Practice of Epidemiology

On Information Coded in Gene-Environment Independence in Case-Control Studies

Hua Yun Chen* and Jinbo Chen

* Correspondence to Dr. Hua Yun Chen, Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL 60612 (e-mail: hychen@uic.edu).

Initially submitted September 2, 2010; accepted for publication April 14, 2011.

For analysis of case-control genetic association studies, it has recently been shown that gene-environment independence in the population can be leveraged to increase efficiency for estimating gene-environment interaction effects in comparison with the standard prospective analysis. However, for the special case in which data on the binary phenotype and genetic and environmental risk factors can be summarized in a $2 \times 2 \times 2$ table, the authors show here that there is no efficiency gain for estimating interaction effects, nor is there an efficiency gain for estimating the genetic and environmental main effects. This contrasts with the well-known result assuming that rare phenotype prevalence and gene-environment independence in the control population for the same data can lead to efficiency gain. This discrepancy is counterintuitive, since the 2 likelihoods are also approximately equal when the phenotype is rare. An explanation for the paradox based on a theoretical analysis is provided. Implications of these results for data analyses are also examined, and practical guidance on analyzing such case-control studies is offered.

case-control studies; epidemiologic methods; logistic models; odds ratio; prospective sampling; rare diseases; retrospective sampling

Abbreviations: G-E, gene-environment; PLR, prospective logistic regression; RML, retrospective maximum likelihood; SNP, single nucleotide polymorphism.

To estimate odds ratio parameters for genetic main effects and gene-environment (G-E) interactions using data from case-control genetic association studies, the traditional prospective likelihood method (1–6) has been the standard approach to analysis. Recently, it has been found that the G-E independence assumption can be leveraged to increase the efficiency of estimation (7–10). These methods were essentially based on maximizing the retrospective likelihood where the G-E independence was appropriately reflected, while the joint distribution of $G$ and $E$ was nonparametric in the standard approach. Some of these methods (7, 8) were based on log-linear models with the G-E independence assumption imposed on the control population, and some (9, 10) were based on logistic regression models with the G-E independence assumption imposed on the general study population. When an additional assumption that the binary phenotype is rare is imposed, the model of Chatterjee and Carroll (9) can be approximated by the model of Umbach and Weinberg (8). In the latter case, the need to estimate the intercept parameter in the logistic regression model is eliminated. Without the rare phenotype assumption and when both the gene and the environment have effects on the phenotype, Chatterjee and Carroll (9) showed an interesting result: The intercept parameter in the logistic model became identifiable under the G-E independence assumption in the general population. This contrasts with the widely known result that the baseline risk for the phenotype is unidentifiable from case-control data using prospective logistic regression analysis.

It turns out that the need to estimate the additional identifiable parameter compromises the efficiency gain from incorporating the G-E independence, and we illustrate this point here. We show that the magnitude of the efficiency gain under the G-E independence assumption depends on the structure of the data on genetic and environmental risk factors. In the
Let $D$ denote the binary phenotype, $G$ denote the genetic factor, and $E$ denote the environmental factor (or the other genetic factor in studies of gene-gene interaction). Let $S = 1$ indicate that a subject in the general study population is selected into the case-control sample. In the case-control design, the sampling probability depends only on the disease status, that is, $P(S = 1 | D, G, E) = P(S = 1 | D)$. Assume that $G$ and $E$ contribute to the risk of phenotype $D$ following a logistic regression model

$$
\log \frac{P(D = 1 | G, E)}{P(D = 0 | G, E)} = \beta_0 + \beta_1 G + \beta_2 E + \beta_{12} GE
$$

(1)

and that in the study population, the gene $G$ and the environmental factor $E$ are independently distributed, that is,

$$
p(G, E) = p(G)p(E).
$$

(2)

The contribution to the retrospective likelihood by a single subject, $p(G, E | D)$, can be expressed as

$$
P(D, G, E | D) = \frac{\eta \{D: (G, E)\} \{1 + e^{m(G, E)}\}^{-1} P(G)p(E)}{\sum_G \sum_E \eta \{D: (G, E)\} \{1 + e^{m(G, E)}\}^{-1} P(G)p(E)},
$$

where $\eta \{D: (G, E)\} = \exp(\beta_1 DG + \beta_2 DE + \beta_{12} DGE)$ and $m(G, E) = \beta_0 + \beta_1 G + \beta_2 E + \beta_{12} GE$. In the case-control sample, the proportions of cases or controls $P(D | S = 1) = n_0 / n$ are fixed by the sampling design, where $n_0$ and $n_1$ are the respective numbers of cases and controls, and $n = n_0 + n_1$. Let $\beta_0 = \beta_0 + \log \left( \frac{P(S = 1 | D = 1)}{P(S = 1 | D = 0)} \right)$ and define $p^*(D) = \exp(\beta_0 / \{1 + \exp(\beta_0)\})$, $p^*(G) = P(G = 0, D = 0)$, and $p^*(E) = P(E | G = 0, D = 0)$. By combining $P(G, E | D, S = 1)$ with $P(D | S = 1)$, the data $(D, G, E)$ for the case-control sample are distributed as follows:

$$
p(D, G, E | S = 1) = \eta \{D: (G, E)\} \eta_1 (G | E = 0) p^*(D) p^*(G) p^*(E) \sum_D \sum_G \sum_E \eta \{D: (G, E)\} \eta_1 (G | E = 0) p^*(D) p^*(G) p^*(E),
$$

(3)

where

$$
\eta_1 (G | E = 0) = \frac{\left(1 + e^{\beta_0 + \beta_1 G}\right) \left(1 + e^{\beta_0 + \beta_2 E}\right)}{\left(1 + e^{\beta_0}\right)\left(1 + e^{\beta_0 + \beta_1 GE + \beta_2 GE}\right)}.
$$

Chen (11) studied the problem of parameter identifiability and estimation for general biased sampling designs that included the case-control design as a special case. Let $\theta$ be parameters in $p(D, G, E | S = 1)$ other than $p^*(D)$ and $p^*(E)$. Lemma 1 below follows from the results presented by Chen (11).

**Lemma 1.** For the case-control design, if both $p^*(D)$ and $p^*(E)$ are unconstrained and variation-independent of each other and of $\theta$, the profile likelihood for parameters in $p(G, E | D)$ other than $p^*(E)$ from the conditional likelihood based on $p(D, G, E | S = 1)$ is the same as the profile likelihood for $\theta$ from the retrospective likelihood based on $p(G, E | D)$.

In a typical case-control design, $P(S = 1 | D = 1) / P(S = 1 | D = 0)$ is unknown. This means that $p^*(D)$ is unconstrained and is variation-independent of $\theta = (\beta_1, \beta_2, \beta_{12}, p^*(G))$ and $p^*(E)$. Under equations 1 and 2 and unconstrained $p(E)$, we can conclude from lemma 1 that the estimator from maximizing the joint conditional likelihood $\prod_{i=1}^n P(D_i, G_i | E_i, S_i = 1)$ is efficient if the estimator from maximizing the retrospective likelihood $\prod_{i=1}^n P(G_i, E_i | D_i)$ is efficient. The traditional method of analyzing case-control data maximizes the prospective likelihood $\prod_{i=1}^n P(D_i | G_i, E_i, S_i = 1)$. The joint conditional likelihood equals the prospective likelihood multiplied by $\prod_{i=1}^n P(G_i | E_i, S_i = 1)$. Lemma 1 implies that any additional information for estimating $\beta_0$ is useful. Due to the incorporation of the population-level G-E independence in the retrospective likelihood is contained in $\prod_{i=1}^n P(G_i | E_i, S_i = 1)$. Below, we analyze this term to characterize the relation between the magnitude of efficiency gain and the structure of the data on $(D, G, E)$. In the simplest case of a $2 \times 2 \times 2$ table, which is frequently used to summarize case-control data, we show that $\prod_{i=1}^n P(G_i | E_i, S_i = 1)$ contains no information for estimating $\beta_0$. Thus, the only difference between the traditional prospective analysis and the method of Chatterjee and Carroll (9) in this case is that $\beta_0$ is identifiable and can be estimated from the observed data when conditions on the identifiability of $\beta_0$ hold.

**RESULTS**

**G-E independence in the general population**

**Theorem 1.** Assume the case-control sampling probabilities are unknown and $p(E)$ is unconstrained. When the case-control data on $(D, G, E)$ can be summarized in a $2 \times 2 \times 2$ table, the
maximum likelihood estimators for \((\beta_1, \beta_2, \beta_{12})\) based on \(p(G, E|D)\) without or with the \(G-E\) independence assumption, that is, those obtained from maximizing \(\prod_{i=1}^{n} P(D_i|G_i, E_i, S_i = 1)\) and \(\prod_{i=1}^{n} P(D_i, G_i|E_i, S_i = 1)\), are identical.

Proof of theorem 1 is given in the Appendix. The result in theorem 1 can be understood as follows. Note that \(p(G|E, S = 1)\) involves the parameters \(\beta_0\) and \(p(G)\), which do not appear in \(p(D|G, E, S = 1)\). For a \(2 \times 2 \times 2\) table, the likelihood \(\prod_{i=1}^{n} P(D_i|G_i, E_i, S_i = 1)\) allows for the estimation of 2 independent parameters, \(p(G = 1|E = 1, S = 1)\) and \(p(G = 1|E = 0, S = 1)\), to which parameters \(\beta_0\) and \(p(G = 1)\) can be mapped. This equivalent parameterization means that there is no information left for estimating \((\beta_1, \beta_2, \beta_{12})\) in equation A4 (Appendix) after \(\beta_0\) and \(p(G = 1)\) are estimated. By analogy, for slightly richer data summarized in a \(2 \times 3 \times 2\) table, if 1 parameter is used to model the \(G-E\) association in the general population (12) and no additional constraint is imposed on \(p(G)\), we would expect no efficiency gain even if the model for \(p(G|E)\) is not saturated.

Theorem 1 has direct implications for detecting interaction effects between a single nucleotide polymorphism (SNP) and binary environmental exposures. A SNP with alleles \(B\) and \(b\) has 3 possible genotypes, \(BB, Bb,\) and \(bb\). Let \(G\) be the numerical coding for the genotype. The classical models for the SNP effect include the dominant model \((G = 1\) for \(BB\) and \(G = 0\) for \(bb\)), the recessive model \((G = 1\) for \(BB\) and \(G = 0\) for \(Bb\) and \(bb\)), and the additive model \((G = 0\) for \(bb, G = 1\) for \(Bb\), and \(G = 2\) for \(BB\)). With the dominant or recessive model for the main effect of \(G\) and a dichotomous environmental factor \(E\), theorem 1 implies that incorporating the \(G-E\) independence in the general population cannot improve statistical power for detecting \(G-E\) interaction effects. For the additive case on a logarithmic scale, some efficiency gain in estimating the odds ratio parameters can be achieved, but it has very limited magnitude.

It is important to note that when the phenotype prevalence rate, \(p(D = 1)\), is known, theorem 1 does not apply. That is, the efficiency of the \((\beta_1, \beta_2, \beta_{12})\) estimator from maximizing the retrospective likelihood may be higher than that from maximizing the prospective likelihood, even in the case of a \(2 \times 2 \times 2\) table. On the other hand, it is recognized that knowing the disease prevalence rate does not help in estimating \((\beta_1, \beta_2, \beta_{12})\) if \(p(G, E)\) is unconstrained, because the additional constraints imposed by fixed \(p(D = 1)\) happen to be equivalent to the constraints imposed by fixing the numbers of cases and controls for the estimation of odds ratio parameters (13). The key reason that theorem 1 does not hold is that conditions for applying lemma 1 are no longer satisfied. When the phenotype prevalence is known, \(\beta_0^*\) is fully determined by \(\beta_0\):

\[
\beta_0^* = \beta_0 + \log \frac{P(D = 1|S = 1)P(D = 0)}{P(D = 0|S = 1)P(D = 1)}
\]

\[
= \beta_0 + \log \frac{n_1P(D = 0)}{n_0P(D = 1)}.
\]

Thus, \(p^*(D)\) is no longer variation-independent of 0 under the \(G-E\) independence. These results suggest that if we would
like to exploit the G-E independence for efficiency gain in estimating the odds ratio parameter, one strategy would be to incorporate knowledge of the disease prevalence rate.

Chatterjee and Carroll (9) noticed that it is challenging to estimate the intercept parameter $\beta_0$ based on their profile likelihood method, despite the fact that $\beta_0$ is theoretically identifiable when $\beta_{12} \neq 0$ or $\beta_1 \neq 0$ and $\beta_2 \neq 0$. (See a related problem discussed in the article by Neuhaus et al. (14).) The difficulty in estimating $\beta_0$ with data summarized in a 2 x 2 table may be explained as follows. Given $(\beta_1, \beta_2, \beta_{12})$, the part of the likelihood useful for estimating $\beta_0$ is

$$
g(\beta_0) = \frac{1 + e^{\beta_0 + \beta_1}}{1 + e^{\beta_0}} \cdot \frac{1 + e^{\beta_0 + \beta_2}}{1 + e^{\beta_0 + \beta_1 + \beta_{12}}}. $$

It can be seen that $\lim_{\beta_0 \to -\infty} g(\beta_0) = \exp(-\beta_{12})$ and $\lim_{\beta_0 \to +\infty} g(\beta_0) = 1$. Thus, $g(\beta_0)$ is a bounded function. Figure 1 shows 2 typical shapes of the function with a positive or negative interaction effect and 2 typical shapes of the function in the absence of an interaction effect.

Since the second part of the right-hand side of equation A9 (i.e., $n_{11} p_{10}/n_{10} p_{01}$) can be any number due to the random variation, the solution to equation 14 may not exist for a data set with a finite sample size. However, this difficulty would be less likely to arise with extremely large sample sizes.

### Rare phenotype assumption and G-E independence in the control population

Assuming a log-linear model for analyzing case-control data $(D, G, E)$, Umbach and Weinberg (8) showed that efficiency for estimating the odds ratio parameters when G-E independence is imposed on the control population is higher than that in the traditional prospective analysis, even for data summarized by a 2 x 2 x 2 table. However, it is well known that a log-linear model approximates a logistic regression model adopted by Chatterjee and Carroll (9) when the phenotype is rare, that is, $p(D = 1) \approx 0$, and G-E independence holds in the general population. We explain these seemingly contradictory results as follows. Note that the assumptions in the article by Umbach and Weinberg (8) are equivalent to equation 1 and $p(G, E|D = 0) = p(G|D = 0)p(E|D = 0)$. By the odds ratio formulation of Chen (11), the joint distribution of $(D, G, E)$ under the case-control design can be rewritten as

$$
p(D, G, E|S = 1) = \frac{\eta \{D; (G, E)\} p^*(D)p(G|D = 0)p(E|D = 0)}{\sum_D \sum_G \sum_E \eta \{D; (G, E)\} p^*(D)p(G|D = 0)p(E|D = 0)}. $$

In this case, parameters in $p(G, E|D)$ other than $p(E|D = 0)$ are $(\beta_1, \beta_2, \beta_{12}, p(G|D = 0))$, which do not include $\beta_0$. If the distribution $p(E|D = 0)$ is not constrained, lemma 1 can be applied to obtain the efficient parameter estimator for $(\beta_1, \beta_2, \beta_{12}, p(G|D = 0))$ by profiling $p(E|D = 0)$ out from the unconstrained joint likelihood based on $p(D, G, E|S = 1)$. The profile likelihood is equivalent to the conditional likelihood

$$
\prod_{i=1}^{n} \eta \{D; (G_i, E_i)\} p^*(D)p(G_i|D = 0). $$

The likelihood equation 5 is approximately the same as equation A2 under the rare phenotype assumption, that is, $\exp(m(G, E)) \approx 0$. Note, however, that the property of likelihood equation 5 is very different from that of likelihood equation A2. First, theorem 1 is no longer applicable to equation 5 and the estimator of $(\beta_1, \beta_2, \beta_{12})$ based on likelihood equation 5 can be more efficient than that based on

### Table 1. Simulation Results for Odds Ratio Estimation Under Gene-Environment (G-E) Independence in the General Population With an Unknown Case Prevalence Rate

<table>
<thead>
<tr>
<th>Effect</th>
<th>Truth</th>
<th>Bias (Estimated Truth)</th>
<th>SD of the Parameter Estimate</th>
<th>Average of the SE Estimates</th>
<th>Bias (Estimated Truth)</th>
<th>SD of the Parameter Estimate</th>
<th>Average of the SE Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
<td>1.0</td>
<td>0.003</td>
<td>0.0511</td>
<td>0.0525</td>
<td>0.003</td>
<td>0.0510</td>
<td>0.0525</td>
</tr>
<tr>
<td>$E$</td>
<td>1.0</td>
<td>0.001</td>
<td>0.0141</td>
<td>0.0139</td>
<td>0.001</td>
<td>0.0141</td>
<td>0.0139</td>
</tr>
<tr>
<td>$G \times E$</td>
<td>0.5</td>
<td>0.000</td>
<td>0.0404</td>
<td>0.0434</td>
<td>0.004</td>
<td>0.0404</td>
<td>0.0433</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.0</td>
<td>-0.003</td>
<td>0.1032</td>
<td>0.1013</td>
<td>-0.003</td>
<td>0.1032</td>
<td>0.1013</td>
</tr>
</tbody>
</table>

Abbreviations: SD, standard deviation; SE, standard error.

a The simulation results are based on 1,000 replicates in a sample of 50,000 with 25,000 cases and 25,000 controls.

b Average of the SE estimates of the parameter estimator.
the traditional prospective likelihood. Although equation 5 can be factored into equation A3 and
\[
\prod_{i=1}^{n} \frac{\sum_{D} \sum_{G} \eta(D; (G, E_i)) p^*(D) p(G_i | D = 0)}{\sum_{D} \sum_{G} \eta(D; (G, E_i)) p^*(D) p(G | D = 0)},
\]
there is only 1 parameter to be estimated from equation 6, aside from \( p^*(D) \), and \( p^*(D) \) needs to be estimated jointly with equation A3. Efficiency gain in estimating \( \beta_0 \) leads to more efficient estimation of \( (\beta_1, \beta_2, \beta_1 \beta_2) \) than with the traditional prospective likelihood estimator. Second, when \( p(D = 1) \) is known, \( p^*(D) \) is determined by \( \beta_0 \), which is not contained in \( p(G, E | D) \) under the \( G-E \) independence in the control population. This means that lemma 1 is still applicable to likelihood equation 4 because the variation-independence condition in lemma 1 is satisfied. As a result, knowing \( p(D = 1) \) does not help increase the efficiency for estimating \( (\beta_1, \beta_2, \beta_1 \beta_2, p(G | D = 0)) \). This is very different from the case where \( G-E \) independence is assumed for the general population. The false contradiction observed earlier is based on the wrong intuition that the maximum likelihood estimators should have approximately equal efficiency if the 2 likelihoods are approximately equal.

### SIMULATION STUDIES

We performed 2 simulation studies to assess the magnitude of efficiency gain under the \( G-E \) independence assumption in the general population, and we performed 1 simulation study to assess bias when this \( G-E \) independence assumption is violated. In the first study, the traditional prospective logistic regression (PLR) analysis based on \( p(D | G, E, S = 1) \) and the prospective analysis based on \( p(D, G | E, S = 1) \) in the article by Chatterjee and Carroll (9) (retrospective maximum likelihood (RML)) are compared. Models 1 and 2 are used to simulate data. Because \( G \) and \( E \) are in symmetric positions, we assume in the first simulation study that \( G \) takes the values 0 and 1 only and \( E \) either has 3 different categories or is a continuous variable. Since the relative efficiency is of primary interest and the intercept parameter is difficult to estimate, we choose a very large sample size.
in the first simulation study. The sample size is 50,000 with 25,000 cases and 25,000 controls. In this simulation, $P(G = 1) = 0.1$, and $E$ is uniformly distributed on $[0, 2]$ or on $[0, 2]$.

The log odds ratio parameters ($\beta_1, \beta_2, \beta_{12}$) are $(1, 1, 0.5)$ and the intercept $\beta_0 = -3$. The relatively large $G$-$E$ effects were primarily chosen to make the computational method more stable. Results based on 1,000 replicates are shown in Table 1.

In the second simulation study, we compared the traditional prospective logistic analysis with the approach proposed by Chatterjee and Carroll (9) assuming a known correct or incorrect phenotype prevalence rate, and as in the article by Umbach and Weinberg (8) assuming $G$-$E$ independence in the control population. We generated independent binary $G$ and $E$ with $p(G = 1) = 0.1$ and $p(E = 1) = 0.5$. The log odds ratio association parameters in model 1 were more realistically set to $(0.2, 0.1, 0.5)$, and we varied the intercept $\beta_0$ from $-1.2$ to $-2.3$ to $-7.0$. Five estimators of the relative risk parameters were computed: the traditional PLR estimator, Chatterjee and Carroll’s approach assuming

### Table 3. Simulation Results for Odds Ratio Estimation with Incorrectly Assumed Gene-Environment ($G$-$E$) Independence in the General Population

<table>
<thead>
<tr>
<th>Method</th>
<th>$G(0.2)$</th>
<th>$E(0.1)$</th>
<th>$G \times E(0.5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias (Estimated Truth)</td>
<td>SD of the Parameter Estimate</td>
<td>Average of the SE Estimates</td>
</tr>
<tr>
<td>PLR</td>
<td>-0.003</td>
<td>0.139</td>
<td>0.138</td>
</tr>
<tr>
<td>UpRML</td>
<td>-0.184</td>
<td>0.138</td>
<td>0.138</td>
</tr>
<tr>
<td>pRML</td>
<td>-0.184</td>
<td>0.128</td>
<td>0.128</td>
</tr>
<tr>
<td>LpRML</td>
<td>-0.165</td>
<td>0.123</td>
<td>0.122</td>
</tr>
<tr>
<td>UW</td>
<td>-0.101</td>
<td>0.120</td>
<td>0.119</td>
</tr>
</tbody>
</table>

Abbreviations: OR, odds ratio; PLR, prospective logistic regression; RML, retrospective maximum likelihood; SD, standard deviation; SE, standard error; UW, Umbach and Weinberg.

*The simulation results are based on 1,000 replicates in a sample of 5,000 with 2,500 cases and 2,500 controls.

Average of the SE estimates of the parameter estimator.

Traditional PLR approach.

RML approach assuming a known wrong disease prevalence rate as twice that of the truth.

RML approach assuming a known true disease prevalence rate.

RML approach assuming a known wrong disease prevalence rate as half that of the truth.

Umbach and Weinberg’s (8) approach, corresponding to LpRML with a known wrong disease prevalence rate of 0.

Notice that our sample size is 10 times that used by Chatterjee and Carroll (9). After adjusting for the sample size difference, we obtain $\sqrt{10}(0.100, 0.011, 0.041) = (0.32, 0.035, 0.130)$, which are close to the $(0.32, 0.037, 0.128)$ obtained by Chatterjee and Carroll (9). Indeed, the efficiency gain for the interaction parameter in this case is much larger than that in our simulation setting. This may suggest that the efficiency gain depends on the underlying covariate distribution.
that the true phenotype prevalence rate (pRML) is known, Chatterjee and Carroll’s approach assuming an incorrect phenotype prevalence rate (twice the true rate (UpRML) and half of the true rate (LpRML)), and Umbach and Weinberg’s approach, which effectively assumes an incorrect phenotype prevalence rate at 0. The simulation results, shown in Table 2, are based on 1,000 repetitions of a sample size of 5,000 with 2,500 cases and 2,500 controls.

The simulation results show that pRML had a negligible bias and a big reduction in variance compared with PLR. Both UpRML and LpRML had smaller bias than Umbach and Weinberg’s estimator. However, the simulation also shows that the variance of the estimator decreases as the assumed incorrect phenotypic prevalence rate decreases. In particular, the UpRML estimator has a larger variance than did the pRML estimator, which may suggest that an upward bias in the assumed incorrect phenotypic prevalence rate can result in loss of efficiency. Umbach and Weinberg’s estimator has the smallest variance among all of the estimators, but it can have relatively large bias when the phenotypic prevalence rate is large. For a very small prevalence rate, all of the estimators with a known prevalence rate appear to have the same performance. This suggests that the additional efficiency gain in Umbach and Weinberg’s approach compared with Chatterjee and Carroll’s approach without knowledge of the prevalence rate is due to the known prevalence rate of approximately 0 in the former. Overall, the result suggests that when we do not have exact information on the prevalence rate, it is still useful to use an inaccurate prevalence rate for estimation and inference.

In the third simulation study, correlated binary G and E are simulated with $P(G = 1, E = 1) = pq$, $P(G = 1, E = 0) = p(1 - q)$, $P(G = 0, E = 1) = (1 - p)(1 - q)$, and $P(G = 0, E = 0) = (1 - pq)$, where $p = 0.1$ and $q = 0.55$. The association between $G$ and $E$ has an odds ratio of approximately 1.49. All other model parameters are the same as in the second simulation setting. In the analysis, $G$ and $E$ are modeled as independent, and the same methods of analysis as in the second simulation are used. Simulation results are shown in Table 3. These results show that, with moderate deviation from the $G$-$E$ independence assumption, large bias occurs with all of the estimators except for the PLR. The results suggest that approaches exploiting the $G$-$E$ independence assumption for efficiency gain are all very sensitive in terms of bias to the deviation from the independence assumption. It is very important to check the validity of the assumption in applying these approaches.

**DISCUSSION**

Here we showed that estimation of the identifiable intercept parameter in the logistic regression model under certain constraints on the covariate distribution could compromise efficiency for estimating odds ratio parameters. Thus, the information coded in the $G$-$E$ independence assumption alone for assessing joint effects of $G$-$E$ variables can become limited with a binary genetic variable, particularly when the environmental variable takes only a small set of unique values. The amount of information increases as the data structure on $(G, E)$ becomes more complex. In this sense, haplotype-based association analyses may be preferable to single-SNP-based analyses. In the same sense, one would probably need to be cautious in deciding whether to dichotomize a continuous environmental exposure. Of course, any analysis would have to be guided by scientific inquiry, rather than statistical efficiency. In addition, implementing the approach of Chatterjee and Carroll (9) without assuming any knowledge of the phenotypic prevalence rate is inherently computationally challenging.

Additional knowledge on the prevalence rate can be leveraged to increase the information coded in $G$-$E$ independence and largely eradicates the difficulty in parameter estimation. The approach of Umbach and Weinberg (8) does not require us to know the exact phenotypic prevalence rate. However, Chatterjee and Carroll (9) noticed the subtlety that not only should the overall phenotype prevalence be small but it should be small in every population subgroup defined by covariates. When the phenotypic prevalence rate is not very low, the approach of Umbach and Weinberg (8) can have substantial bias. The approach of Chatterjee and Carroll (9) assuming a known phenotypic prevalence rate can have a substantial efficiency gain over the traditional prospective analysis. When the phenotypic prevalence rate is very low and is known, the methods of Umbach and Weinberg (8) and Chatterjee and Carroll (9) have comparable performance.

When the phenotypic prevalence rate is not known, attempting to estimate the intercept from the case-control data under the $G$-$E$ independence assumption may not be a good choice. We see from these simulation results that, even if we do not know the exact phenotypic prevalence rate, assuming an incorrect phenotypic prevalence rate can still lead to a substantial reduction in variance with some bias in the parameter estimator. The exact cutoff in the bias and variance trade-off requires further research. Finally, caution should be exercised in using the approaches for efficiency gain because of their sensitivity to deviation from the independence assumption.

**ACKNOWLEDGMENTS**

Author affiliations: Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, Illinois (Hua Yun Chen); and Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania (Jinbo Chen).

This work was partially supported by the Division of Mathematical Science of the National Science Foundation (grant DMS 1007726 to H. Y. C.) and the National Institutes of Health (grant R01 ES016626 to J. C.).

Conflict of interest: none declared.

**REFERENCES**


APPENDIX

**Proof of Theorem 1**

Lemma 1 implies that the maximum likelihood estimator of $0$ can be obtained by maximizing the conditional likelihood

$$
\prod_{i=1}^{n} \sum_{D} \sum_{G} \eta(D; (G; E)) p^*(D) \eta_1(G; E; D = 0) p^*(G),
$$

where the second part is based on $P(D; G, E, S = 1)$, which is the likelihood employed by the traditional prospective analysis and equals

$$
\prod_{i=1}^{n} \sum_{D} \sum_{G} \eta(D; (G; E)) p^*(D) \{1 + e^{m(G;E)}\}^{-1} p^*(G).
$$

This likelihood can be factored into 2 parts. The first part is based on $P(D; G, E, S = 1)$, which is the likelihood employed by the traditional prospective analysis and equals

$$
\prod_{i=1}^{n} \sum_{D} \sum_{G} \eta(D; (G; E)) p^*(D) p^*(G).
$$

The second part is based on $P(G| E, S = 1)$ and equals

$$
\prod_{i=1}^{n} \sum_{D} \sum_{G} \eta(D; (G; E)) p^*(D) \{1 + e^{m(G;E)}\}^{-1} p^*(G).
$$

We note that parameters $\beta_0$ and $p(G)$ are only contained in the second part of equation A4, so that the maximum likelihood estimator for $\beta_0$ and $p(G)$ can be obtained by maximizing equation A4. Let $A_{GE} = \sum_{D} \sum_{G} \eta(D; (G; E)) p^*(D) \{1 + e^{m(G;E)}\}^{-1} p^*(G)$, $p_G = p(G)$, and $n_{ij}$ be the observed number of subjects with $G = k$ and $E = j$, $k = 0, 1$ and $j = 0, 1$, respectively. When data on $(D, G; E)$ can be summarized in a $2 \times 2$ table, likelihood equation A4 can be rewritten as

$$
\frac{\left( \frac{p_1 A_{11}}{p_1 A_{11} + p_0 A_{11}} \right)^{n_{11}} \left( \frac{p_0 A_{01}}{p_1 A_{11} + p_0 A_{01}} \right)^{n_{01}}}{\left( \frac{p_1 A_{10}}{p_1 A_{10} + p_0 A_{10}} \right)^{n_{10}} \left( \frac{p_0 A_{00}}{p_1 A_{10} + p_0 A_{00}} \right)^{n_{00}}}. \tag{A5}
$$

The possible maximum value that equation A5 can achieve is

$$
\left( \frac{n_{11}}{n_{11} + n_{10}} \right)^{n_{11}} \left( \frac{n_{01}}{n_{11} + n_{10}} \right)^{n_{01}} \left( \frac{n_{00}}{n_{00} + n_{10}} \right)^{n_{00}} \left( \frac{n_{00}}{n_{00} + n_{10}} \right)^{n_{00}}. \tag{A6}
$$

This requires that at least 1 solution exist for $(p_0, p_1)$ and $\beta_0$ for the following 2 equations:

$$
\frac{p_1}{p_0} = A_{01} n_{11} / A_{11} n_{11}, \tag{A7}
$$

$$
A_{11} A_{00} / A_{10} A_{01} = n_{11} n_{00} / n_{10} n_{10}. \tag{A8}
$$

where the second equation can be rewritten as

$$
\frac{(1 + e^{\beta_0 + \beta_1})(1 + e^{\beta_0 + \beta_2})}{(1 + e^{\beta_0})(1 + e^{\beta_0 + \beta_1} + e^{\beta_0 + \beta_2})} = \frac{(1 + e^{\beta_0 + \beta_1})(1 + e^{\beta_0 + \beta_2})}{(1 + e^{\beta_0})(1 + e^{\beta_0 + \beta_1} + e^{\beta_0 + \beta_2})},
$$

and

$$
\frac{(1 + e^{\beta_0 + \beta_1})(1 + e^{\beta_0 + \beta_2})}{(1 + e^{\beta_0})(1 + e^{\beta_0 + \beta_1} + e^{\beta_0 + \beta_2})} \converges to
\frac{(1 + e^{\beta_0 + \beta_1})(1 + e^{\beta_0 + \beta_2})}{(1 + e^{\beta_0})(1 + e^{\beta_0 + \beta_1} + e^{\beta_0 + \beta_2})},
$$

asymptotically, there exists a $\beta_0$ satisfying equation A9 for any given $(\beta_1, \beta_2, \beta_1, \beta_2)$ close to the true values. This further implies that $(p_0, p_1)$ values satisfying equation A7 exist. It follows that likelihood equation A5 asymptotically achieves the maximum of equation A6, which is independent of $(\beta_1, \beta_2, \beta_1, \beta_2)$. Whenever solutions to the equations A7 and A8 exist, the profile likelihood for $(\beta_1, \beta_2, \beta_1, \beta_2)$ based on equation A2 is the same as that based on equation A3. This completes the proof.