation, the common number was set to 150, 300, 600, or 1000; in the unequal situation, the ratio of case pairs to control pairs was set to be 1 : 2, and the number of case pairs was set to be 100, 200, 400, or 700.

In significance tests, to illustrate the impact of the correlation between maternal genotype and environmental risk factors, the value of η was set to range between 0 and log(2.5), and the values of β_c, β_x, and β_{cx} were fixed as above, β_m was set as log(1.3), β_{mx_2} was set to log(1.2), and the value of β_{mx_1} ranged between $-\log(1.5)$ and log(1.5). The significance level was fixed at 5%, and the sample size was $n_1=n_0=150$, 300, or 1000.

3.2 Estimation results

Five considered methods, namely IndHWE, IndHWD, Dep, EB1, and EB2 were applied to the generated data. To compare the estimation efficiencies of two estimators, we introduced the so-called efficiency gain of estimator A against estimator B as

$$1-\frac{\mathrm{MSE(A)}}{\mathrm{MSE(B)}},$$

where MSE(A) and MSE(B) are the mean squared errors of Methods A and B, respectively. The estimation results are summarized in Tables 1–3 and the Supporting Information

DOI: 10.52396/JUSTC-2022-0007
JUSTC, 2022, 52(5): 3

Tables S1–S9. The efficiency gains with equal numbers of case and control pairs are presented in Supporting Information Tables S10–S15.

As shown in Tables 1 and S4, all five considered methods are essentially unbiased for all sample size combinations when both HWE and independence assumptions are satisfied. According to MSEs, IndHWE is more efficient than EB1 and EB2, and the latter two methods are more efficient than Dep. For example, when $n_1 = n_0 = 150$, the efficiency gains of EB1 and EB2 against Dep were approximately 0.95% – 24.29% and 2.16% – 37.98%, respectively. IndHWE appears to be much more efficient than IndHWD with efficiency gains ranging between 9.04% and 44.01% for $n_1 = n_0 = 150$, although IndHWD incorporates only one more parameter ρ compared with IndHWE (Supporting Information Table S10). When the independence assumption holds but HWE is violated, IndHWE could produce considerable biases on maternal genetic effects but not on the other effects in our simulation situation, so that IndHWD outperforms IndHWE in estimating maternal genetic effects (Supporting Information Tables S1, S7, and S13). However, the estimation results of the remaining four methods appeared to be unaffected by the violation of HWE.

When HWE holds, but the G^m-X_1 independence assumption is violated ($\eta = \log(1.5)$), both IndHWE and IndHWD could be seriously biased, especially for main effect of X_1 and the interaction effect between and G^m and X_1 (Tables 2 and S5). As expected, Dep, which does not require HWE or independence assumptions, is inherently unbiased. Correspondingly, the MSEs of the above two effects were generally much larger for IndHWE than for Dep. Furthermore, EB1 and EB2 can effectively balance the bias and efficiency. The efficiency gains of EB1 and EB2 against IndHWE are considerable for the above two effects. For example, when $n_1 = n_0 = 150$, the efficiency gains of EB1 and EB2 against IndHWE for the two effects are approximately 29.67% – 40.04% and 68.67% – 73.91%, respectively. For the other effects, EB1 and EB2 are less efficient than IndHWE but more efficient than Dep, with efficiency gains against Dep ranging between 0.22% and 10.56% for EB1 and 0.86% and 33.52% for EB2 when the sample size is 150 (Supporting Information Table S11). In addition, EB1 and EB2 are slightly more efficient than IndHWD in the simulation situations. When both HWE and the independence assumption are violated, both IndHWE and IndHWD are biased, and their performances are poorer than those of the other three methods (Supporting Information Tables S2, S8, and S14).

With a larger η of log(2.5), IndHWE and IndHWD are more biased toward the maternal effect, X_1 environmental effect, and G^m-X_1 interaction (Tables 3 and S6), in contrast with the results for $\eta = \log(1.5)$ (Tables 2 and S5). Consequently, the efficiency gains of EB1 and EB2 against IndHWE and IndHWD increased. For example, with a sample size of 150, the efficiency gains of EB1 and EB2 against IndHWE for the three effects increased to 23.00% – 26.30%, 64.75% – 72.99%, and 91.51% – 92.62%, respectively (Supporting Information Table S12). When HWE was violated, the efficiency gains of EB1, EB2, and Dep against IndHWE and IndHWD were even greater (Supporting Information Tables S3,

Table 1. MSE and bias (in parenthesis) of various estimators when case and control pairs are equal and both the independence assumption ($\eta = 0$) and HWE ($\rho = 0$) hold.

n_0 (n_1)	Method	β_c	β_m	β_{x1}	β_{x2}	β_{cx1}	β_{cx2}	β_{mx1}	β_{mx1}
	IndHWE	0.056(0.004)	0.051(0.006)	0.024(–0.002)	0.110(–0.021)	0.024(–0.007)	0.101(–0.008)	0.023(0.009)	0.100(0.014)
	IndHWD	0.099(0.006)	0.088(–0.038)	0.027(–0.001)	0.124(–0.082)	0.037(–0.017)	0.166(0.007)	0.031(0.017)	0.155(0.039)
150	Dep	0.067(0.008)	0.104(0.033)	0.032(–0.012)	0.133(–0.024)	0.025(–0.010)	0.103(–0.009)	0.053(0.027)	0.193(0.019)
	EB1	0.062(0.008)	0.082(0.024)	0.029(–0.010)	0.123(–0.024)	0.024(–0.009)	0.102(–0.009)	0.040(0.022)	0.152(0.018)
	EB2	0.057(0.006)	0.068(0.021)	0.026(–0.007)	0.113(–0.023)	0.024(–0.007)	0.101(–0.008)	0.033(0.019)	0.127(0.017)
	IndHWE	0.025(0.006)	0.025(0.004)	0.011(–0.002)	0.054(0.005)	0.011(–0.009)	0.047(–0.006)	0.011(0.005)	0.045(0.006)
	IndHWD	0.041(0.004)	0.044(–0.044)	0.012(0.003)	0.064(–0.061)	0.015(–0.013)	0.068(0.016)	0.014 (0.004)	0.063(0.033)
300	Dep	0.032(0.009)	0.048(0.015)	0.015(–0.006)	0.066(0.004)	0.012(–0.011)	0.048(–0.007)	0.025(0.012)	0.091(0.007)
	EB1	0.029(0.010)	0.038(0.010)	0.013(–0.005)	0.061(0.005)	0.012(–0.011)	0.048(–0.007)	0.019(0.011)	0.070(0.007)
	EB2	0.026(0.008)	0.031(0.008)	0.012(–0.004)	0.056(0.005)	0.011(–0.010)	0.047(–0.006)	0.016(0.010)	0.058(0.005)
	IndHWE	0.013(0.004)	0.013(0.008)	0.006(–0.000)	0.026(–0.002)	0.006(–0.005)	0.025(0.004)	0.006(0.001)	0.021(–0.003)
	IndHWD	0.019(–0.014)	0.020(–0.024)	0.007(0.004)	0.035(–0.065)	0.007(–0.005)	0.032(0.031)	0.006(–0.001)	0.030(0.022)
600	Dep	0.017(0.004)	0.027(0.012)	0.008(–0.004)	0.030(–0.005)	0.007(–0.006)	0.025(0.005)	0.012(0.008)	0.048(0.003)
	EB1	0.015(0.005)	0.021(0.010)	0.007(–0.003)	0.028(–0.004)	0.007(–0.006)	0.025(0.004)	0.010(0.006)	0.036(0.001)
	EB2	0.014(0.004)	0.017(0.010)	0.006(–0.002)	0.026(–0.002)	0.006(–0.005)	0.025(0.004)	0.008(0.005)	0.029(–0.001)
	IndHWE	0.008(0.004)	0.008(–0.001)	0.003(0.001)	0.016(–0.002)	0.004(–0.005)	0.015(0.004)	0.004(0.001)	0.014(0.003)
	IndHWD	0.012(–0.013)	0.012(–0.021)	0.004(0.004)	0.023(–0.056)	0.004(–0.004)	0.020(0.026)	0.004(–0.001)	0.019(0.025)
1000	Dep	0.010(0.003)	0.015(0.007)	0.004(–0.001)	0.018(–0.002)	0.004(–0.005)	0.015(0.004)	0.008(0.004)	0.028(0.000)
	EB1	0.009(0.003)	0.012(0.005)	0.004(–0.000)	0.017(–0.001)	0.004(–0.005)	0.015(0.004)	0.006(0.004)	0.022(0.001)
	EB2	0.008(0.003)	0.010(0.004)	0.004(0.000)	0.016(–0.002)	0.004(–0.005)	0.015(0.004)	0.005(0.003)	0.018(0.003)

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

Table 2. MSE and bias (in parenthesis) of various estimators when case and control pairs are equal and the independence assumption is violated ($\eta = \log(1.5)$) but HWE holds ($\rho = 0$).

n_0 (n_1)	Method	β_c	β_m	β_{x_1}	β_{x_2}	β_{cx_1}	β_{cx_2}	β_{mx_1}	β_{mx_2}
	IndHWE	0.055(–0.003)	0.063(–0.071)	0.051(–0.167)	0.099(0.006)	0.022(0.007)	0.114(–0.005)	0.190(0.408)	0.100(–0.001)
	IndHWD	0.096(0.005)	0.100(–0.100)	0.046(–0.128)	0.112(–0.064)	0.033(0.006)	0.167(0.021)	0.162(0.341)	0.141(0.031)
150	Dep	0.066(0.011)	0.110(0.018)	0.031(–0.011)	0.120(–0.008)	0.023(–0.009)	0.115(–0.001)	0.046(0.012)	0.191(0.017)
	EB1	0.064(0.011)	0.099(0.006)	0.031(–0.028)	0.115(–0.006)	0.023(–0.007)	0.115(–0.002)	0.050(0.057)	0.171(0.014)
	EB2	0.056(0.003)	0.076(–0.023)	0.036(–0.086)	0.101(0.002)	0.022(0.005)	0.114(–0.005)	0.060(0.097)	0.127(0.007)
	IndHWE	0.027(–0.004)	0.031(–0.078)	0.038(–0.164)	0.049(–0.006)	0.011(0.009)	0.049(–0.006)	0.176(0.406)	0.048(0.012)
	IndHWD	0.041(–0.018)	0.049(–0.110)	0.037(–0.146)	0.063(–0.082)	0.013(0.013)	0.069(0.024)	0.162(0.376)	0.069(0.052)
300	Dep	0.032(0.002)	0.053(0.001)	0.015(0.001)	0.061(–0.017)	0.011(–0.004)	0.049(–0.002)	0.022(0.009)	0.098(0.029)
	EB1	0.032(0.002)	0.049(–0.005)	0.015(–0.016)	0.059(–0.016)	0.011(–0.003)	0.049(–0.002)	0.024(0.037)	0.091(0.028)
	EB2	0.027(–0.002)	0.036(–0.035)	0.019(–0.061)	0.050(–0.009)	0.011(0.008)	0.049(–0.006)	0.028(0.058)	0.063(0.020)
	IndHWE	0.013(–0.006)	0.019(–0.080)	0.032(–0.162)	0.026(–0.007)	0.005(0.011)	0.025(–0.000)	0.170(0.406)	0.024(0.004)
	IndHWD	0.017(–0.030)	0.028(–0.104)	0.033(–0.161)	0.038(–0.081)	0.006(0.018)	0.031(0.028)	0.168(0.398)	0.035(0.045)
600	Dep	0.015(–0.002)	0.025(–0.001)	0.007(–0.002)	0.032(–0.015)	0.006(–0.001)	0.024(0.004)	0.010(0.006)	0.046(0.014)
	EB1	0.015(–0.002)	0.024(–0.005)	0.007(–0.009)	0.031(–0.015)	0.006(–0.000)	0.024(0.004)	0.011(0.021)	0.045(0.014)
	EB2	0.013(–0.004)	0.020(–0.038)	0.009(–0.038)	0.027(–0.010)	0.005(0.009)	0.025(0.000)	0.012(0.031)	0.031(0.009)
	IndHWE	0.007(–0.001)	0.015(–0.082)	0.029(–0.161)	0.016(0.002)	0.003(0.009)	0.016(–0.006)	0.168(0.405)	0.014(0.006)
	IndHWD	0.010(–0.027)	0.021(–0.100)	0.031(–0.164)	0.025(–0.066)	0.004(0.015)	0.020(0.025)	0.169(0.405)	0.021(0.405)
1000	Dep	0.009(0.002)	0.017(0.001)	0.004(–0.000)	0.020(–0.001)	0.003(–0.002)	0.016(–0.002)	0.006(0.004)	0.029(0.006)
	EB1	0.009(0.002)	0.016(–0.001)	0.004(–0.004)	0.020(–0.001)	0.003(–0.001)	0.016(–0.002)	0.007(0.013)	0.028(0.006)
	EB2	0.008(0.000)	0.014(–0.035)	0.005(–0.024)	0.017(0.001)	0.003(0.008)	0.016(–0.006)	0.007(0.019)	0.018(0.005)

Table 3. MSE and bias (in parenthesis) of various estimators when case and control pairs are equal and the independence assumption is violated ($\eta = \log(2.5)$) but HWE holds ($\rho = 0$).

n_0 (n_1)	Method	β_c	β_m	β_{x_1}	β_{x_2}	β_{cx_1}	β_{cx_2}	β_{mx_1}	β_{mx_2}
	IndHWE	0.061(–0.053)	0.161(–0.326)	0.104(–0.295)	0.097(0.024)	0.017(0.039)	0.123(–0.009)	0.539(0.720)	0.110(–0.013)
	IndHWD	0.100(–0.057)	0.176(–0.270)	0.104(–0.256)	0.107(–0.044)	0.024(0.039)	0.171(0.035)	0.471(0.624)	0.163(–0.003)
150	Dep	0.072(–0.003)	0.128(0.033)	0.028(–0.012)	0.121(0.013)	0.018(–0.009)	0.118(0.006)	0.037(0.029)	0.205(–0.009)
	EB1	0.071(–0.004)	0.124(0.021)	0.028(–0.021)	0.119(0.014)	0.018(–0.008)	0.118(0.005)	0.040(0.051)	0.200(–0.009)
	EB2	0.063(–0.036)	0.118(–0.104)	0.037(–0.084)	0.100(0.021)	0.017(0.027)	0.123(–0.008)	0.046(0.076)	0.136(–0.012)
	IndHWE	0.028(–0.052)	0.144(–0.341)	0.091(–0.287)	0.050(–0.287)	0.010(0.042)	0.053(–0.007)	0.504(0.703)	0.047(–0.003)
	IndHWD	0.049(–0.092)	0.144(–0.323)	0.107(–0.296)	0.063(–0.079)	0.013(0.060)	0.071(0.040)	0.500(0.685)	0.067(0.009)
300	Dep	0.031(–0.012)	0.059(0.009)	0.013(–0.002)	0.064(–0.018)	0.009(–0.004)	0.049(0.010)	0.017(0.011)	0.093(0.006)
	EB1	0.031(–0.012)	0.058(0.003)	0.013(–0.007)	0.064(–0.018)	0.009(–0.003)	0.049(0.010)	0.018(0.022)	0.091(0.006)
	EB2	0.028(–0.038)	0.067(–0.094)	0.016(–0.043)	0.052(–0.008)	0.009(0.029)	0.052(–0.006)	0.019(0.034)	0.060(0.001)
	IndHWE	0.016(–0.044)	0.130(–0.339)	0.087(–0.287)	0.023(0.011)	0.006(0.043)	0.029(–0.013)	0.494(0.699)	0.026(–0.008)
	IndHWD	0.029(–0.104)	0.131(–0.331)	0.109(–0.318)	0.031(–0.072)	0.009(0.062)	0.039(0.045)	0.519(0.712)	0.033(0.005)
600	Dep	0.016(–0.005)	0.031(0.006)	0.006(–0.003)	0.029(–0.001)	0.005(–0.001)	0.027(0.004)	0.009(0.007)	0.044(0.000)
	EB1	0.016(–0.005)	0.031(0.003)	0.006(–0.006)	0.029(–0.001)	0.005(–0.001)	0.027(0.004)	0.009(0.013)	0.043(–0.000)
	EB2	0.015(–0.030)	0.037(–0.062)	0.007(–0.025)	0.024(0.008)	0.005(0.027)	0.029(–0.011)	0.009(0.019)	0.031(–0.005)
	IndHWE	0.010(–0.044)	0.121(–0.336)	0.084(–0.286)	0.015(0.007)	0.004(0.043)	0.017(–0.004)	0.485(0.694)	0.015(–0.013)
	IndHWD	0.022(–0.109)	0.125(–0.339)	0.107(–0.322)	0.025(–0.083)	0.007(0.065)	0.024(0.054)	0.519(0.717)	0.019(0.005)
1000	Dep	0.011(–0.007)	0.018(0.007)	0.004(–0.001)	0.020(–0.008)	0.003(–0.001)	0.017(0.012)	0.005(0.002)	0.026(–0.001)
	EB1	0.011(–0.007)	0.018(0.005)	0.004(–0.003)	0.020(–0.008)	0.003(–0.000)	0.017(0.012)	0.005(0.005)	0.026(–0.001)
	EB2	0.010(–0.030)	0.021(–0.039)	0.004(–0.014)	0.016(0.003)	0.003(0.023)	0.017(–0.003)	0.005(0.009)	0.018(–0.007)

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

S9, and S15).

In summary, IndHWE is the preferred method when the assumptions of HWE and maternal gene-environment independence hold. However, when any one of the two assumptions is violated, serious estimation biases can be produced for maternal genetic effects, environmental effects, and maternal gene-environment interactions. Dep is the most robust with respect to the two assumptions, but it is much less efficient than IndHWE, and its performance is much poorer than IndHWE in terms of MSE when independence is satisfied. In contrast, EB1 and EB2 can balance estimation bias and efficiency by appropriately weighting IndHWE and Dep. In the small-sized situation, EB1 is generally more unbiased than EB2, although the former is slightly less efficient. Consequently, EB1 may be preferable to EB2 with a small sample size; however, EB2 may be preferable to EB1 with a moderate or large sample size.

The five considered methods require specification of the prevalence f. Therefore, it would be interesting to study the impact of misspecifying f on the estimation results. A sensitivity analysis was conducted for this purpose. The true parameters (including f) were the same as those of Table 1, but f was misspecified to be 0.01 or 0.1 for all of the five methods. The simulation results shown in Supporting Information Tables S16–S21 suggest that all methods were not affected by misspecifying the prevalence, except that the MSEs were slightly larger, with the prevalence being overspecified. This finding is consistent with those of Refs. [14, 20].

3.3 Significance test results

In this subsection, the significance test results (type-I error rates and powers at a significance level of 0.05) of the five considered methods are summarized for maternal gene-environment interactions in the simulation settings described in Section 3.1. Fig. 1 and Supporting Information Table S22 show type-I error rates for the significance tests of β_{mx_1} for various η ($0-\log(2.5)$), sample sizes ($n_1 = n_0 = 150$, 300, and 1000), and ρ (0 for HWE and 0.1 for HWD). As expected, type-I error rates of IndHWE and IndHWD increase dramatically with η, because both methods rely on the independence assumption. In contrast, type-I error rates of both EB1 and EB2 are only slightly inflated (type I error rates: $4.4\%-7.5\%$ and $3.9\%-9.9\%$ for EB1 and EB2, respectively, $n_1 = n_0 = 150$ and $\rho = 0$), controlled around the nominal level even for large η, demonstrating the robustness of EB1 and EB2. Not surprisingly, Dep has well-controlled type-I error rates ($4.4\%-6.2\%$, $n_1 = n_0 = 150$, and $\rho = 0$).

Figs. 2 and 3 show the powers for testing the G^m-X_1 interaction for various β_{mx_1} (from $-\log(1.5)$ to $\log(1.5)$), sample sizes ($n_1 = n_0 = 150$, 300, and 1 000), and ρ (0 for HWE and 0.1 for HWD). Because Type I error rates of IndHWE and IndHWD cannot be controlled when $\eta > 0$, powers are not reported for IndHWE and IndHWD in such a situation. As ex-

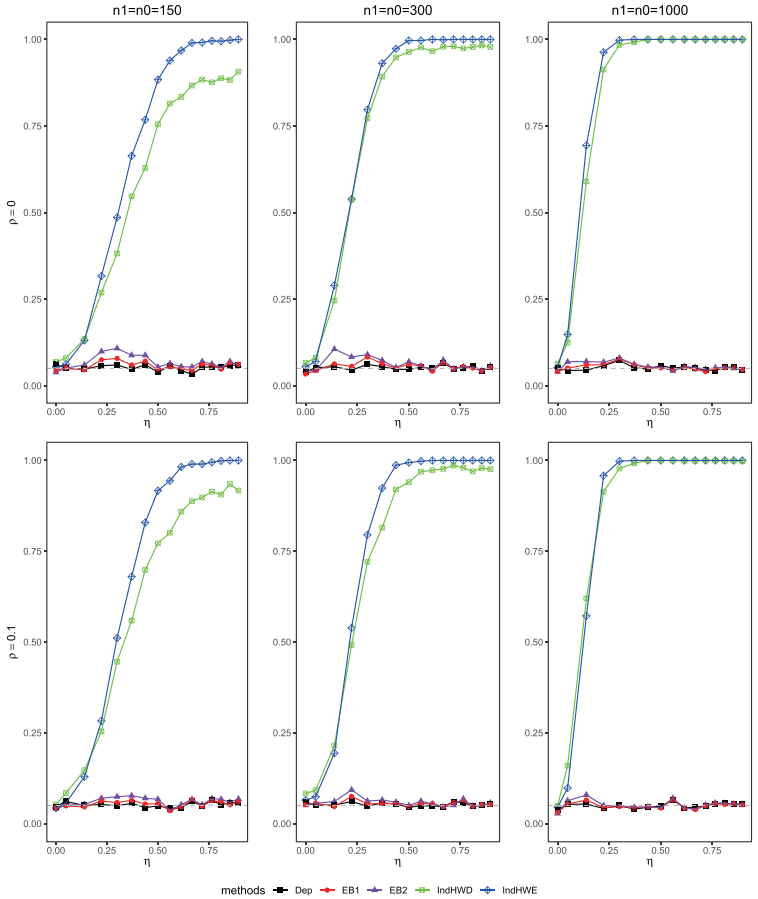


Fig. 1. Type-I error rates for the significance tests of β_{mx_1} with $\rho = 0$ (HWE) or $\rho = 0.1$ (HWD), various sample sizes ($n_1 = n_0 = 150$, $n_1 = n_0 = 300$, and $n_1 = n_0 = 1000$), and various η values (0 through $\log(2.5)$). The other parameters were fixed: $\beta_c = \log(1.8)$, $\beta_m = \log(1.3)$, $\beta_{x_1} = \beta_{x_2} = -\log(1.2)$, $\beta_{cx_1} = \beta_{cx_2} = -\log(1.5)$, and $\beta_{mx_2} = \log(1.2)$.

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

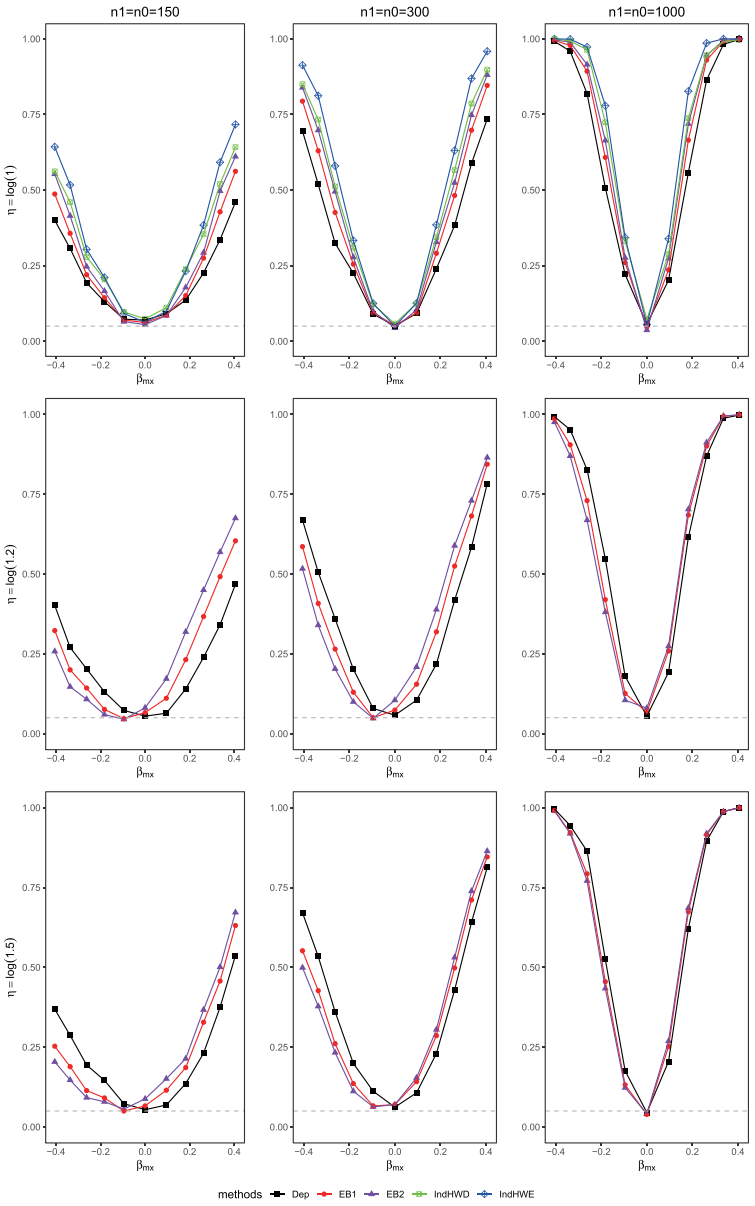


Fig. 2. Powers for the significance tests of maternal gene-environment interaction (β_{mx_1}) with HWE ($\rho = 0$), various sample sizes ($n_1 = n_0 = 150$, $n_1 = n_0 = 300$, and $n_1 = n_0 = 1000$), and various η values (0, log(1.2), and log(1.5)). The other parameters were fixed: $\beta_c = \log(1.8)$, $\beta_m = \log(1.3)$, $\beta_{x_1} = \beta_{x_2} = -\log(1.2)$, $\beta_{cx_1} = \beta_{cx_2} = -\log(1.5)$, and $\beta_{mx_2} = \log(1.2)$.

pected, IndHWE is the most powerful among the five methods when the independence assumption holds ($\eta = 0$), followed by IndHWD ranks the second. Evidently, EB1 and EB2 were more powerful than Dep in this situation. When the independence assumption does not hold, EB1 and EB2 can be either more or less powerful than Dep, depending on the sign of β_{mx_1}.

The diagrams of Type I errors and powers for the significance tests of the combined effect of maternal gene-environment interactions are shown in the Supporting Information Figures S1–S3, which have patterns similar to those of the G^m-X_1 interaction effect.

4 Application to the Danish National Birth Cohort study

Previous studies have shown that spontaneous preterm birth (SPTB) is the single largest cause of premature birth and is thought to be caused by a variety of factors, including infection or inflammation, stress, malnutrition, and genetic factors[5]. Using case-control mother-child pair data nested in the Danish National Birth Cohort (DNBC), we examined the association between SPTB and candidate SNPs by adjusting for maternal pre-pregnancy BMI (pp-BMI).

The DNBC is a cohort study that began in 1996. Extensive biomaterials and epidemiological data were collected from more than 10 0000 mothers and their children in Denmark. It has been well established that both fetal and maternal genotypes contribute to the risk of preterm birth, and genetic loci associated with SPTB have been identified through genetic association studies[21–24]. Furthermore, low maternal pp-BMI was believed to be associated with a high risk of SPTB[25]. In this study, 720 mothers gave birth prematurely (gestational

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

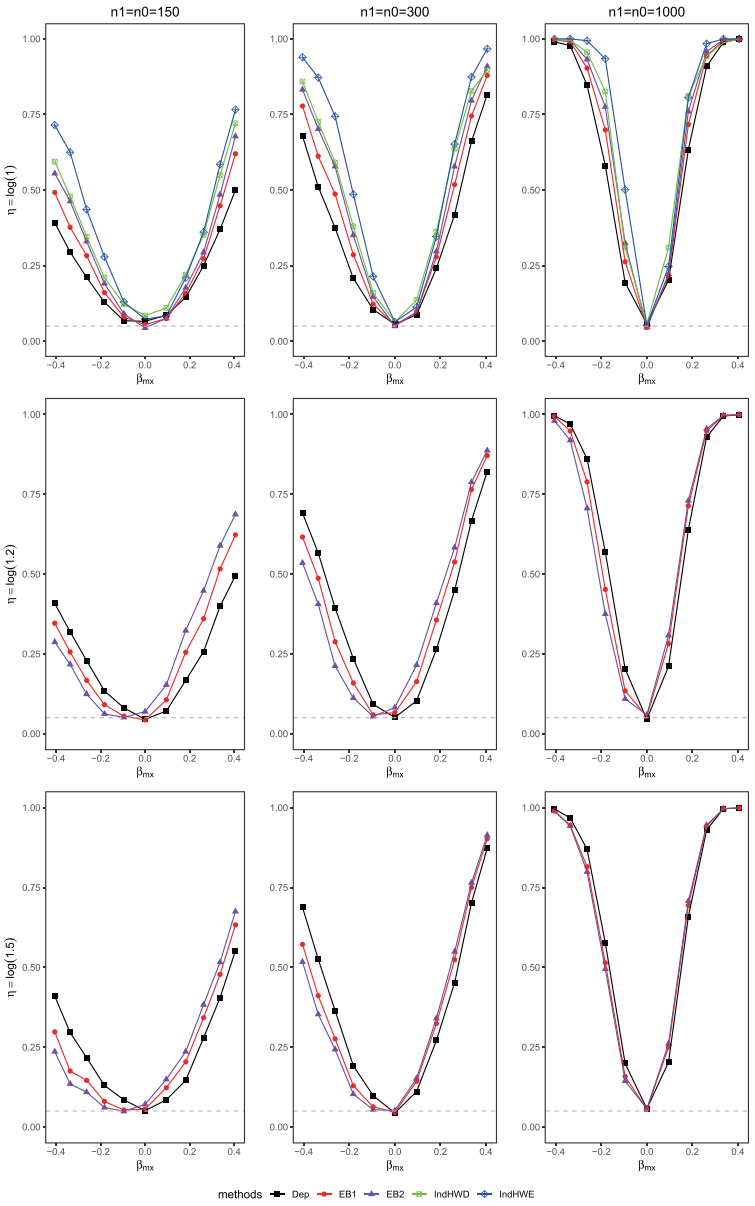


Fig. 3. Powers for the significance tests of maternal gene-environment interaction (β_{mx_1}) with HWD ($\rho = 0.1$), various sample sizes ($n_1 = n_0 = 150$, $n_1 = n_0 = 300$, and $n_1 = n_0 = 1000$), and various η values (0, log(1.2), and log(1.5)). The other parameters were fixed: $\beta_c = \log(1.8)$, $\beta_m = \log(1.3)$, $\beta_{x_1} = \beta_{x_2} = -\log(1.2)$, $\beta_{cx_1} = \beta_{cx_2} = -\log(1.5)$, and $\beta_{mx_2} = \log(1.2)$.

days < 259) and 906 term deliveries (280 ⩽ gestational days ⩽ 286) among 1626 eligible mother-child pairs.

The SNPs rs9939609 in genes FTO and rs2684811 in gene IGF1R have been identified as associated with BMI and SPTB, respectively[26, 23]. In the current study, we focused on the common SNPs located within 50kb of rs9939609 (eight SNPs in this region) and rs2684811 (24 SNPs in this region). All these 32 SNPs appear to be in HWE according to our analysis results by PLINK[27]; refer to Supporting Information Table S24 for the estimated MAFs and the p-values for testing HWE. Moreover, maternal pp-BMI can be assumed to be independent of the child genotype conditional on the maternal genotype since it is a maternal phenotype. Independent tests were also performed between pp-BMI and maternal genotype using the Kruskal-Wallis test. We applied the five methods mentioned in Section 3 to this dataset and adopted an additive mode of inheritance for each method. As the five methods are robust with respect to misspecification of disease prevalence according to our simulation results described in Section 3, we fixed the preterm birth prevalence at 7% because the prevalence of preterm birth in Denmark is approximately 5%–9% [28, 29] according to a survey of birth data in developed countries.

All the considered methods are expected to be nearly unbiased if G^m is independent of X, and EB1 and EB2 should be preferable to Dep in this situation. On the contrary, if G^m is strongly correlated to X, then IndHWE is generally biased, and EB1 and EB2 should be preferable to IndHWE and IndHWD. To verify this assertion, we studied the relationship between the shrinkage strengths of EB estimators and the association strength between X and G^m. First, a linear transformation was performed such that the transformed effects of Dep and IndHWE were 0 and 1, respectively. Then, the

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

shrinkage strengths of the two EB estimators were measured using the regression coefficient between the transformed EB estimators and the p-value for testing the independence between X and G^m. As shown in Supporting Information Fig. S4, the desired shrinkage strength of EB1 is significant (regression coefficient = 0.398; 95% confidence interval = [0.130, 0.665]), which is effect-independent (refer to Supporting Information Appendix S5 for a rigorous proof). In contrast, the shrinkage strength of EB2 is significant only for the G^m-X interaction effect β_{mx} (regression coefficient = 0.277; 95% confidence interval = [0.057, 0.497]). In the following, we describe the analysis results for two selected SNPs (rs16945088 and rs9941349) in detail, which are also summarized in Table 4. The results of the analysis of the other 30 candidate SNPs are summarized in Supporting Information Tables S25–S28.

For the first SNP rs16945088, the maternal genotype is shown to be independent of maternal pp-BMI (p-value = 0.636). Consequently, the effect estimates of Dep and IndHWE are close to each other, IndHWE is generally more efficient than Dep in terms of standard errors, and the estimates of EB1 and EB2 are much closer to those of IndHWE than to Dep. For example, the β_{mx} estimates of Dep, IndHWE, EB1, and EB2 are −0.081, −0.103, −0.102, and −0.103, respectively, indicating that the estimates of EB1 and EB2 are much closer to that of IndHWE than those of Dep. In addition, the standard errors of the β_{mx} estimates by IndHWE, EB1, and EB2 are 0.114, 0.116, and 0.115, respectively, compared to 0.160 by Dep, indicating that the former three methods are more efficient than Dep. For the second SNP rs9941349, maternal genotype is significantly associated with pp-BMI (p-value = 0.039). As a result, the estimates of EB1 are closer to those of Dep compared with IndHWE, whereas the estimates of EB2 are closer to those of IndHWE compared with Dep except that the β_{mx} estimate by EB2 lies in the middle of the β_{mx} estimates of Dep and IndHWE. For example, the β_{mx} estimates of Dep, EB1, and EB2 are 0.073, 0.084, and 0.113, respectively, compared to 0.176 by IndHWE. Furthermore, the G^m-X interaction is not significant for Dep, EB1, and EB2 (p-values > 0.3), whereas it is significant for IndHWE (p-value = 0.017). This suggests that the seemingly significant result of IndHWE might be a false positive finding.

5 Conclusions

In this study, two EB-type estimators (EB1 and EB2) are proposed by appropriately weighting an independent model estimator and a dependent-model estimator, which can adaptively balance robustness and efficiency. The independent-model estimator is an efficient estimator that improves efficiency by imposing HWE and G^m-X independence assumptions in the model, but it could produce a serious estimation bias if at least one of the assumptions is violated. In contrast, the dependent-model estimator Dep is a robust estimator that does not require the above two assumptions but might not be efficient. The EB-type estimators EB1 and EB2 developed in this study adaptively utilize the HWE assumption and the independence assumption between maternal genotype and maternal environmental risk factors, which appropriately balance robustness and efficiency. As in Dep, EB1 and EB2 fully utilize Mendelian inheritance and the conditional independence assumption between maternal environmental risk factors and children's genotype given maternal genotype.

These properties were verified by our simulation results in various situations. Specifically, if both HWE and G^m-X independence assumptions hold, EB1 and EB2 are more efficient than Dep, although they are less efficient than IndHWE. In contrast, IndHWE could be seriously biased if HWE or G^m-X independence assumptions are violated, while EB1 and EB2 still perform well. If the sample size is small, EB1 is more robust than EB2, whereas EB2 appears more suitable for moderate or large samples in terms of estimation efficiency. The desired properties of EB1 and EB2 are also demonstrated using a real-data application. In practice, it is difficult to obtain prior information regarding the relationship between maternal genotype and environmental risk factors. In this situation, EB1 and EB2 serve as robust alternatives to IndHWE and Dep.

Our novel method has been designed to analyze common genetic variants (MAF⩾ 5%) and the low-dimensional covariate X. In practice, there could be many mother-related characteristics, such as age, pre-pregnancy BMI, and smoking. The dimension of the parameters could be at least three times that of X when studying gene-environment interactions in case-

Table 4. Genetic effect estimates of two common SNPs in genes FTO and IGF1R (Danish National Birth Cohort).

SNP[a]	log(OR)	Dep			IndHWE			EB1			EB2		
		Est[b]	SE[c]	P[d]	Est[b]	SE[c]	P[d]	Est[b]	SE[c]	P[d]	Est[b]	SE[c]	P[d]
rs16945088	β_c	–0.151	0.140	0.283	–0.146	0.127	0.250	–0.146	0.127	0.249	–0.146	0.127	0.250
(0.636)	β_m	0.055	0.147	0.709	0.049	0.125	0.694	0.050	0.125	0.692	0.049	0.125	0.694
	β_{cx}	0.045	0.118	0.702	0.044	0.115	0.703	0.044	0.116	0.703	0.044	0.115	0.703
	β_{mx}	–0.081	0.160	0.613	–0.103	0.114	0.367	–0.102	0.116	0.382	–0.103	0.115	0.373
rs9941349	β_c	0.053	0.081	0.511	0.109	0.071	0.127	0.059	0.079	0.455	0.091	0.077	0.236
(0.039)	β_m	0.034	0.086	0.693	–0.031	0.074	0.676	0.027	0.083	0.747	–0.007	0.082	0.929
	β_{cx}	–0.070	0.073	0.341	–0.071	0.073	0.327	–0.070	0.073	0.340	–0.071	0.073	0.327
	β_{mx}	0.073	0.083	0.380	0.176	0.074	0.017	0.084	0.082	0.304	0.113	0.085	0.180

[Note] [a] Selected SNPs (the first one is BMI-related candidate SNP in gene FTO, and the second one is SPTB-related candidate SNP in gene IGF1R). Presented in parentheses are p-values for testing the independence between the SNPs and maternal pp-BMI using Kruskal-Wallis test.[b] Estimated effects. [c] Estimated standard errors.[d] Significance test p-values.

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3

control mother-child pair data. With high-dimensional parameters, we can adopt any regularization method to select associated covariates. However, increasing evidence shows that rare genetic variants (MAF< 1%) play an important role in complex diseases and traits[30–32]. Recently, some methods have been developed to analyze family based multilocus rare variants[33–36]. However, these methods are not based on retrospective likelihood functions; therefore, they cannot fully utilize family information, such as Mendelian inheritance and gene-environment independence. Further investigation is thus needed to extend our method to analyze multilocus rare variants under a case-control mother-child pair design.

Data availability statement

The DNBC data are available at the dbGAP website (study accession: phs000103.v1.p1).

Supporting information

The supporting information for this article is available online at https://doi.org/10.52396/JUSTC-2022-0007. It includes Appendices S1–S5, Figures S1–S4, and Tables S1–S28.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (11771096, 12171451, and 72091212). Funding support for the GWAS of Prematurity and Its Complications study was provided by the NIH Genes, Environment, and Health Initiative [GEI] (U01HG004423). The GWAS of Prematurity and its Complications study is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as general study coordination, was provided by the GENEVA Coordinating Center (U01HG004446). The National Center for Biotechnology Information provided assistance with data cleaning. Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438) and the NIH contract "High-throughput Genotyping for Studying the Genetic Contributions to Human Disease" (HHSN268200782096C).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Yanan Zhao is currently a PhD student under the tutelage of Prof. Hong Zhang at the University of Science and Technology of China. Her research interests focus on empirical Bayes procedures and statistical genetics.

Hong Zhang is a Full Professor with the University of Science and Technology of China (USTC). He received his Bachalor's degree in Mathematics and Ph.D. degree in Statistics from USTC in 1997 and 2003, respectively. His major reseach interests include statistical genetics, causal inference, and machine learning.

References

[1] Goddard K A, Tromp G, Romero R, et al. Candidate-gene association study of mothers with pre-eclampsia, and their infants, analyzing 775 SNPs in 190 genes. *Human Heredity,* **2007**, *63* (1): 1–16.
[2] Kanayama N, Takahashi K, Matsuura T, et al. Deficiency in $p57^{Kip2}$ expression induces preeclampsia-like symptoms in mice. *Molecular Human Reproduction,* **2002**, *8* (12): 1129–1135.
[3] Saftlas A F, Beydoun H, Triche E. Immunogenetic determinants of preeclampsia and related pregnancy disorders: A systematic review. *Obstetrics and Gynecology,* **2005**, *106* (1): 162–172.
[4] Wangler M F, Chang A S, Moley K H, et al. Factors associated with preterm delivery in mothers of children with Beckwith-Wiedemann syndrome: A case cohort study from the BWS registry. *American Journal of Medical Genetics Part A,* **2005**, *134* (2): 187–191.
[5] Goldenberg R L, Culhane J F, Iams J D, et al. Epidemiology and causes of preterm birth. *The Lancet,* **2008**, *371* (9606): 75–84.
[6] Zhang G, Feenstra B, Bacelis J, et al. Genetic associations with gestational duration and spontaneous preterm birth. *The New England Journal of Medicine,* **2017**, *377* (12): 1156–1167.
[7] Hong X, Hao K, Ji H, et al. Genome-wide approach identifies a novel gene-maternal pre-pregnancy BMI interaction on preterm birth. *Nature Communications,* **2017**, *8* (1): 15608.
[8] Chen J, Zheng H, Wilson M L. Likelihood ratio tests for maternal and fetal genetic effects on obstetric complications. *Genetic Epidemiology,* **2009**, *33* (6): 526–538.
[9] Fu W, Li M, Sun K, et al. Testing maternal-fetal genotype incompatibility with mother-offspring pair data. *Journal of Proteomics and Genomics Research,* **2013**, *1* (2): 40–56.
[10] Chen J, Lin D, Hochner H. Semiparametric maximum likelihood methods for analyzing genetic and environmental effects with case-control mother-child pair data. *Biometrics,* **2012**, *68* (3): 869–877.
[11] Lin D, Weinberg C R, Feng R, et al. A multi-locus likelihood method for assessing parent-of-origin effects using case-control mother-child pairs. *Genetic Epidemiology,* **2013**, *37* (2): 152–162.
[12] Prentice R L, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika,* **1979**, *66* (3): 403–411.
[13] Shi M, Umbach D M, Vermeulen S H, et al. Making the most of case-mother/control-mother studies. *American Journal of Epidemiology,* **2008**, *168* (5): 541–547.
[14] Zhang H, Mukherjee B, Arthur V, et al. An efficient and computationally robust statistical method for analyzing case-control mother-offspring pair genetic association studies. *Annals of Applied Statics,* **2020**, *14* (2): 560–584.
[15] Chen Y H, Chatterjee N, Carroll R J. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association,* **2009**, *104* (485): 220–233.
[16] Owen A B. Empirical Likelihood. New York: Chapman and Hall/CRC, 2001.
[17] Zhang H, Chatterjee N, Rader D, et al. Adjustment of nonconfounding covariates in case-control genetic association studies. *Annals of Applied Statistics,* **2018**, *12* (1): 200–221.
[18] Casella G, Berger R L. Statistical Inference. 2nd edition. Boston, MA: Cengage Learning, 2001.
[19] 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.
[20] Zhang K, Zhang H, Hochner H, et al. Covariate adjusted inference of parent-of-origin effects using case-control mother-child paired multilocus genotype data. *Genetic Epidemiology,* **2021**, *45* (8): 830–847.
[21] Engel S A M, Erichsen H C, Savitz D A, et al. Risk of spontaneous preterm birth is associated with common proinflammatory cytokine

polymorphisms. *Epidemiology,* **2005**, *16* (4): 469–477.
[22] Frey H A, Stout M J, Pearson L N, et al. Genetic variation associated with preterm birth in African-American women. *American Journal of Obstetrics and Gynecology,* **2016**, *215* (2): 235.e1–235.e8.
[23] Haataja R, Karjalainen M K, Luukkonen A, et al. Mapping a new spontaneous preterm birth susceptibility gene, *IGF1R*, using linkage, haplotype sharing, and association analysis. *PLoS Genetics,* **2011**, *7* (2): e1001293.
[24] Menon R, Velez D R, Simhan H, et al. Multilocus interactions at maternal tumor necrosis factor-α, tumor necrosis factor receptors, interleukin-6 and interleukin-6 receptor genes predict spontaneous preterm labor in European-American women. *American Journal of Obstetrics and Gynecology,* **2006**, *194* (6): 1616–1624.
[25] Hendler I, Goldenberg R L, Mercer B M, et al. The preterm prediction study: Association between maternal body mass index and spontaneous and indicated preterm birth. *American Journal of Obstetrics and Gynecology,* **2005**, *192* (3): 882–886.
[26] Frayling T M, Timpson N J, Weedon M N, et al. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science,* **2007**, *316* (5826): 889–894.
[27] Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics,* **2007**, *81* (3): 559–575.
[28] Hamilton B E, Martin J A, Ventura S J. Births: Preliminary data for 2005. *National Vital Statistics Reports,* **2006**, *55* (11): 1–18.
[29] Slattery M M, Morrison J J. Preterm delivery. *The Lancet,* **2002**, *360* (9344): 1489–1497.
[30] Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics,* **2008**, *40* (6): 695–701.
[31] Lee S, Abecasis G R, Boehnke M, et al. Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics,* **2014**, *95* (1): 5–23.
[32] Schork N J, Murray S S, Frazer K A, et al. Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics and Development,* **2009**, *19* (3): 212–219.
[33] Ionita-Laza I, Lee S, Makarov V, et al. Family-based association tests for sequence data, and comparisons with population-based association tests. *European Journal of Human Genetics,* **2013**, *21* (10): 1158–1162.
[34] Jiang D, McPeek M S. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genetic Epidemiology,* **2014**, *38* (1): 10–20.
[35] Wang X, Lee S, Zhu X, et al. GEE-based SNP set association test for continuous and discrete traits in family-based association studies. *Genetic Epidemiology,* **2013**, *37* (8): 778–786.
[36] Wang X, Zhang Z, Morris N, et al. Rare variant association test in family-based sequencing studies. *Briefings in Bioinformatics,* **2016**, *18* (6): 954–961.

DOI: 10.52396/JUSTC-2022-0007
JUSTC, **2022**, 52(5): 3