Published in final edited form as:

Stat Med. 2013 April 30; 32(9): 1494–1508. doi:10.1002/sim.5613.

Bias correction to secondary trait analysis with case–control design

Hua Yun Chen^{a,*†}, Rick Kittles^b, and Wei Zhang^c

^aDivision of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, IL, U.S.A

^bDepartment of Medicine, College of Medicine, Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, IL, U.S.A

^cDepartment of Pediatrics, College of Medicine, University of Illinois at Chicago, Chicago, IL, U.S.A

Abstract

In genetic association studies with densely typed genetic markers, it is often of substantial interest to examine not only the primary phenotype but also the secondary traits for their association with the genetic markers. For more efficient sample ascertainment of the primary phenotype, a case–control design or its variants, such as the extreme-value sampling design for a quantitative trait, are often adopted. The secondary trait analysis without correcting for the sample ascertainment may yield a biased association estimator. We propose a new method aiming at correcting the potential bias due to the inadequate adjustment of the sample ascertainment. The method yields explicit correction formulas that can be used to both screen the genetic markers and rapidly evaluate the sensitivity of the results to the assumed baseline case-prevalence rate in the population. Simulation studies demonstrate good performance of the proposed approach in comparison with the more computationally intensive approaches, such as the compensator approaches and the maximum prospective likelihood approach. We illustrate the application of the approach by analysis of the genetic association of prostate specific antigen in a case–control study of prostate cancer in the African American population.

Keywords

extreme-value sampling design; semi-parametric likelihood; odds ratio model; sensitivity analysis

1. Introduction

Biased sample ascertainment such as the case–control design for a qualitative trait or the extreme-value sampling design for a quantitative trait is frequently used for its sampling efficiency in genome-wide association studies with densely typed genetic markers. Data on secondary traits are often collected in such studies, and it is of substantial interest to explore the association of genetic markers with a secondary trait. The secondary trait sample thus ascertained may be neither a random sample nor a case–control sample from the study population because the induced sampling design for the secondary trait potentially depends on the genotype and the environmental factor in addition to the secondary trait itself. As a

result, analysis of the secondary trait by the traditional prospective logistic regression [1–4] or its extensions [5–8] may yield a biased estimator of the true genetic association. This problem had been previously studied in the context of reusing case–control data [9–12]. It was further considered in the secondary trait analysis in [13–18].

Methods proposed in the literature for the analysis of secondary traits are subject to problems. The secondary trait analysis using controls only requires a low case-prevalence rate of the primary phenotype in the general population and is inefficient. Analyses that ignore or condition on the primary phenotype [11–13], although simple, may yield a biased estimator of the genetic association. The method viewing the case–control sample of the primary phenotype as the second-stage sample in a two-stage sampling design [10–13, 16] is applicable only when additional data on the cohort sample wherein the case–control design is nested are available. The retrospective likelihood approach based on a joint model of the first and secondary traits [15, 19] performs well when supplementary information in terms of the population prevalence rate of the primary phenotype is known. However, when the population prevalence rate of the primary phenotype is unknown or known inaccurately, the retrospective likelihood approach can be computationally difficult to implement [20]. In addition, the use of the known population prevalence rate can be questionable when pre-selection of subjects through eligible criteria for subject inclusion is involved in the case–control sample ascertainment. The bias correction approach in [18] requires us to know the population prevalence rates for both the primary phenotype and the secondary trait.

In this article, we propose a new approach to bias correction in the secondary trait analysis. Our approach yields closed-form formulas for correcting bias in the marginal or conditional analysis of secondary traits. The proposed approach first performs a supplemental analysis of the primary phenotype given the secondary trait, the genetic marker, and possible environmental factors. The correction can then be carried out on the basis of the estimates from the supplemental analysis using the derived explicit formulas. The simple correction formulas enable us to rapidly perform the sensitivity analysis to the baseline case-prevalence rate for tens of thousands of genetic markers. We also propose more accurate but computationally slower compensator approaches and the maximum prospective likelihood approach with fixed baseline case-prevalence rate. We can use these approaches to validate the fast approximation approach proposed.

2. Correction to conditional analysis

2.1. Distribution of sample ascertained by the case–control design

Consider the secondary trait as a binary phenotype Y . Assume the association of secondary trait Y with genetic marker G , environment factor E , and possibly the interaction of G and E in the study population by the logistic regression,

$$p(Y=1|G, E) = \frac{\exp(\theta_0 + \theta_1 G + \theta_2 E + \theta_{12} GE)}{1 + \exp(\theta_0 + \theta_1 G + \theta_2 E + \theta_{12} GE)}. \quad (1)$$

Define the odds ratio function between Y and (G, E) as

$$\eta_2\{Y; (G, E)\} = \exp(\theta_1 Y G + \theta_2 Y E + \theta_{12} Y G E). \quad (2)$$

We can rewrite the logistic regression model as

$$p(Y|G, E) = \frac{\eta_2\{Y; (G, E)\} \exp(Y\theta_0)/\{1+\exp(\theta_0)\}}{\sum_{Y=0}^1 \eta_2\{Y; (G, E)\} \exp(Y\theta_0)/\{1+\exp(\theta_0)\}}, \quad (3)$$

where $\exp(Y\theta_0)/\{1 + \exp(\theta_0)\} = P(Y|G=0, E=0)$. (θ_1, θ_{12}) characterizes the association between the secondary trait and the genetic marker. We can carry out testing for no genetic association by testing $(\theta_1, \theta_{12}) = (0, 0)$.

In the analysis of a secondary trait, the sample on the secondary trait is not a random sample from the study population when the sample is ascertained by the case-control status of the primary phenotype. It also differs from a case-control sample ascertained on the status of the secondary trait. We need to carefully consider the sample ascertainment mechanism to avoid bias in the analysis. This entails us to take the primary phenotype D into consideration in the analysis of the secondary trait. Assume that the primary phenotype given the secondary trait, the genotype, and the environment factor follows the logistic regression model,

$$P(D=1|Y, G, E) = \frac{\exp(\gamma_0 + \gamma_{01}Y + \gamma_1G + \gamma_2E + \gamma_{12}GE)}{1 + \exp(\gamma_0 + \gamma_{01}Y + \gamma_1G + \gamma_2E + \gamma_{12}GE)}. \quad (4)$$

Similarly, define the odds ratio function between D and (Y, G, E) as

$$\eta_1\{D; (Y, G, E)\} = \exp(\gamma_{01}DY + \gamma_1DG + \gamma_2DE + \gamma_{12}DGE). \quad (5)$$

We can rewrite the logistic regression model as

$$p(D|Y, G, E) = \frac{\eta_1\{D; (Y, G, E)\} \exp(D\gamma_0)/\{1+\exp(\gamma_0)\}}{\sum_{D=0}^1 \eta_1\{D; (Y, G, E)\} \exp(D\gamma_0)/\{1+\exp(\gamma_0)\}}, \quad (6)$$

where $\exp(D\gamma_0)/\{1 + \exp(\gamma_0)\} = P(D|Y=0, G=0, E=0)$.

Let $S=1$ denote a subject being included in the case-control sample. We model the sampling probability as

$$P(S=1|D, Y, G, E) = \pi_1(D)\pi_2(Y)\pi_3(G, E), \quad (7)$$

which is a little more general than the conventional case-control design. For example, it accommodates pre-selection of subjects on the basis of (G, E) , which frequently happens in practice. However, using the case-prevalence rate in the general population as supplemental information in the retrospective likelihood approach under this general design [15] may lead to bias in the parameter estimation. The sampling design reduces to case-control ascertainment when $\pi_2(Y) = \pi_3(G, E) = 1$. Assume that $\pi_k, k=1, 2, 3$ are unknown. The sample ascertained by such a scheme has (D, Y) jointly distributed as

$$p(D, Y|G, E, S=1) = \frac{\pi_1(D)p(D|Y, G, E)\pi_2(Y)P(Y|G, E)}{\sum_D \sum_Y \pi_1(D)p(D|Y, G, E)\pi_2(Y)P(Y|G, E)} \quad (8)$$

Under the sampling design (7), Chen [20] showed that maximizing this prospective joint likelihood yields an association parameter estimator identical to that from maximizing the retrospective likelihood [15] on the basis of $p(Y, G, E|D)$. Plugging (3) and (6) into the right-hand side of expression (8), we obtain

$$p(D, Y|G, E, S=1) = \frac{\eta_1\{D; (Y, G, E)\}T(Y, G, E)\eta_2\{Y; (G, E)\}q_1(D)q_2(Y)}{\sum_D \sum_Y \eta_1\{D; (Y, G, E)\}T(Y, G, E)\eta_2\{Y; (G, E)\}q_1(D)q_2(Y)} \quad (9)$$

where $q_1(D) = p\{D|Y=0, G=0, E=0, S=1\}$, $q_2(Y) = p\{Y|D=0, G=0, E=0, S=1\}$, and

$$T(Y, G, E) = \{1 + \exp(\gamma_0 + \gamma_{01}Y + \gamma_1G + \gamma_2E + \gamma_{12}GE)\}^{-1}, \quad (10)$$

Directly maximizing the likelihood on the basis of (9) is computationally difficult because it involves the intercept γ_0 , which is weakly identifiable [20]. In Appendix A, we show that $T(Y, G, E)$ in (9) can be replaced by the odds ratio function

$$\eta_3\{Y; (G, E)\} = \exp\{-\tau^2\gamma_{01}(\gamma_1YG + \gamma_2YE + \gamma_{12}YGE)\}, \quad (11)$$

where $(\gamma_{01}, \gamma_1, \gamma_2, \gamma_{12})$ are the odds ratio parameters in (5) and $\tau^2 = r_0(1 - r_0)$ with $r_0 = \{1 + \exp(-\gamma_0)\}^{-1}$ being the baseline prevalence rate of the primary phenotype. We give more general formulas in Appendix A for non-zero reference point for (D, Y, G, E) and for a quantitative primary and/or secondary trait.

2.2. Correction formula for the conditional analysis

When we analyze the secondary trait conditional on the primary phenotype, the distribution of the secondary trait from (9) approximately has the form

$$p(Y|D, G, E, S=1) = \frac{\eta_1(D; Y|G, E)\eta_3\{Y; (G, E)\}\eta_2\{Y; (G, E)\}q_2(Y)}{\sum_Y \eta_1(D; Y|G, E)\eta_3\{Y; (G, E)\}\eta_2\{Y; (G, E)\}q_2(Y)}, \quad (12)$$

which we can rewrite as

$$p(Y|D, G, E, S=1) = \frac{\exp\{Y(\theta_0^* + \gamma_{01}D + \psi_1G + \psi_2E + \psi_{12}GE)\}}{1 + \exp(\theta_0^* + \gamma_{01}D + \psi_1G + \psi_2E + \psi_{12}GE)}, \quad (13)$$

where $\theta_0^* = \theta_0 + \log\pi_2(1)/\pi_2(0)$, and

$$\begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_{12} \end{pmatrix} \approx \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_{12} \end{pmatrix} - \tau^2\gamma_{01} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_{12} \end{pmatrix}. \quad (14)$$

From (14), we can obtain a correction formula for the conditional analysis as

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_{12} \end{pmatrix} \approx \begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_{12} \end{pmatrix} + \tau^2 \gamma_{01} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_{12} \end{pmatrix}. \quad (15)$$

This approximation suggests that the magnitude of bias in the conditional analysis is influenced by the baseline variance of the primary phenotype (determined by the baseline prevalence rate), the strength of association between the primary and the secondary traits, and the strength of association between the primary phenotype and the genetic marker and the environmental factor. When the primary phenotype is rare in the study population ($\tau^2 \approx 0$) or when no association exists between the primary phenotype and the secondary trait ($\gamma_{01} = 0$) given the genetic marker and the environmental factor, or when no association exists between the primary phenotype and the genetic marker ($(\gamma_1, \gamma_{12}) = (0, 0)$) controlled for the secondary trait and the environment factor, the conditional analysis yields approximately correct estimates of association between the secondary trait and the genetic marker.

We can carry out the correction in the following way. First, we run a logistic regression analysis of Y given D , G , and E to obtain estimates of $(\psi_1, \psi_2, \psi_{12})$. We then run a supplemental logistic regression of D given Y , G , and E to obtain estimates of $(\gamma_{01}, \gamma_1, \gamma_2, \gamma_{12})$. We can easily compute the corrected estimates of $(\theta_1, \theta_2, \theta_{12})$ using (15) when τ^2 is known. When τ^2 is unknown or is known inaccurately, a plausible range of τ^2 is first chosen. We can then compute the estimates of $(\theta_1, \theta_2, \theta_{12})$ at different τ^2 values in the plausible range. We can obtain a sensitivity plot of $(\theta_1, \theta_2, \theta_{12})$ estimates against the plausible range of τ^2 . We show in Appendix A that the correction formula (15) is also applicable when either or both the primary phenotype and the secondary trait are quantitative.

Although the correction to the conditional analysis is easy to obtain, obtaining the variance estimate can be more involved. We can combine the two logistic regressions into a single joint analysis based on (9) with $T(Y, G, E)$ replaced by (11) so that we can obtain the variance by routine calculation in a maximum likelihood approach. The joint analysis yields estimates of $(\psi_1, \psi_2, \psi_{12})$ and $(\gamma_{01}, \gamma_1, \gamma_2, \gamma_{12})$ simultaneously, and we can then obtain the association parameter $(\theta_1, \theta_2, \theta_{12})$ by the correction formula (15). We give the detailed variance estimate for the parameter estimator in Appendix A.

2.3. Testing no association between the secondary trait and a genetic marker

Let $\theta(\tau)$ denote θ for a given τ through the formula (15). Testing no association between the secondary trait and a genetic marker corresponds to test $H_0: \theta_1(\tau_0) = \theta_{12}(\tau_0) = 0$, where τ_0 is fixed at the true value. For the case with known τ_0 , we can perform a χ^2 test statistic with two degrees of freedom:

$$Q(\tau_0) = n \left\{ \hat{\theta}(\tau_0) - \theta(\tau_0) \right\}^T c^T \{ c \hat{V}(\tau_0) c^T \}^{-1} c \{ \hat{\theta}(\tau_0) - \theta(\tau_0) \} \sim \chi^2_2$$

where $\hat{V}(\tau_0)$ is the variance estimate for $\hat{\theta}(\tau_0)$ and

$$c = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

For the case with unknown or inaccurately known τ_0 , we can perform a sensitivity analysis on the basis of the χ^2 statistic. Alternatively, we can perform a test based on the minimal statistic in the following way. Assume that τ_0 is in $[\tau_a, \tau_b]$. For example, if the primary phenotype is binary, then $r_0 \in (0, 1/2]$ and $[\tau_a, \tau_b] \in (0, 1/4]$. We set the nominal p -value as

where χ_{df}^2 denotes the chi-square distribution with degrees of freedom df . Because the chi-square distribution is a monotone function, we can calculate the p -value as . The actual type I error α for a nominal type I error p for individual tests is

This approach to controlling type I error can be too conservative when the range of $[\tau_a, \tau_b]$ is large. In this case, we may choose a prior distribution for τ_0 and apply an alternative approach as follows. Suppose that τ has a normal distribution with mean τ_0 and variance ω^2 . It then follows that

$$\sqrt{n} \{ \hat{\theta}(\tau_0) - \theta(\tau_0) \} \rightarrow N\{0, V(\tau_0) + v_1\},$$

where $v_1 = n\omega^2\gamma_{01}^2$. We can carry out a test for no association between the genetic marker and the secondary trait by using

$$n \{ \hat{\theta}(\tau_0) - \theta(\tau_0) \}^T c^T [c \{ \hat{V}(\tau_0) + \hat{v}_1 \} c^T]^{-1} c \{ \hat{\theta}(\tau_0) - \theta(\tau_0) \} \sim \chi_2^2$$

3. Other correction approaches

3.1. Correction to the marginal analysis

When the marginal analysis for the secondary trait is used, it follows from (9) that the correct model is

$$p(Y|G, E, S=1) = \frac{W(Y, G, E) \exp(\theta_0^* + \theta_1 YG + \theta_2 YE + \theta_{12} YGE)}{\sum_Y W(Y, G, E) \exp(\theta_0^* + \theta_1 YG + \theta_2 YE + \theta_{12} YGE)}, \quad (16)$$

where

$$W(Y, G, E) = \frac{1 + \exp(\gamma_0^* + \gamma_{01}Y + \gamma_1G + \gamma_2E + \gamma_{12}GE)}{1 + \exp(\gamma_0 + \gamma_{01}Y + \gamma_1G + \gamma_2E + \gamma_{12}GE)}, \quad (17)$$

and $\gamma_0^* = \gamma_0 + \log\{\pi_1(1)/\pi_1(0)\}$. By a similar method that was used in obtaining the replacement $\eta_3\{Y; (G, E)\}$ for $T(Y, G, E)$, we can obtain a replacement for $W(Y, G, E)$ in (20) as

$$\eta_3^*\{Y;(G, E)\}=\exp\{(\tau^{*2}-\tau^2)\gamma_{01}(\gamma_1YG+\gamma_2YE+\gamma_{12}YGE)\}, \quad (18)$$

where $\tau^{*2}=r^*(1-r^*)$, the baseline variance of the sample, and $r^*=\{1+\exp(-\gamma_0^*)\}^{-1}$. We can estimate parameters ($\gamma_0^*, \gamma_1, \gamma_2, \gamma_{12}$) by fitting a logistic regression model to the case-control sample with γ_0^* being the intercept of the regression model. An explicit correction formula for the association of the genetic marker and the secondary trait appears as

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_{12} \end{pmatrix} \approx \begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_{12} \end{pmatrix} + (\tau^2 - \tau^{*2})\gamma_{01} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_{12} \end{pmatrix}. \quad (19)$$

We give the variance estimate of the corrected estimator in Appendix A.

3.2. Correction approaches using compensators

We can rewrite expression (16) as

$$p(Y|G, E, S=1) = \frac{\exp\{Y(\theta_0^*(G, E) + \theta_1G + \theta_2E + \theta_{12}GE)\}}{1 + \exp(\theta_0^*(G, E) + \theta_1G + \theta_2E + \theta_{12}GE)} \quad (20)$$

where

$$\theta_0^*(G, E) = \theta_0^* + \log \frac{W(1, G, E)}{W(0, G, E)} \quad (21)$$

If we know $W(Y, G, E)$, (20) and (21) suggest an alternative approach to obtain an estimator of θ through a logistic regression with a known compensator $\log \{W(1, G, E)/W(0, G, E)\}$.

We can estimate the parameters ($\gamma_0^*, \gamma_1, \gamma_2, \gamma_{12}$) in the compensator by fitting a prospective logistic regression model to the case-control sample. When parameter γ_0 is known, we can straightforwardly compute the compensator. We can then fit a logistic regression model based on (20) with a known compensator to the data to obtain the estimator of θ directly. When γ_0 is unknown or known inaccurately, we can perform a sensitivity analysis of the association parameter of the secondary trait with the genetic marker to the assumed γ_0 values based on (20) and (21).

We can carry out a similar compensator approach by the conditional analysis based on

$$p(Y|D, G, E, S=1) = \frac{\exp\{Y(\theta_0^{**}(G, E) + \gamma_{01}D + \theta_1G + \theta_2E + \theta_{12}GE)\}}{1 + \exp(\theta_0^{**}(G, E) + \gamma_{01}D + \theta_1G + \theta_2E + \theta_{12}GE)}, \quad (22)$$

where

$$\theta_0^{**}(G, E) = \theta_0^* + \log \frac{T(1, G, E)}{T(0, G, E)} \quad (23)$$

We can estimate parameters ($\gamma_{01}, \gamma_1, \gamma_2, \gamma_{12}$) in the compensator $\log \{T(1, G, E)/T(0, G, E)\}$ by fitting a logistic regression model to the case-control sample. When the other parameter γ_0 is known, we can straightforwardly compute the compensator. When γ_0 is

unknown or known inaccurately, we can perform a sensitivity analysis of the association parameters to the assumed γ_0 values. Likewise, we can similarly perform a joint analysis with compensator.

Methods based on compensators along with the maximum joint prospective likelihood approach based on (9) with γ_0 fixed can yield more accurate estimators. In terms of sensitivity analysis, however, they require much more computation, and the inference can be more difficult to carry out because of lack of explicit functional relationship between the association parameters and the baseline prevalence rate for the primary phenotype. As a result, we may use the compensator approaches and/or the maximum likelihood approach jointly with the approximation approach as follows. First, we may use the approximation approach to screen a large number of genetic markers in a genome-wide association study. We may then further study a relatively small number of identified markers by the more accurate compensator approaches or the maximum likelihood approach.

4. Simulation studies

In the first simulation study, we evaluate the impact of different choices of the reference point (Appendix A) in the approximation approaches. We simulated a binary primary phenotype and a binary secondary trait. The model for generating the primary phenotype is (4), and the model for generating the secondary trait is (1). Neither includes the environment factor E . We simulate genetic markers as single-nucleotide polymorphisms (SNPs) following the Hardy–Weinberg equilibrium with minor allele frequencies (MAF) 20%. However, we do not make this assumption in the analysis of the simulated data. In model (4), we assume the effect of the secondary phenotype on the first phenotype to be relatively large ($\gamma_{01} = 1$) because it enables us to assess the performance of the explicit correction formulas. The genetic effect is multiplicative, and the effect size is set to $\gamma_1 = 0.5$ and $\theta_1 = 0.5$. We simulated the baseline prevalence rate ranging from $r_0 = 0.001$ to 0.5 with an increment of 0.01. For each r value, we simulated 5000 repetitions of a sample size 1200 with 600 cases and 600 controls.

We computed three sets of corrections. In the first set, $(Y_0, G_0) = (0, 0)$. In the second set, (Y_0, G_0) is the sample mean. In the third set, $(Y_0, G_0) = (0.5, 1)$, which is the middle point of the ranges of the variables. We display the simulation results in Figure 1. We see from Figure 1a that the conditional analysis has little bias when the baseline prevalence rate r is very small. But the bias increases very rapidly when $r = 1 \sim 20\%$. The bias remains at a similar magnitude for $r = 20 \sim 50\%$. For the marginal analysis, the bias is very large when r is very small. But the bias decreases to nearly zero as r decreases from 1% to 20% and mostly stays at the same level for $r = 20 \sim 50\%$. Figure 1b–d respectively shows the corrections with $(Y_0, G_0) = (0, 0)$, (Y_0, G_0) being the sample mean and $(Y_0, G_0) = (0.5, 1)$ the middle point of the ranges of the variables. We can see that bias is reduced in all three choices. However, the best choice of the reference point appears to be the middle point of the ranges of the variables. We repeated the simulation with MAF being 30% and 10% and with different genetic effects; the relative merit (results not shown) remains the same.

In the second simulation study, we used the same data-generation models with the inclusion to both models of an environment variable and the gene–environment interaction. The simulated environment factor is continuous and uniformly distributed on $(-2, 2)$. We fixed the log-odds ratios for genetic, environmental, and gene–environment interaction in the primary phenotype model at $(0.5, 0.5, 0.3)$. We fixed the log-odds ratio for genetic, environmental, and gene–environment interaction in the secondary trait model at $(0.5, -0.5, -0.3)$. We fixed the baseline prevalence rate of the secondary trait at approximately 10%. The simulated baseline prevalence rates for the primary phenotype are at 1%, 5%, 10%, and

15%, which are in the range of rates mostly encountered in practice. We use several methods to analyze the simulated data. They include the analysis using controls only; the marginal and conditional analyses ignoring the biased sampling design; the corrected marginal, conditional, and joint analysis using the explicit correction formulas; and the marginal, the conditional, and the joint correction estimators using the compensators. The joint correction estimator refers to the maximization of likelihood based on (9) with γ_0 in T fixed at the truth. We take the reference point used in the correction analysis at the average value of each variable. Note that the analysis under the rare case-prevalence assumption is the same, as the conditional analysis ignores the biased sampling design.

We list the simulation results based on 5000 repetitions of a sample size 1200 with 600 cases and 600 controls in Table I. Because the conditional analyses yield results very close to those of the joint analysis, we do not include results on the conditional analyses in Table I. The analyses presented in Table I include the analysis based on controls only (Controls), marginal analysis (Marginal), conditional analysis based on the odds ratio model without correction (Condit.), corrected joint analysis (Correct.), the compensator approach (Comp.) based on the conditional analysis, and the analysis using the maximum prospective joint likelihood (MLE). The compensator approach based on the marginal analysis yielded very close results as the Comp. and is thus suppressed. We see from the results in Table I that the controls-only approach can be substantially biased when the cases are not very rare in the population, and the odds ratio estimator has large variations in general due to the reduced sample size. MLE with γ_0 fixed at the true value has the best performance. Comp. with γ_0 fixed at the true value is almost indistinguishable from the MLE. Correct. has a little larger bias than either Comp. or MLE. The variance of Comp. or MLE is the same as or slightly smaller than the variance of the Correct. The other estimators listed in the table have clearly worse performance in at least some of the simulated cases. When computation complexity is not an issue, Comp. or MLE appears to be the favorite. When we consider computation complexity in conjunction with performance, Correct. is favored.

We also evaluate the size and the power of the proposed tests on the basis of the corrected joint analysis of the approximate likelihood. We calculate the empirical sizes on the basis of the analysis of the simulated data with $(\gamma_1, \gamma_2, \gamma_{12}) = (0, -0.5, 0)$, that is, no genetic effect but with environmental effect. We list results on tests in Table II. In each scenario, we perform five tests, and we list the estimated size/power for the five tests for the nominal type I error rates: 5% and 1%. The estimated type I errors are close to the nominal type I error. The estimated size and power for different tests suggested that for a large range of the baseline prevalence rate, the size and power of the different tests do not differ much. This may suggest that the tests that control the type I error for a range of baseline prevalence rate may have reasonably good performance. We also varied MAF and the genetic and gene–environment interaction effects in the simulation; similar results (not shown) were obtained.

5. Genetic association of the prostate specific antigen

Prostate specific antigen (PSA) level in blood is a reasonably good predictor for prostate cancer, although it is not a precise marker for prostate cancer [21]. In genetic association studies of prostate cancer, PSA is often measured. It is of substantial interest to investigate the association of PSA level with genetic variants. For example, in performing the genome-wide association studies of prostate cancer through combining case–control studies that included thousands of prostate cancer cases and health controls, Eeles *et al.* [22, 23] performed secondary analyses on PSA level in blood on the basis of the control sample in an attempt to identify SNPs that may be associated with PSA. The control-only analysis can have substantially lower power in detecting possible association than the analyses that include both cases and controls. In addition, such an analysis that requires sparse disease in

the population can be biased because the prevalence rate of prostate cancer in the elder male population can be high [21].

The proposed approach in this article can be more appropriate for the analysis of the PSA data in [22, 23]. Here, we use a similar case-control study of prostate cancer in the African American population to illustrate the use of our approach. The sample in this study was collected and genotyped in one of the coauthors' (Dr. Kittles) lab. There are 260 cases and 345 controls whose genomes were typed on over one million SNPs using Illumina array. PSA level was measured on 218 cases and 325 controls. In this analysis, we concentrate on the 543 subjects with PSA measured. In a preliminary analysis of the prostate cancer status, we fit a logistic regression model without any SNPs in the model. We found that the logarithm transformed PSA, that is, $\log(1 + \text{PSA})$, age of the subject, and the interaction between the two of a subject are all strongly associated with the case-control status of the subject. In a preliminary analysis of the PSA, we fit a linear regression model for the logarithm transformed PSA. We found that the case-control status and the age of the subject are strongly associated with the logarithm transformed PSA, but there was no age and status interaction. These models form the basis of our genome-wide association analysis of the PSA as a secondary trait. Our model for the primary phenotype, the prostate cancer status, is

$$\log[\eta_1\{D;(\text{lpsa}, G, \text{age})\}] = D * (\gamma_{01}\text{lpsa} + \gamma_{02}\text{lpsa} * \text{age} + \gamma_1 G + \gamma_2 \text{age}),$$

where D is the prostate cancer status and $\text{lpsa} = \log(1 + \text{PSA})$. Our model for the secondary trait is

$$\log[\eta_2\{\text{lpsa};(G, \text{age})\}] = \text{lpsa} * (\theta_1 G + \theta_2 \text{age}).$$

The correction formula for the genetic effect remains the same. See Appendix B for more details.

The number of the typed SNPs is over one million. The sample size is relatively small for detecting possible small genetic effects hidden in the large number of SNPs. We randomly sampled 10% of the typed SNPs in the 22 autosomes for the purpose of illustration. After excluding those SNPs with a MAF less than 5% and/or with a p -value less than 0.0001 in Hardy-Weinberg equilibrium test, a total of 90,232 SNPs were selected. We performed analyses examining baseline case-prevalence rates ranges from 0% to 10% with an evaluation interval 0.2%. We chose the fixed point as $(\text{age}, G, \text{lpsa}) = (65, 1, 4)$, the baseline case-prevalence rate should be interpreted as for this subpopulation. In each case, we applied the genomic control approach in [24] to check for possible population stratification. The estimated variance inflation factors range from $\lambda = 0.964$ to $\lambda = 0.970$. The adjustment for population stratification appears to be not imperative. We displayed the top SNPs in the p -value spectrum to show the impact of different baseline case-prevalence rates on the test results.

Table III displays the test results of genetic association whose smallest p -value in the baseline rate range of 0% to 10% is smaller than 0.0001 for the secondary trait PSA. Test 1 shows the results under the sparse disease assumption, an approach routinely used for the analysis of secondary traits. Test 2 and Test 3 are computed under the 5% and 10% baseline case-prevalence rates, respectively. We can see that the impact of a non-zero case-prevalence rate can range from relatively small (for example, rs10197545) to very large (for example, rs9316673). The impact is unevenly distributed over different case-prevalence

rates because the correction not only affects the odds ratio estimation but also the variance estimates of the odds ratio estimator. The test (MinT) controlling the type I error for the baseline rate ranging from 0% to 10% is a more robust test for the secondary trait analysis when the sparse disease assumption is suspected. MinT can yield the same results as Test 1 in some cases or different results from Test 1 in other cases. Even when MinT yields the same results as Test 1, the analysis is still useful in that it provides information on the reliability of the results. For example, the sensitivity analysis indicates that the significant SNP tested under sparse disease assumption, rs2255684, is only moderately affected by the non-zero baseline case-prevalence rates. This analysis ensures us the reliability of the test result even if the sparse disease assumption does not hold. On the other hand, the test based on sparse disease assumption for rs9316673 can be much less reliable. We can easily perform such a sensitivity analysis using our approach, whereas it can be very time-consuming to perform using other approaches. We also performed a prospective maximum likelihood analysis of the selected SNPs; the results confirmed the effectiveness of the correction approach.

6. Discussion

For the analysis of secondary traits, the retrospective likelihood approach proposed in Lin and Zeng [15] included three scenarios. The first assumes a known population case-prevalence rate. The second assumes rare disease that corresponds to a disease prevalence rate in the general population close to zero. The third scenario does not assume any known supplemental information on the disease prevalence rate in the general population. The third scenario is the most important one, whereas the first two may be regarded as special cases of the third one. In the first two scenarios, the retrospective likelihood approach has good performance when the population level case-prevalence rate is accurate and usable. In the third scenario, the retrospective likelihood approach can perform poorly. This is primarily due to the presence of weakly identifiable parameters [20] that cannot be estimated well with practical attainable sample sizes.

The approach we proposed mainly tackles the problem in scenario 3. We consider an approximate model that does not involve estimating the weakly identifiable parameter. In carrying out the task, we work with the simpler joint prospective likelihood, which was shown [20] to be equivalent to the retrospective likelihood in scenario 3. Following the estimation, we recover the parameters of interest by a sensitivity analysis to the baseline case-prevalence rate. Lin and Zeng [15] also suggested a sensitivity analysis to the population case-prevalence rate in scenario 3. However, the validity of their approach under scenarios 1 and 3 rests on the assumption that the covariate distribution in the general population can be identified from the case-control sample, which is violated when pre-selection of subjects through eligible criteria, routinely performed in biomedical studies, is involved. Our approach is not subject to this limitation. Furthermore, the sensitivity analysis using explicit correction formulas we proposed is computationally much faster, which can be critical for performing genome-wide analysis. The explicit correction formulas also suggest that the magnitude of bias in the conditional or marginal analysis is determined by three factors: the baseline prevalence rate of the primary phenotype, the strength of association between the primary phenotype and the secondary trait, and the strength of association between the primary phenotype and the genetic marker. Such dependence has largely been qualified but not quantified previously in the literature. Of note is that correction of the bias does not require us to know the prevalence rate of the secondary trait in the general population. Our approach can also be applied to scenarios 1 and 2 when the baseline case-prevalence rate is used instead.

Our proposed approaches are directly applicable to the secondary trait analysis even if either or both of the primary and secondary traits are quantitative and the sample is ascertained through an extreme-value sampling design rather than a case-control design. Both the environment factor and the gene-environment interaction can be readily included in the model for analysis. In particular, when the primary phenotype is quantitative and a normal model is used to link the phenotype to the genotype and the environmental factor, the correction formula for the joint analysis becomes exact. Our analysis also suggests that when the case-prevalence rate for the primary phenotype is low, the uncorrected conditional analysis is much better than the uncorrected marginal analysis.

In deriving the correction formulas in Appendix A, we have assumed that there is no interaction between the secondary trait and the genotype or the environmental factors. If such interactions exist, the idea we used to obtain the correction formulas can still be applied. However, the correction formulas need to be modified accordingly. See Appendix B. We chose to mainly deal with the simple situation in this article to keep the presentation easy to follow.

Acknowledgments

This research was supported by NSF grant DMS 1007726 (to H. Y. C.). We thank two anonymous reviewers and an associate editor for very helpful comments on the previous versions of this paper.

References

1. Anderson JA. Separate sample logistic discrimination. *Biometrika*. 1972; 59:19–35.
2. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrics*. 1976; 66:403–411.
3. Breslow, NE.; Day, N. Statistical methods in cancer research. Volume I: The analysis of case-control studies. IARC Scientific Publications; Lyon: IARC; 1980.
4. Breslow NE. Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*. 1996; 91:14–28. [PubMed: 12155399]
5. Rabinowitz D. A note on efficient estimation from case-control data. *Biometrika*. 1997; 84:486–488.
6. Roeder K, Carroll RJ, Lindsay BC. A nonparametric mixture approach to case-control studies with errors in variables. *Journal of the American Statistical Association*. 1996; 91:722–732.
7. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. *Biometrika*. 1997; 84:57–71.
8. Chen HY. A note on prospective analysis of outcome-dependent samples. *Journal of Royal Statistical Society, Ser B*. 2003; 65:575–584.
9. Nagelkerke NJD, Moses S, Plummer FA, Brunham RC, Fish D. Logistic regression in case-control studies: the effect of using independent as dependent variables. *Statistics in Medicine*. 1995; 14:769–755. [PubMed: 7644857]
10. Lee AJ, McMurchy L, Scott AJ. Re-using data from case-control studies. *Statistics in Medicine*. 1997; 16:1377–1389. [PubMed: 9232759]
11. Jiang Y, Scott AJ, Wild CJ. Secondary analysis of case-control data. *Statistics in Medicine*. 2006; 25:1323–1339. [PubMed: 16220494]
12. Reilly M, Torrang A, Klint A. Reuse of case-control data for analysis of new outcome variables. *Statistics in Medicine*. 2005; 24:4009–4019. [PubMed: 16320270]
13. Richardson DB, Rzehak P, Klenk J, Weiland SK. Analysis of case-control data for additional outcomes. *Epidemiology*. 2007; 18:441–445. [PubMed: 17473707]
14. Kraft P. Analysis of genome-wide association scans for additional outcomes. *Epidemiology*. 2007; 18:838. [PubMed: 18049198]

15. Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*. 2009; 33:256–265. [PubMed: 19051285]
16. Monsees GM, Tamimi RM, Kraft P. Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology*. 2009; 33:718–728.
17. Li H, Gail M, Berndt S, Chatterjee N. Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology*. 2010; 34:427–433. [PubMed: 20583284]
18. Wang J, Shete S. Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genetic Epidemiology*. 2011; 35:190–200. [PubMed: 21308766]
19. Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association*. 2006; 101:89–104.
20. Chen HY. A unified framework for parameter identifiability and estimation in biased sampling designs. *Biometrika*. 2011; 98:163–175.
21. Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, Minasian LM, Ford LG, Lippman SM, Crawford ED, Crowley JJ, Coltman CA. Prevalence of prostate cancer among men with a prostate-specific antigen level ≤ 4.0 ng per milliliter. *The New England Journal of Medicine*. 2004; 350:2239–2246. [PubMed: 15163773]
22. Eeles RA, Kote-Jarai Z, Giles GG, Olama AAA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, Field HI, Southey MC, Severi G, Donovan JL, Hamdy FC, Dearnaley DP, Muir KR, Smith C, Bagnato M, Arden-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Cox A, Lewis S, Brown PM, Jhavar SG, Tymrakiewicz M, Lophatananon A, Bryant SL, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Fisher C, Jamieson C, Cooper CS, English DR, Hopper JL, Neal DE, Easton DF. Multiple newly identified loci associated with prostate cancer susceptibility. *Nature Genetics*. 2008; 40:316–321. [PubMed: 18264097]
23. Eeles RA, Kote-Jarai Z, Olama AAA, Giles GG, Guy M, Severi G, Muir K, Hopper JL, Henderson BE, Haiman CA, Schleutker J, Hamdy FC, Neal DE, Donovan JL, Stanford JL, Ostrander EA, Ingles SA, John EM, Thibodeau SN, Schaid D, Park JY, Spurdle A, Clements J, Dickinson JL, Maier C, Vogel W, Dork T, Rebbeck TR, Cooney KA, Cannon-Albright L, Chappuis PO, Hutter P, Zeegers M, Kaneva R, Zhang HW, Lu YJ, Foulkes WD, English DR, Leongamornlert DA, Tymrakiewicz M, Morrison J, Arden-Jones AT, Hall AL, O'Brien LT, Wilkinson RA, Saunders EJ, Page EC, Sawyer EJ, Edwards SM, Dearnaley DP, Horwich A, Huddart RA, Khoo VS, Parker CC, Van As N, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Cooper CS, Southey MC, Lophatananon A, Liu JF, Kolonel LN, Le Marchand L, Wahlfors T, Tammela TL, Auvinen A, Lewis SJ, Cox A, FitzGerald LM, Koopmeiners JS, Karyadi DM, Kwon EM, Stern MC, Corral R, Joshi AD, Shahabi A, McDonnell SK, Sellers TA, Pow-Sang J, Chambers S, Aitken J, Gardiner RA, Batra J, Kedda MA, Lose F, Polanowski A, Patterson B, Serth J, Meyer A, Luedeke M, Stefflova K, Ray AM, Lange EM, Farnham J, Khan H, Slavov C, Mitkova A, Cao GW, Easton DF. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nature Genetics*. 2009; 41:1116–1121. [PubMed: 19767753]
24. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999; 55:997–1004. [PubMed: 11315092]
25. Chen HY. A semiparametric odds ratio model for measuring association. *Biometrics*. 2007; 63:413–421. [PubMed: 17688494]
26. Chen HY. Compatibility of conditionally specified models. *Statistics & Probability Letters*. 2010; 80:670–677. [PubMed: 20436935]

Appendix A. Derivation of the formulas

We derive general correction formulas for the marginal, conditional, and joint analyses for both discrete and continuous (D , Y) with an arbitrary reference point (D_0 , Y_0 , G_0 , E_0). If D is a binary phenotype, we assume it follows model (4). If D is a quantitative trait, we model it by

$$D = \beta_0 + \beta_{01}Y + \beta_1G + \beta_2G + \beta_2E + \beta_{12}GE + \varepsilon, \quad (\text{A.1})$$

where $\varepsilon \sim N(0, \tau^2)$. Note that the following derivation can also be modified to include interactions between Y and other variables. To keep the presentation simple, we only deal with the case without Y interaction. Define an odds ratio function for D versus (Y, G, E) relative to the reference point as

$$\eta_1\{D; (Y, G, E)\} = \frac{p(D|Y, G, E)p(D_0|Y_0, G_0, E_0)}{p(D_0|Y, G, E)p(D|Y_0, G_0, E_0)}. \quad (\text{A.2})$$

The odds ratio function for model (A.1) is

$$\log \eta_1\{D; (Y, G, E)\} = \gamma_{01}(D - D_0)(Y - Y_0) + \gamma_1(D - D_0)(G - G_0) + \gamma_2(D - D_0)(E - E_0) + \gamma_{12}(D - D_0)(GE - G_0E_0),$$

where $(\gamma_{01}, \gamma_1, \gamma_2, \gamma_{12}) = (\beta_{01}, \beta_1, \beta_2, \beta_{12})/\tau^2$. The density of D under model (A.1) Can be rewritten as

$$p(D|Y, G, E) = \frac{\eta_1\{D; (Y, G, E)\}p\{D|Y_0, G_0, E_0\}}{\int \eta_1\{D; (Y, G, E)\}p\{D|Y_0, G_0, E_0\}dD}. \quad (\text{A.3})$$

This odds ratio representation also works for binary phenotype D [25, 26]. Expression (6) is the special case with the reference point $(D_0, Y_0, G_0, E_0) = (0, 0, 0, 0)$. If Y is a quantitative trait and is modeled by

$$Y = \alpha_0 + \alpha_1G + \alpha_2E + \alpha_{12}GE + e, \quad (\text{A.4})$$

where $e \sim N(0, \sigma^2)$, the following odds ratio representation

$$p(Y|G, E) = \frac{\eta_2\{Y; (G, E)\}p\{Y|G_0, E_0\}}{\sum_Y \eta_2\{Y; (G, E)\}p\{Y|G_0, E_0\}} \quad (\text{A.5})$$

holds, where

$$\eta_2\{Y; (G, E)\} = \frac{p(Y|G, E)p(Y_0|G_0, E_0)}{p(Y_0|G, E)p(Y|G_0, E_0)}. \quad (\text{A.6})$$

For the quantitative trait modeled by (A.4),

$$\log \eta_2\{Y; (G, E)\} = \theta_1(Y - Y_0)(G - G_0) + \theta_2(Y - Y_0)(E - E_0) + \theta_{12}(Y - Y_0)(GE - G_0E_0),$$

where $(\theta_1, \theta_2, \theta_{12}) = (\alpha_1, \alpha_2, \alpha_{12})/\sigma^2$. The representation (A.5) also works for binary phenotype Y . Expression (3) is the special case with the reference point $(Y_0, G_0, E_0) = (0, 0, 0)$.

When the sampling design is (7),

$$p(D, Y|G, E, S=1) = \frac{P(S=1|D, Y, G, E)p(D|Y, G, E)p(Y|G, E)}{\sum_D \sum_Y P(S=1|D, Y, G, E)p(D|Y, G, E)p(Y|G, E)}. \quad (\text{A.7})$$

By replacing $P(S=1|D, Y, G, E)$ with (7), $p(D|Y, G, E)$ with (A.3), and $p(Y|G, E)$ with (A.5), we can rewrite (A.7) as

$$p(D, Y|G, E, S=1) = \frac{\eta_1\{D;(Y, G, E)\}T(Y, G, E)\eta_2\{Y;(G, E)\}p(Y|D_0, G_0, E_0)p\{D|Y_0, G_0, E_0, S=1\}}{\sum_D \sum_Y \eta_1\{D;(Y, G, E)\}T(Y, G, E)\eta_2\{Y;(G, E)\}p(Y|D_0, G_0, E_0)p\{D|Y_0, G_0, E_0, S=1\}} \quad (\text{A.8})$$

Let

$$\eta_T\{Y;(G, E)\} = \frac{T(Y, G, E)T(Y_0, G_0, E_0)}{T(Y_0, G, E)T(Y, G_0, E_0)}.$$

When D is a quantitative trait and is modeled by (A.1),

$$\log \eta_T\{Y;(G, E)\} = -\tau^2 \gamma_{01}(Y - Y_0)\{\gamma_1(G - G_0) + \gamma_2(E - E_0) + \gamma_{12}(GE - G_0E_0)\}$$

by direct calculation. When D is binary phenotype and is modeled by (4),

$$\eta_T\{Y;(G, E)\} = \frac{[1 + \exp\{\gamma_0 + m(Y_0, G, E)\}][1 + \exp\{\gamma_0 + m(Y, G_0, E_0)\}]}{[1 + \exp\{\gamma_0 + m(Y, G, E)\}][1 + \exp\{\gamma_0 + m(Y_0, G_0, E_0)\}]},$$

where $m(y, G, E) = \gamma_{01}y + \gamma_1G + \gamma_2E + \gamma_{12}GE$. From the second order Taylor expansion,

$$\log(1 + e^t) \approx \log(1 + e^{t_0}) + \frac{e^{t_0}}{1 + e^{t_0}}(t - t_0) + \frac{e^{t_0}}{2(1 + e^{t_0})^2}(t - t_0)^2$$

it follows by expanding $\log[1 + \exp\{\gamma_0 + m(y, G, E)\}]$ around $t_0 = \gamma_0 + m(y_0, G_0, E_0)$ that

$$\log \eta_T\{Y;(G, E)\} \approx -\tau^2 \gamma_{01}(Y - Y_0)\{\gamma_1(G - G_0) + \gamma_2(E - E_0) + \gamma_{12}(GE - G_0E_0)\},$$

where $\tau^2 = r_0(1 - r_0)$, which is the conditional variance of D given $Y = Y_0$, $G = G_0$, and $E = E_0$, and $r_0 = e^{t_0} / (1 + e^{t_0})$, which is the baseline case-prevalence rate for the primary phenotype. Note that τ^2 , which was used to denote the residual variance in the regression model for the primary trait, can also be interpreted as the conditional variance of D given $Y = Y_0$, $G = G_0$, and $E = E_0$. If the third order Taylor expansion is used, that is,

$$\log(1+e^t) \approx \log(1+e^{t_0}) + \frac{e^{t_0}}{1+e^{t_0}}(t-t_0) + \frac{e^{t_0}}{2(1+e^{t_0})^2}(t-t_0)^2 + \frac{e^{t_0}(1-2e^{t_0})}{6(1+e^{t_0})^3}(t-t_0)^3,$$

the approximation yields

$$\log \eta_T\{Y; (G, E)\} \approx -\tau^2 \gamma_{01}(Y-Y_0)\{\gamma_1(G-G_0) + \gamma_2(E-E_0) + \gamma_{12}(GE-G_0E_0)\} - c_3 \gamma_{01} \gamma_1 \gamma_2 (Y-Y_0)(G-G_0)(E-E_0)$$

when we exclude higher order terms from the approximation, where $c_3 = r_0(1-r_0)(1-2r_0)$. We can obtain the approximation with interactions between Y and (G, E) by changing $m(y, G, E)$ in the derivation accordingly.

When we replace $T(Y, G, E)$ by approximation to $\eta_T\{Y; (G, E)\}$, we can rewrite (A.8) as

$$p(D, Y|G, E, S=1) = \frac{\xi(D, Y, G, E)q_1(D)q_2(Y)}{\sum_D \xi(D, Y, G, E)q_1(D)q_2(Y)} \quad (\text{A.9})$$

where $q_1(D) = p(D|Y_0, G_0, E_0, S=1)$,

$$q_2(Y) = \frac{T(Y, G_0, E_0)p(Y|G_0, E_0)}{\sum_Y T(Y, G_0, E_0)p(Y|G_0, E_0)},$$

$$\begin{aligned} \log \xi(D, Y, G, E) = & \gamma_{01}(D-D_0)(Y-Y_0) + \gamma_1(Y-Y_0)(G-G_0) + \gamma_2(Y-Y_0)(E-E_0) \\ & + \gamma_{12}(D-D_0)(GE-G_0E_0) + \psi_1(Y-Y_0)(G-G_0) \\ & + \psi_2(Y-Y_0)(E-E_0) + \psi_{12}(Y-Y_0)(GE-G_0E_0), \end{aligned}$$

and

$$\begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_{12} \end{pmatrix} \approx \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_{12} \end{pmatrix} - \gamma_{01} \begin{pmatrix} \tau^2 \gamma_1 - c_3 \gamma_1 \gamma_2 E_0 \\ \tau^2 \gamma_2 - c_3 \gamma_1 \gamma_2 G_0 \\ \tau^2 \gamma_{12} + c_3 \gamma_1 \gamma_2 \end{pmatrix} \quad (\text{A.10})$$

(A.10) becomes (14) when $c_3 = 0$. From (A.9), we can straightforwardly obtain the correction formula (15) for the conditional analysis, which is the same as (A.10). To obtain the correction formula (19) for the marginal analysis, expand

$$U(Y, G, E) = \sum_D \eta_1\{D; (Y, G, E)\} p(D|Y_0, G_0, E_0, S=1)$$

in the same way as for $T(Y, G, E)$. The general formula appears as

$$\begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_{12} \end{pmatrix} \approx \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_{12} \end{pmatrix} - \gamma_{01} \begin{pmatrix} (\tau^2 - \tau^{*2})\gamma_1 - (c_3 - c_3^*)\gamma_1\gamma_2 E_0 \\ (\tau^2 - \tau^{*2})\gamma_2 - (c_3 - c_3^*)\gamma_1\gamma_2 G_0 \\ (\tau^2 - \tau^{*2})\gamma_{12} - (c_3 - c_3^*)\gamma_1\gamma_2 \end{pmatrix}, \quad (\text{A.11})$$

which leads to (18) when $c_3=0=c_3^*$.

We can estimate parameters in the joint correction formula by maximizing the likelihood on the basis of (A.9). We can approximate the variance of the parameter estimator by the inverse of the observed information matrix. We use the δ method to find the approximate variance estimate for the corrected odds ratio estimator for the association of the secondary trait and the genetic marker. Specifically, let

$$\sqrt{n}\{(\hat{\psi}, \hat{\gamma}) - (\psi, \gamma)\} \rightarrow N\{(0, 0), V\},$$

where $(\hat{\psi}, \hat{\gamma})$ is the estimate of (ψ, γ) , and $\psi = (\psi_1, \psi_2, \psi_{12})$ and $\gamma = (\gamma_{01}, \gamma_1, \gamma_2, \gamma_{12})$. We can write correction formulas based on (A.10) or (A.11) in the form of $\theta = \psi + g(\gamma)$. The variance for $\hat{\theta}$ has the form

$$V_{\psi\psi} + V_{\psi\gamma} \frac{\partial g}{\partial \gamma} + \left(\frac{\partial g}{\partial \gamma} \right)^T V_{\gamma\psi} + \left(\frac{\partial g}{\partial \gamma} \right)^T V_{\gamma\gamma} \frac{\partial g}{\partial \gamma},$$

where

$$V = \begin{pmatrix} V_{\psi\psi} & V_{\psi\gamma} \\ V_{\gamma\psi} & V_{\gamma\gamma} \end{pmatrix}.$$

For the marginal analysis, we can use the same steps except that we now have $\gamma = (\gamma_0^*, \gamma_{01}, \gamma_1, \gamma_2, \gamma_{12})$, where γ_0^* is the intercept of the logistic regression model from the case-control sample. For different approximations, g is different, and its derivatives with respect to γ can be found accordingly.

Appendix B. More general correction formulas

When there are interactions between the secondary trait and the gene and/or environment factors, that is $\eta_1\{D; (Y, G, E)\} = \exp\{(D - D_0)m(Y, G, E)\}$, where

$$\begin{aligned}
m(Y, G, E) = & \gamma_{01}(Y \\
& - Y_0) + \gamma_{02}(Y G \\
& - Y_0 G_0) \\
& + \gamma_{03}(Y E \\
& - Y_0 E_0) + \gamma_{023}(Y G E \\
& - Y_0 G_0 E_0) \\
& + \gamma_1(G \\
& - G_0) + \gamma_2(E \\
& - E_0) + \gamma_{12}(G E - G_0 E_0).
\end{aligned}$$

We can modify the correction formulas for the conditional or the joint analysis to

$$\begin{aligned}
\theta_1^* &= \psi_1 + (\mu - D_0)\gamma_{02}^* + \tau^2\gamma_{01}^*\gamma_1^*, \\
\theta_2^* &= \psi_2 + (\mu - D_0)\gamma_{03}^* + \tau^2\gamma_{01}^*\gamma_2^*, \\
\theta_{12} &= \psi_{12} + (\mu - D_0)\gamma_{023} + \tau^2(\gamma_{01}^*\gamma_{12}^* + \gamma_1^*\gamma_{03}^* + \gamma_2^*\gamma_{02}^*) + c_3\gamma_{01}^*\gamma_1^*\gamma_2^*,
\end{aligned}$$

where

$$\theta_1^* = \theta_1 + \theta_{12}G_0, \theta_2^* = \theta_2 + \theta_{12}E_0, \gamma_{12}^* = \gamma_{12} + \gamma_{023}Y_0, \gamma_{02}^* = \gamma_{02} + \gamma_{023}E_0, \gamma_{03}^* = \gamma_{03} + \gamma_{023}G_0 \text{ and}$$

$$\begin{aligned}
\gamma_1^* &= \gamma_1 + \gamma_{02}Y_0 + \gamma_{12}E_0 + \gamma_{023}Y_0E_0 \\
\gamma_2^* &= \gamma_2 + \gamma_{03}Y_0 + \gamma_{12}G_0 + \gamma_{023}Y_0G_0 \\
\gamma_{01}^* &= \gamma_2 + \gamma_{02}G_0 + \gamma_{03}E_0 + \gamma_{023}G_0E_0
\end{aligned}$$

We can obtain the variance similarly for the simpler case by the δ method.

Figure 1. Conditional and marginal analyses and the correction estimators.

Table 1

Parameter estimates for simulated data with a binary primary phenotype and a binary secondary trait.

Methods	$\theta_1 = 0.5$			$\theta_2 = -0.5$			$\theta_{12} = -0.3$		
	Bias	A.std	E.std	Bias	A.std	E.std	Bias	A.std	E.std
r_0 approximates 1% (or $\beta_0 = -4.6$)									
Controls	-0.039	0.242	0.243	-0.007	0.151	0.152	-0.034	0.212	0.211
Marginal	0.073	0.122	0.124	0.116	0.091	0.094	0.018	0.105	0.109
Condit.	-0.025	0.125	0.127	-0.006	0.097	0.099	-0.029	0.109	0.111
Correct.	-0.008	0.124	0.126	0.011	0.095	0.097	-0.019	0.108	0.111
Comp	0.002	0.124	0.127	0.001	0.096	0.098	-0.002	0.109	0.112
MLE	0.002	0.124	0.126	0.001	0.096	0.098	-0.002	0.109	0.112
r_0 approximates 5% (or $\beta_0 = -2.95$)									
Controls	-0.081	0.267	0.265	-0.033	0.159	0.157	-0.076	0.216	0.227
Marginal	0.033	0.129	0.129	0.085	0.096	0.095	-0.018	0.112	0.117
Condit.	-0.066	0.132	0.132	-0.035	0.100	0.100	-0.067	0.116	0.119
Correct.	-0.004	0.129	0.130	0.028	0.096	0.096	-0.029	0.113	0.117
Comp.	0.008	0.130	0.133	0.003	0.097	0.100	-0.006	0.115	0.119
MLE	0.008	0.130	0.131	0.003	0.097	0.097	-0.006	0.115	0.120
r_0 approximates 10% (or $\beta_0 = -2.2$)									
Controls	-0.104	0.295	0.286	-0.071	0.157	0.162	-0.072	0.243	0.241
Marginal	0.001	0.127	0.135	0.050	0.096	0.097	-0.022	0.121	0.123
Condit.	-0.096	0.130	0.138	-0.070	0.101	0.101	-0.071	0.125	0.125
Correct.	-0.003	0.128	0.136	0.026	0.096	0.097	-0.014	0.123	0.123
Comp.	0.000	0.128	0.138	-0.004	0.098	0.101	-0.016	0.125	0.126
MLE	0.000	0.129	0.136	0.004	0.098	0.098	-0.015	0.125	0.124
r_0 approximates 15% (or $\beta_0 = -1.73$)									
Controls	-0.144	0.314	0.305	-0.093	0.165	0.167	-0.086	0.257	0.253
Marginal	-0.003	0.135	0.141	0.030	0.099	0.099	-0.027	0.127	0.127
Condit.	-0.100	0.137	0.144	-0.090	0.103	0.103	-0.076	0.129	0.130
Correct.	0.011	0.136	0.141	-0.023	0.099	0.099	-0.009	0.129	0.128
Comp.	0.003	0.136	0.144	-0.007	0.101	0.103	-0.008	0.130	0.130

Methods	$\theta_1 = 0.5$			$\theta_2 = -0.5$			$\theta_{12} = -0.3$		
	Bias	A.std	E.std	Bias	A.std	E.std	Bias	A.std	E.std
MLE	0.003	0.136	0.142	-0.008	0.101	0.099	-0.008	0.130	0.129

Baseline case-prevalence rate π_0 is assumed known in Correct., Comp., and MLE.

Bias, estimate-truth; A.std, square root of the averaged estimated variance; E.std, empirical standard deviation from the repeated simulations; θ_1 , genetic effect on the secondary phenotype; θ_2 , environment effect on the secondary trait; θ_{12} , gene-environment interaction on the secondary trait; Controls, analysis based on controls only; Marginal, marginal analysis; Condit., conditional analysis; Correct., corrected joint analysis based on the approximate odds ratio model; Comp., joint analysis by the compensator approach; MLE, the maximum likelihood approach.

Table II

Size and power of the tests under different assumptions on the baseline prevalence rate for the primary phenotype.

r_0	$(\theta_1, \theta_2, \theta_{12}) = (0, -0.5, 0)$, Size with type I error 5%					$(\theta_1, \theta_2, \theta_{12}) = (0.5, -0.5, -0.3)$, Power with type I error 5%				
	Test 1	Test 2	Test 3	Test 4	Test 5	Test 1	Test 2	Test 3	Test 4	Test 5
1%	5.80	6.14	8.38	10.44	11.80	<i>99.1</i>	99.6	99.7	99.8	99.8
	5.96	<i>5.10</i>	5.50	6.10	6.64	86.6	78.5	71.7	67.0	<i>64.7</i>
5%	8.40	4.68	4.40	5.20	5.56	99.2	99.7	99.7	99.7	99.7
	7.46	5.16	4.98	4.86	4.74	88.1	81.7	76.6	73.0	70.9
10%	8.36	5.30	4.60	4.60	4.76	<i>99.1</i>	99.5	99.5	99.6	99.6
	8.24	5.68	4.94	4.86	4.92	86.2	79.6	73.7	70.1	<i>68.0</i>
15%	8.70	5.52	4.80	4.78	4.90	98.9	99.3	99.5	99.6	99.7
	7.64	5.22	4.58	4.42	4.54	84.1	77.5	72.1	68.5	<i>66.4</i>

	Size with type I error 1%					Power with type I error 1%				
	Test 1	Test 2	Test 3	Test 4	Test 5	Test 1	Test 2	Test 3	Test 4	Test 5
1%	1.16	<i>1.12</i>	1.86	2.92	3.68	95.9	97.8	98.5	98.9	99.0
	1.36	<i>0.94</i>	1.24	1.36	1.54	65.8	54.4	46.0	41.6	38.9
5%	1.44	0.68	0.82	1.12	1.24	96.1	98.0	98.3	98.6	98.8
	1.56	0.84	0.86	0.82	0.84	70.7	60.6	52.9	48.3	<i>46.1</i>
10%	1.90	0.90	0.92	1.12	1.20	95.8	97.5	98.2	98.3	98.4
	1.94	1.16	1.00	<i>1.00</i>	1.02	67.3	56.6	48.9	44.7	<i>42.6</i>
15%	2.18	0.98	0.84	0.88	0.88	95.6	97.1	97.7	98.0	98.2
	1.86	1.12	0.82	0.86	0.88	64.2	54.4	47.0	42.6	<i>40.2</i>

The first row under each baseline prevalence rate r_0 is the test of no genetic effects, and the second row is the test of no gene-environment interaction effect. The entries under each test are the rejection rate times 100. Entries with bold face are tests with correct baseline prevalence rate assumed. The values in *italics* are the minimum p -values or power achieved among a range of r_0 .

Test 1, Wald test assuming $\tau^2 = 0.01 \times 0.99$; Test 2, Wald test assuming $\tau^2 = 0.05 \times 0.95$; Test 3, Wald test assuming $\tau^2 = 0.1 \times 0.9$; Test 4, Wald test assuming $\tau^2 = 0.15 \times 0.85$; Test 5, $\tau^2 = 0.2 \times 0.8$.

Table III

Analysis of the secondary trait *prostate specific antigen* in the case-control study of prostate cancer: test of no genetic effect.

CHR	SNP	%n _T	Test 1 ($\times 10^{-4}$)	Test 2 ($\times 10^{-4}$)	Test 3 ($\times 10^{-4}$)	MinT ($\times 10^{-4}$)
2	rs10197545	0.704	0.953 0.953	0.552 0.400	0.729 0.460	0.953
3	rs17034929	-0.037	2.20 2.20	0.467 0.674	0.150 0.446	2.20
6	rs9448387	0.687	1.40 1.40	0.93 0.449	1.33 0.480	1.40
10	rs2255684	0.550	0.421 0.421	0.479 0.478	1.00 0.590	1.00
12	rs7135149	-0.150	1.46 1.46	0.675 1.50	0.465 1.44	1.46
12	rs11068651	0.414	0.600 0.600	0.246 0.250	0.199 0.240	0.600
13	rs9594159	-0.022	7.41 7.41	1.98 3.91	0.776 2.40	7.41
13	rs9316673	0.462	0.215 0.215	0.048 0.054	0.025 0.040	0.215
19	rs350134	0.095	4.06 4.06	1.66 2.20	0.990 1.90	4.06

CHR, Chromosome. SNP, single nucleotide polymorphism; Test 1, p -value of the Wald test assuming $r^2 = 0$; Test 2, p -value of the Wald test assuming $r^2 = 0.05 \times 0.95$; Test 3, p -value of the Wald test assuming $r^2 = 0.1 \times 0.9$. MinT, p -value of the minimum test for the baseline prevalence rate ranging from 0% to 10% evaluated at an interval of 0.2%. The second entry is the maximum likelihood approach.