Published in final edited form as:

Ann Hum Genet. 2012 November ; 76(6): 484–499. doi:10.1111/j.1469-1809.2012.00725.x.

Analysis of Secondary Phenotype Involving the Interactive Effect of the Secondary Phenotype and Genetic Variants on the Primary Disease

Jian Wang and Sanjay Shete

Departments of Biostatistics and Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas

Summary

A genome-wide association (GWA) study is usually designed as a case-control study, where the presence and absence of the primary disease defines the cases and controls, respectively. Using the existing data from GWA studies, investigators are also trying to identify the association between genetic variants and secondary phenotypes, which are defined as traits associated with the primary disease. However, recent studies have shown that bias arises in the estimation of marker-secondary phenotype association using originally collected data. We recently proposed a bias correction approach to accurately estimate the odds ratio (OR) for marker-secondary phenotype association. In this communication, we further investigated whether our bias correction approach is robust for a scenario involving the interactive effect of the secondary phenotype and genetic variants on the primary disease. We found that in such a scenario, our bias correction approach also provides an accurate estimation of OR for marker-secondary phenotype association. We investigated accuracy of our approach using simulation studies and showed that the approach better controlled for type I errors than the existing approaches. We also applied our bias correction approach to the real data analysis of association between an *N*-acetyltransferase gene, *NAT2*, and smoking on the basis of colorectal cancer data.

Keywords

odds ratio; bias; secondary phenotype; SNP; genome-wide association study; frequency-matched study design

Introduction

Genome-wide association (GWA) study is a popular approach for detecting genetic variants associated with primary diseases of interest. A GWA study is usually designed as a case-control study, where the primary disease defines cases (subjects with the primary disease) and controls (subjects without the primary disease). In a GWA study, information about secondary phenotypes, defined as traits associated with the primary disease, is also collected. Using the existing data from GWA studies, investigators are trying to identify the association between genetic variants and secondary phenotypes. However, recent studies have shown that bias arises in the estimations of marker-secondary phenotype association because the presence and absence of the secondary phenotype (i.e., cases and controls with respect to the secondary phenotype) are not sampled following the principles of case-control

study design. Several approaches were proposed to provide unbiased estimators for measures of marker-secondary phenotype association, including the inverse-probability-of-sampling-weighted (IPW) regression approach (Monsees et al., 2009; Richardson et al., 2007), the likelihood-based approach (Lin & Zeng, 2009), the adaptively weighted (AW) method (Li et al., 2010), and the bias-correction approach that solves a series of non-linear estimating equations (Wang & Shete, 2011a).

The study conducted by Li and colleagues is the only one that investigated the scenario in which the secondary phenotype and the genetic variants have an interactive effect on the primary disease (Li et al., 2010). They found that this interactive effect could have an impact on the odds ratio (OR) estimators of marker-secondary phenotype association and proposed an AW method using both primary disease cases and controls. However, their study only focused on very rare primary diseases and an unmatched case-control study design, which assumed either a dominant or recessive genetic model for the genetic variants (i.e., the genetic variant could not be modeled as additive).

We (Wang & Shete, 2011a) also proposed a bias correction approach to evaluate the accurate OR estimations for genetic variants associated with a secondary phenotype. Briefly, our approach provides an accurate OR estimation by solving a system of non-linear equations iteratively, in which the prevalence values of primary disease and secondary phenotype are taken into account. In the original study, we showed that the bias correction approach can provide more accurate OR estimations and is more powerful than those obtained using the standard approaches (logistic regressions, with or without adjustment for the primary disease status) or the IPW regression approach, especially when the original case-control study of primary disease is frequency-matched with respect to the secondary phenotype of interest (Wang & Shete, 2011a; Wang & Shete, 2011b). However, our original study did not investigate the special situation where the secondary phenotype and the genetic variant have an interactive effect on the primary disease, as reported in (Li et al., 2010). Therefore, in the current study we investigated whether our approach is robust in the case of an interactive effect of the secondary phenotype and genetic variants on the primary disease, and we found that our bias correction approach does provide accurate OR estimation for marker-secondary phenotype association. Also, previously we used the standard error from logistic regression, without adjusting for primary disease status, in our bias-correction approach. However, we found that when the secondary phenotype and the single nucleotide polymorphism (SNP) have an interactive effect on the primary disease, the standard error from standard logistic regression without adjusting for primary disease status is also biased. Thus, the estimated standard error cannot be used as the standard error of the natural logarithm of the corrected OR, as suggested in our previous study (Wang & Shete, 2011b). Therefore, in this communication, we developed a new approach, using asymptotic formulas, to estimate standard error for the corrected OR when the secondary phenotype and genetic variants interact. We further proposed a bootstrapping approach to evaluate empirical p values corresponding to the corrected OR.

We investigated our approach using simulation studies and compared the results to those of standard logistic regression approaches, the controls-only (CO) approach (using only subjects without the primary disease), the AW method, and the IPW regression approach. When the secondary phenotype and the marker have an interactive effect on the primary disease, the standard logistic regression approaches provided biased OR estimators for all scenarios; the IPW regression approach could be biased when the original study of primary disease was frequency-matched on the secondary phenotype but was unbiased for unmatched studies. The AW method was proposed for this particular scenario involving a secondary phenotype-marker interactive effect on the primary disease; however, we found that the method works only when the primary disease is very rare (e.g., prevalence of

disease 4×10^{-5} , as in the Li et al. (2010) study, data not shown) but could be biased when the disease is relatively more common (e.g., prevalence of disease 5%). The CO approach performed very similarly to the AW approach for all the scenarios. Indeed, from our simulation studies, we observed that when the primary disease was very rare, the CO approach, AW method, IPW regression approach, and our bias correction approach performed very similarly, and all provided an accurate estimation of OR for secondary phenotype-marker association in all scenarios. Therefore, in this study, we focused on less common to common primary diseases (e.g., 5% to 30% prevalence of disease). The simulation results showed that our bias correction approach uniformly provided accurate OR estimations for marker-secondary phenotype association compared to the other approaches, whether or not there was an interactive effect of the genetic variant and secondary phenotype on the primary disease. The bias correction approach also better controlled type I errors and was robust for both unmatched and frequency-matched case-control study designs for primary disease and for different primary disease prevalence values (e.g., rare, less common, and common primary diseases). Finally, we applied the bias correction approach to a real data analysis of association between the *N*-acetyltransferase gene *NAT2* and smoking using existing colorectal cancer data, where *NAT2* and smoking have an interactive effect on colorectal adenoma.

Methods

We considered a scenario where both the secondary phenotype and genetic variant are associated with the primary disease and, furthermore, have an interactive effect on the primary disease. We assumed binary random variables for the primary disease and secondary phenotype, denoted as $Y = \{0, 1\}$ and $T = \{0, 1\}$, respectively, with 0 representing individuals without the primary disease/secondary phenotype and 1 representing individuals with the primary disease/secondary phenotype. We considered a SNP locus with two alleles: deleterious allele A and normal allele a . We first assumed a dominant or recessive genetic model for the genetic variant and also denoted it as a binary random variable $X = \{0, 1\}$. For a dominant genetic model, 0 represents genotype (a, a) and 1 represents genotypes (A, a) and (A, A) ; for a recessive genetic model, 0 represents genotypes (a, a) and (A, a) and 1 represents genotype (A, A) . Note that if an additive genetic model was assumed, a categorical random variable $X = \{0, 1, 2\}$ was denoted to represent genotypes (a, a) , (A, a) , and (A, A) , respectively. Given three random variables, X , T , and Y , the association among all random variables can be expressed using the following conditional probabilities with logistic models. To include the interaction effect of X and T , we added an interaction term of X and T in Equation (2).

$$\Pr(T=1|X=i) = \frac{\exp(\alpha_0 + \alpha_1 i)}{1 + \exp(\alpha_0 + \alpha_1 i)}, \text{ and} \quad (1)$$

$$\Pr(Y=1|X=i, T=k) = \frac{\exp(\beta_0 + \beta_1 i + \beta_2 k + \beta_{12} ik)}{1 + \exp(\beta_0 + \beta_1 i + \beta_2 k + \beta_{12} ik)}, \quad (2)$$

where α_0 , β_0 , α_1 , β_1 , β_2 , and β_{12} are regression coefficients and $i, k = 0, 1$. The $\text{OR} = \exp(\alpha_1)$ is the OR for the association between the genetic variant and the secondary phenotype, which is the parameter of interest. To obtain the accurate OR estimation of SNP-secondary phenotype association using the bias correction approach, we followed the same procedure proposed in our original study (Wang & Shete, 2011a), but now employed the conditional probabilities shown above to include the interactive effect of the secondary phenotype and the SNP on the primary disease. Basically, the biased OR estimated from the logistic regression without adjusting for the primary disease, the prevalence values of the

primary disease, and the prevalence values of the secondary phenotype can be expressed using non-linear equations with the unknown regression coefficients of α_0 , β_0 , and α_1 . The prevalence values are assumed to be known from the literature (our approach is robust to this assumption, as shown in the Sensitivity Analysis section). We estimated the biased OR and values of the other parameters necessary for the non-linear equations from the data (e.g., regression coefficients $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_{12}$). The solution of the system of non-linear equations gives us the corrected OR for the SNP-secondary phenotype association (see (Wang & Shete, 2011a) for analytical details of these non-linear equations).

For the purpose of comparison, in this study we considered different approaches to assess the OR for SNP-secondary phenotype association (i.e., $OR = \exp(\alpha_1)$). We used notations similar to those in our original study: (1) $\widehat{OR}$ was obtained using logistic regression, without adjustment for the primary disease status; (2) $\widehat{OR}_{adj}$ was obtained using logistic regression, with adjustment for the primary disease status; (3) OR_{CO} was obtained using the CO approach; (4) OR_{AW} was obtained using the AW method; (5) OR_{IPW} was obtained using the extended IPW regression method; and (6) $\widehat{OR}$ and $\widehat{OR}_m$ were obtained using the bias correction approach, as proposed in the original study (Wang & Shete, 2011a), for the unmatched and frequency-matched case-control study designs for primary disease, respectively. Accordingly, we denoted random variables $\widehat{\alpha}_1$, $\widehat{\alpha}_{1adj}$, $\widehat{\alpha}_{1CO}$, α_{1AW} , α_{1IPW} , and $\tilde{\alpha}_1$ (or $\tilde{\alpha}_{1m}$) to represent the natural logarithm ORs of $\log(\widehat{OR})$, $\log(\widehat{OR}_{adj})$, $\log(OR_{CO})$, $\log(OR_{AW})$, $\log(OR_{IPW})$, and $\log(\widehat{OR})$ (or $\log(\widehat{OR}_m)$), respectively, and denoted $\hat{s}$, $\hat{s}_{adj}$, s_{CO} , s_{AW} , s_{IPW} , and $\tilde{s}$ (or $\tilde{s}_m$) as the corresponding standard errors.

In the follow-up study of our bias correction approach (Wang & Shete, 2011b), we showed that the empirical distribution of $\tilde{\alpha}_1 = \log(\widehat{OR})$ follows a normal distribution under the null hypothesis. Also, we found that the standard error $\tilde{s}$ is close to $\hat{s}$ obtained using standard logistic regression without adjusting for primary disease status, and therefore, we applied $\hat{s}$ as the estimation of $\tilde{s}$. However, when the secondary phenotype and the SNP have an interactive effect on the primary disease, the estimation of the standard error $\hat{s}$ is also biased. Therefore, instead of using the biased $\hat{s}$ as the estimation of $\tilde{s}$, in the current study we employed the asymptotic standard error for the natural logarithm OR to estimate $\tilde{s}$. Given a sample with N_I cases and N_0 controls, the asymptotic $\tilde{s}$ can be evaluated using the following formula:

$$\tilde{s} = \sqrt{\sum_{ki} \frac{1}{n_{ki}}} = \sqrt{\sum_{ki} \frac{1}{E_{ki}}}, \quad (3)$$

where n_{ki} is the number of individuals in the sample with the secondary phenotype random variable $T = k$ and the SNP random variable $X = i$. The E_{ki} is the expected number of individuals n_{ki} conditional on N_I and N_0 . The details for evaluating E_{ki} are given in the Appendix I. Note that now the conditional probabilities embedded in the formulas for assessing E_{ki} include the interaction term of secondary phenotype and genetic variant on the primary disease. Given that corrected $\tilde{\alpha}_1 = \log(\widehat{OR})$, and $\tilde{s}$, the significance of $\tilde{\alpha}_1$ can be evaluated using the standard Wald test. When the study was frequency-matched with respect to the secondary phenotype of interest, the corresponding standard error $\tilde{s}_m$ could be evaluated in a similar way. In this situation, however, the values of E_{ki} are calculated differently (see Appendix I).

Additive Genetic Model

In the original study, we also proposed a method when the SNP was assumed to act additively. We employed the property that the biased $\widehat{OR}$ obtained using logistic regression, without adjusting for primary disease status, is given by the per-allele OR. We applied two different approaches to evaluate the per-allele OR and obtained two corrected $\widehat{OR}_S$. The final corrected $\widehat{OR}$ was then estimated as the average of the two $\widehat{OR}_S$. However, we noticed that, when there is an interactive effect of secondary phenotype and SNP on primary disease, the $\widehat{OR}$ obtained using logistic regression does not give the per-allele OR. Therefore, the original bias correction approach for additive genetic model could not be adapted directly. To obtain the true per-allele OR, we employed two ways to assess OR. First, we obtained the biased $\widehat{OR}_1$ by calculating the OR of SNP random variable $X=1$ versus $X=0$, which gives OR for heterozygous genotype against wild-type homozygous genotype. Second, we obtained the biased $\widehat{OR}_2$ by calculating the OR of SNP random variable $X=2$ versus $X=0$, which gives OR for homozygous genotype for variant allele against wild-type homozygous genotype and is denoted as $\widehat{OR}_2$ in this communication. Based on $\widehat{OR}_1$ and $\widehat{OR}_2$, following the different formulas in the original study, we can obtain two $\widehat{OR}_S$: $\widehat{OR}_1$ and $\widehat{OR}_2$. The final $\widehat{OR}$ for additive genetic model is given as: $\widehat{OR} = \exp((\log(\widehat{OR}_1) + \log(\widehat{OR}_2))/2)$ which is slightly more efficient than the estimator proposed in our previous studies. Denote the standard errors of random variables $\log(\widehat{OR}_1)$ and $\log(\widehat{OR}_2)$ as $\tilde{s}_1$ and $\tilde{s}_2$, which can be evaluated based on Equation (3) using corresponding expected numbers of individuals, then the asymptotic standard error of random variable $\tilde{\alpha}_1 = \log(\widehat{OR})$ can be evaluated as:

$$\tilde{s} = \frac{1}{2} \sqrt{\tilde{s}_1^2 + \tilde{s}_2^2/4 + \tilde{c}_{12}},$$

where $\tilde{c}_{12}$ denotes the covariance of random variables $\log(\widehat{OR}_1)$ and $\log(\widehat{OR}_2)$. Applying the delta-method, we estimated asymptotic $\tilde{c}_{12} = 1/n_{00} + 1/n_{11} = 1/E_{00} + 1/E_{11}$. The detail of derivation of $\tilde{s}$ is shown in the Appendix II. We also applied the inverse variance weight to estimate the final $\widehat{OR}$ for an additive genetic model of the SNP, but obtained similar results. Therefore, in this communication, we will only report the results obtained on the basis of the average of two estimators, $\log(\widehat{OR}_1)$ and $\log(\widehat{OR}_2)$.

Bootstrapping: Empirical P Value

Although the Wald test based on asymptotic standard error is simple and straightforward, the deviations of underlying assumptions might make the theoretical p values too conservative or too liberal. Therefore, in addition to the Wald test, we proposed a bootstrapping approach to compute the empirical p values (Efron & Tibshirani, 1993). Given a dataset of interest with N_I cases and N_O controls, the empirical p value corresponding to the estimated $\widehat{OR}$ (or $\widehat{OR}_m$) for secondary phenotype-SNP association was obtained as follows:

1. Given the regression coefficients $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_{12}$ obtained from logistic regression without adjusting for the primary disease, the corrected coefficients $\tilde{\alpha}_0$ and $\tilde{\beta}_0$ obtained from our bias correction approach, and other information obtained from the dataset (e.g., MAF, case-control ratio, sample size, genetic model, and study design), we generate B bootstrap datasets under the null hypothesis that there is no association between secondary phenotype and genetic variant (i.e., $\alpha_1=0$). Each

bootstrap dataset has N_I cases and N_O controls. Details of the data simulation are described in the next subsection, Simulation Approach.

2. For each bootstrap dataset, we evaluate the marker-secondary phenotype association $\widehat{OR}_u^*$, $u=1, 2, \dots, B$. We then calculate $\widehat{OR}_u^*$, $u=1, 2, \dots, B$ using our bias correction approach.
3. The empirical p value of corrected $\widehat{OR}$ (or $\widehat{OR}_m$) can be obtained as (1) the proportion of the $\widehat{OR}_u^* < \widehat{OR}$ or $\widehat{OR}_u^* \geq 1/\widehat{OR}$ when $\widehat{OR} < 1$ or (2) the proportion of the $\widehat{OR}_u^* \geq \widehat{OR}$ or $\widehat{OR}_u^* < 1/\widehat{OR}$ when $\widehat{OR} \geq 1$, $u=1, 2, \dots, B$.

Simulation Approach

We performed simulation studies to investigate the performance of our bias correction approach for the situation where the secondary phenotype and SNP have an interactive effect on the primary disease; we also compared the bias correction approach to the standard logistic regressions, CO approach, AW method, and IPW regression approach. We assumed a single di-allele SNP with a minor allele frequency (MAF) of 40%. We fixed the regression coefficients at $\beta_1 = 0.1823$ and $\beta_2 = 0.5596$, which correspond to ORs of 1.2 and 1.75, respectively. For the null model (no association between the SNP and the secondary phenotype), we set $\alpha_1 = 0$, which corresponds to an OR of 1; for the alternative model (association between the SNP and the secondary phenotype), we set $\alpha_1 = 0.4055$, which corresponds to an OR of 1.5. We also fixed the prevalence value of the secondary phenotype f_T as 10%, since we found in our original study that f_T has a relatively small impact on the bias in OR estimation for SNP-secondary phenotype association. We considered different scenarios using different prevalence values of primary disease f_D (5%, 10%, and 30%) and a range of regression coefficients for the interaction term β_{12} (± 1.0986 , ± 0.4055 , and 0), which corresponds to different ORs (0.333, 0.667, 1.000, 1.500, 3.000). Note that the pre-defined primary disease prevalence values cover less common to common diseases, and the values of β_{12} correspond to both protective and deleterious interactive effects. On the basis of these pre-defined parameters, we obtained the values for the intercept regression coefficients α_0 and β_0 .

First, we generated genotypes for a SNP with the use of the genotype frequencies, assuming that the SNP is in Hardy-Weinberg proportion. The secondary phenotype values were then generated on the basis of the dataset of realizations of SNP using Equation (1). Conditioned on the secondary phenotype and SNP values, we used Equation (2) to generate the values of the primary disease assuming three different genetic models for the SNP: additive, dominant, and recessive. In this way, we simulated a large amount of data on the population of interest and then randomly sampled 3000 cases (individuals with the primary disease) and 3000 controls (individuals without the primary disease). When a frequency-matched case-control study design with respect to the secondary phenotype of interest was considered, the 3000 primary disease cases were still sampled randomly. However, the 3000 controls were sampled so that the proportion of the presence of the secondary trait in the controls was approximately equal to that in the cases (Rothman & Greenland, 1998). For the matched study, we assumed that the difference between the proportions of the presence of the secondary trait in cases and controls was within $\pm 2\%$, with equal probability. The average results and type I error rates reported for the simulation studies were based on 1000 replicate datasets. For each replicate, we evaluated the $\widehat{OR}$, $\widehat{OR}_{adj}$, OR_{CO} , OR_{AW} , OR_{IPW} , and $\widehat{OR}$ (or $\widehat{OR}_m$), and corresponding p values for SNP-secondary phenotype association, for the different approaches studied.

Results

We report the medians of estimated ORs and the 95% confidence intervals (CIs) of ORs for six different approaches to identifying the SNP-secondary phenotype association: (1) logistic regression without adjustment for primary disease; (2) logistic regression with adjustment for primary disease; (3) the CO approach; (4) the AW method; (5) the extended IPW regression; and (6) the bias correction approach proposed in our original study (Wang & Shete, 2011a). We investigated both unmatched and frequency-matched case-control study designs for the study of primary disease.

Tables 1-3 show the simulation results for the frequency-matched study design under the assumption of a dominant, additive, or recessive genetic model for the SNP, respectively. In each table, the top panel shows the results obtained under the null hypothesis that there is no marker-secondary phenotype association ($OR = 1$), and the bottom panel shows results obtained under the alternative hypothesis that there is a marker-secondary phenotype association ($OR = 1.5$). The results of the AW method are not reported for the additive genetic model (Table 2) because the method is not applicable to an additive model.

For all genetic models and all scenarios with different prevalence of disease f_D and regression coefficient β_{12} , we observed that logistic regressions, with and without adjusting for primary disease status, provided biased OR estimations (see Tables 1-3). The magnitude of the bias in the OR estimations increased when the absolute interactive effect was larger. The results for both logistic regressions were very similar, owing to the frequency-matched study design. The CO and AW methods over-estimated the ORs when the interactive effects were protective but under-estimated the ORs when the interactive effects were deleterious. Furthermore, bias in the OR estimation using the CO and AW methods increased when the primary disease was common. Meanwhile, we found that the IPW regression provided relatively accurate OR estimation in scenarios when the interactive effect on the primary disease was protective; however, it still resulted in highly biased OR estimates in most of the scenarios in which the interactive effect on the primary disease was deleterious. When there was no interactive effect ($\beta_{12} = 0$) on the primary disease, the logistic regressions, the CO approach, the AW method, and the IPW regression performed very similarly, as expected.

Our bias correction approach uniformly provided accurate OR estimations for all the scenarios tested. For example, for the dominant genetic model (Table 1), when $f_D = 30\%$ and $\beta_{12} = 1.0986$ (corresponding $OR = 3$), we found that (1) under the null hypothesis, the median $\widehat{OR} = 0.89$ (95% CI = (0.78–1.01)), median $\widehat{OR}_{adj} = 0.89$ (95% CI = (0.78–1.01)), median $OR_{CO} = 0.54$ (95% CI = (0.45–0.65)), median $OR_{AW} = 0.54$ (95% CI = (0.45–0.65)), and median $OR_{IPW} = 0.72$ (95% CI = (0.63–0.84)), all of which under-estimated the true OR of 1, whereas based on our approach, the median $\widehat{OR}_m = 1.00$ (95% CI = (0.83–1.21)), which agreed well with the true OR, and (2) under the alternative hypothesis, the median $\widehat{OR} = 1.29$ (95% CI = (1.13–1.49)), median $\widehat{OR}_{adj} = 1.30$ (95% CI = (1.13–1.49)), median $OR_{CO} = 0.81$ (95% CI = (0.67–0.97)), median $OR_{AW} = 0.81$ (95% CI = (0.68–0.97)), and median $OR_{IPW} = 1.06$ (95% CI = (0.91–1.23)), all of which under-estimated the true OR of 1.5, whereas the median $\widehat{OR}_m = 1.51$ (95% CI = (1.24–1.83)), which agreed well with the true OR.

We found that for the unmatched study design, the logistic regressions and the CO approach resulted in biased OR estimations. The AW method was also biased for almost all of the scenarios tested owing to the violation of the rare disease assumption. However, the extended IPW regression method and our own bias correction approach provided accurate

estimations of the OR for SNP-secondary phenotype association (simulation results reported in the Supporting Information Tables S1-S3).

We also performed simulation studies for very rare diseases to investigate the performance of different approaches. We employed the same parameters used by Li et al. (Li et al., 2010). Specifically, we set $\beta_1 = \beta_2 = 0$, $f_I = 30\%$, and $MAF = 16\%$. The regression coefficients were $\alpha_1 = 0$ (OR=1) for the null hypothesis and $\alpha_1 = 0.25$ (OR=1.28) for the alternative hypothesis. The value of the interaction coefficient β_{12} ranged from -2 to 2. By fixing the intercept coefficient β_0 as -10, we obtained primary disease prevalence f_D values from $\sim 4.0 \times 10^{-5}$ to $\sim 7.5 \times 10^{-5}$ for different scenarios with different interaction coefficient β_{12} . We considered dominant and recessive genetic models. All the results were obtained based on 1000 replicates, each with 1000 primary disease cases and controls. We found that, for very rare diseases, both standard logistic regressions still performed poorly, whereas the CO approach, the AW method, the extended IPW regression approach, and our bias correction approach performed very similarly and provided accurate OR estimations for the association between secondary phenotype and genetic variants for both unmatched and frequency-matched case-control studies. For example, for the frequency-matched study design, when $\alpha_1 = 0$, $\beta_{12} = 1$ and $f_D = 5.2 \times 10^{-5}$, we observed that the median $\widehat{OR} = 1.68$ and the median $\widehat{OR}_{adj} = 1.70$, which highly over-estimated the true OR of 1, while the median $OR_{CO} = 1.01$, $OR_{AW} = 1.02$, median $OR_{IPW} = 1.01$, and median $\widehat{OR}_m = 1.01$, which all agreed well with the true OR.

Type I Error Estimate

We also investigated the type I error probabilities for different approaches (see Table 4). We employed the approaches described in the Methods section to evaluate p values associated with corrected $\widehat{OR}$ and $\widehat{OR}_m$, including the Wald test and the bootstrapping approach. The type I error probabilities were estimated at a nominal significance level of 0.05. As expected, in the scenarios where the OR estimations were biased for the logistic regressions, CO approach, AW method, and IPW approach, the type I error probabilities were dramatically inflated. In some scenarios, when the absolute magnitude of the interactive effect was large (e.g., $\beta_{12} = \pm 1.0986$), the type I error rates for these approaches could be equal to 1 because of the significantly biased estimation of the OR. We observed that using the Wald test based on asymptotic standard error for our bias correction approach resulted in type I error estimates that were conservative in most of the scenarios and liberal in a few scenarios, which could have been due to the violation of the assumptions underlying derivation of asymptotic standard errors. However, the bootstrapping approach provided more accurate type I error estimates for all the scenarios.

For example, in the frequency-matched study design with a dominant genetic model (Table 4), when $f_D = 10\%$ and $\beta_{12} = 0.4055$ (corresponding OR = 1.5), the type I error rates were 0.289, 0.280, 0.154, 0.141, and 0.107 for logistic regression without adjustment, logistic regression with adjustment, the CO approach, the AW method, and IPW regression, respectively, which were substantially higher than the nominal significance level of 0.05. In contrast, our bias correction approach can control type I errors. Specifically, the Wald test based on asymptotic standard error had a type I error probability of 0.016 and the bootstrapping approach had a type I error of 0.049, which were well controlled at a nominal significance level of 0.05. The type I errors were also investigated in unmatched studies (data not reported). We found that, in this situation also, our bias correction approach, as well as the extended IPW method, had well-controlled type I error rates, but the CO approach, the AW method, and standard logistic regression-based approaches did not.

Sensitivity Analysis

In the simulation studies, we assumed that the prevalences of the primary disease and secondary phenotype were known. However, these quantities are generally estimated, with associated confidence intervals. Therefore, we assessed the sensitivity of the corrected $\widehat{OR}$ and $\widehat{OR}_m$ to the estimated prevalence values of the primary disease $\hat{f}_D$ and the secondary phenotype $\hat{f}_T$. Specifically, we considered a range of prevalence values centered on the true prevalences for both primary disease and secondary phenotype, that is, $(f_D - \Delta, f_D + \Delta)$ and $(f_T - \Delta, f_T + \Delta)$, where Δ is the error term. In our simulation study, for sensitivity analyses we set $\Delta = 1\%$. The choice of Δ was based on confidence lengths associated with the prevalence values of several diseases and phenotypes. For example, the estimated prevalence of current smokers in US adults was 22.5% (95% CI = (21.9% - 23.1%)) (CDC, 2004); the estimated prevalence of obesity in US adults age 20 and over was 28.4% (95% CI = (27.7% - 29.1%)) (Cohen & Martinez, 2010); and the estimated prevalence of asthma in the US was 7.8% (95% CI = (7.6% - 8.0%)) (Moorman et al., 2011). From these examples, we noted that the lengths of the confidence intervals for the estimated prevalence values were relatively short (i.e., $<2\%$).

Therefore, we considered a range of $\pm 1\%$ of the true prevalence values for the sensitivity analysis. Eight different misspecifications were assessed: (1) $\hat{f}_D = f_D + \Delta$ and $\hat{f}_T = f_T + \Delta$; (2) $\hat{f}_D = f_D + \Delta$ and $\hat{f}_T = f_T - \Delta$; (3) $\hat{f}_D = f_D - \Delta$ and $\hat{f}_T = f_T + \Delta$; (4) $\hat{f}_D = f_D - \Delta$ and $\hat{f}_T = f_T - \Delta$; (5) $\hat{f}_D = f_D$ and $\hat{f}_T = f_T + \Delta$; (6) $\hat{f}_D = f_D$ and $\hat{f}_T = f_T - \Delta$; (7) $\hat{f}_D = f_D + \Delta$ and $\hat{f}_T = f_T$; and (8) $\hat{f}_D = f_D - \Delta$ and $\hat{f}_T = f_T$. The results of the sensitivity analysis for the corrected $\widehat{OR}_m$ are listed in Tables 5-7. All the results were very similar to those obtained using the exact prevalences. For example, for the dominant genetic model (Table 5), when $f_D = 30\%$, $f_T = 10\%$, and $\beta_{12} = 1.0986$ (corresponding OR = 3), we found that (1) under the null hypothesis, the corrected $\widehat{OR}_m$ was estimated as 1.00 when using the real prevalence values and as 0.98~1.02 when the prevalence values were misspecified; (2) under the alternative hypothesis, the corrected $\widehat{OR}_m$ was estimated as 1.50 when using the true prevalences and as 1.48~1.54 when the prevalence values were misspecified. For both hypotheses, the misspecifications of prevalence values of primary disease and secondary phenotype did not have a major impact on the corrected OR estimations. We also performed the sensitivity analysis for the corrected $\widehat{OR}$ for the unmatched study design and observed similar results (Supporting Information Tables S4-S6).

We further performed the sensitivity analyses of our approach when the Δ value, error associated with estimation of prevalence value, is higher. We now fixed the Δ at 20% of the true prevalence values when the true prevalence of primary disease or secondary phenotype was 5% or 10%; and at 10% of the true prevalence value when the true prevalence value was 30%. The results from the sensitivity analyses for these higher values of error in estimation of prevalence values were reported in the Supporting Information: Tables S7-S9 for the unmatched study design with dominant, additive and recessive genetic model, respectively; and Tables S10-S12 for the frequency-matched study design with different genetic model, respectively. We observed that the proposed approach is overall robust but could be slightly sensitive when the interaction effect is also very large (e.g., $\beta_{12} = 1.0986$).

Real Data Analysis

We next applied the bias correction approach to the case-control association study of the *N*-acetyltransferase gene *NAT2* and smoking based on data from an actual case-control study of colorectal adenoma. Colorectal adenoma is a benign tumor, but it is considered to be a precursor of colorectal cancer (Li et al., 2010; Moslehi et al., 2006). Both smoking and the *NAT2* gene variants related to a low acetylator phenotype are associated with an increased

risk of colorectal adenoma (Giovannucci, 2001; Ji et al., 2006; Moslehi et al., 2006). We used the same real data set used in the study by Li et al. (2010), which involved $N_I = 605$ colorectal adenoma cases and $N_0 = 610$ controls. The data were originally from the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial at the National Cancer Institute. We employed the same coding as that used by Li et al.: $Y = 1$ denotes individuals with colorectal adenoma and $Y = 0$ denotes individuals without colorectal adenoma; $T = 1$ denotes current smokers and $T = 0$ denotes former and never smokers; and $X = 1$ denotes haplotypes corresponding to rapid acetylation and $X = 0$ denotes other haplotypes. We also noted that the proportions of current smokers among colorectal adenoma cases and controls were 64% and 56%, respectively. The results obtained using standard logistic regressions, the CO approach, the AW method, and the extended IPW regression are also reported.

We first performed logistic regressions (Equations (1) and (2)) to estimate the biased $\widehat{OR}$, as well as the regression coefficients to be substituted into the nonlinear equations, and obtained the following values: $\widehat{\beta}_1 = 0.8357$, $\widehat{\beta}_{12} = -1.5876$, and the biased $\widehat{OR} = 0.8375$. The haplotype frequency relating to $X = 1$ was also estimated from the data as 5.27%. The prevalence values of colorectal adenoma ($\hat{f}_D$) and current smokers ($\hat{f}_T$) in the general population were obtained from the literature as 4.43% and 20%, respectively (CDC, 2005; Moslehi et al., 2006). Furthermore, the OR of smoking status associated with colorectal adenoma was set as 1.71 (i.e., $\widehat{\beta}_2 = 0.5365$), as shown in a previous study (Erhardt et al., 2002). All the results obtained using different approaches are reported in Table 8. Although all the approaches showed that there is no statistically significant association between the gene *NAT2* and smoking behavior, the OR estimators obtained suggested different direction (protective or risk) for associations between them. Both logistic regression methods (with and without adjusting for primary disease status) gave very similar results and showed that there is a negative association between the gene *NAT2* and smoking behavior: $\widehat{OR} = 0.8375$ and $\widehat{OR}_{adj} = 0.8418$. That is, $\widehat{OR}$ and $\widehat{OR}_{adj}$ suggested that the individuals with haplotypes corresponding to rapid acetylation are about 16% less likely to be smokers than those with other haplotypes. On the other hand, the CO approach, the AW method, the extended IPW regression method, and our bias correction approach suggested that there is a positive association between the gene *NAT2* and smoking behavior (OR = 1.8502, 1.7664, 1.7212, and 1.7377, respectively). That is, CO, AW, IPW, and our method concluded that the haplotypes corresponding to rapid acetylation are a risk factor for smoking behavior, and the individuals who carry these haplotypes are 72~85% more likely to be smokers. As expected, the extended IPW regression and bias correction approaches performed very similarly, and the CO and AW methods provided a slightly larger OR estimation. It should be noted that based on the simulation studies, in the scenario where the primary disease is less common (5%) and the interactive effect of the secondary phenotype and SNP is highly protective (e.g., $\beta_{12} = -1.5041$), the extended IPW regression and bias correction approaches provided similarly accurate OR estimations, while the CO and AW methods slightly overestimated the OR.

Discussion

In this communication, we investigated several methods of estimating the OR for marker-secondary phenotype association in the special scenario where the secondary phenotype and the genetic variant have an interactive effect on the primary disease. We performed simulation studies to investigate the performance of the bias correction approach (Wang & Shete, 2011a) and compared it to the standard logistic regression methods (with and without adjusting for primary disease status), the CO approach, the AW method, and the extended IPW regression. We examined multiple scenarios with different primary disease study designs (unmatched and frequency-matched), different genetic model assumptions for the

genetic variant (additive, dominant, and recessive), different prevalences of the primary disease (less common to common), and different interactive effects of the secondary phenotype and genetic variant. In almost all scenarios, the logistic regression, the CO approach and AW methods performed poorly, with biased OR estimation and inflated type I errors. The extended IPW regression approach performed similarly to our bias correction approach when the original study of primary disease was an unmatched case-control study. However, when the original study of primary disease was frequency-matched on the secondary phenotype of interest, the extend IPW regression gave biased OR estimations with inflated type I errors in many situations. We observed a trend toward a higher magnitude of bias when the IPW regression approach was used to estimate the OR in studies where the interactive effect was deleterious. On the other hand, the simulation results show that our bias correction approach provides very accurate OR estimations of marker-secondary phenotype association in all scenarios.

To assess the significance of the corrected $\widehat{OR}$, we considered different approaches, including the Wald test based on asymptotic standard errors and a bootstrapping-based approach. Through simulation studies, we found that the Wald test is conservative for most of the scenarios and liberal in a few scenarios, while the bootstrapping approach can control the type I error rate better than all the other approaches studied. Therefore, we recommend using the bootstrapping approach to assess the significance of our bias correction approach. Although the bootstrapping approach is usually time-consuming, it is feasible with the recent advances in computational power.

In this study, we employed the standard IPW regression approach, where the weights were constructed on the basis of primary disease prevalence. For a study frequency-matched on the secondary phenotype, one may consider constructing the weights on the basis of the joint distribution of both primary disease and secondary phenotype in the strata defined by both phenotypes. This is beyond the focus of our current study, but could be an interesting topic for future investigation.

Our bias correction approach assesses the corrected OR for the marker-secondary phenotype association by solving a non-linear equation system with the use of iterative algorithms, provided that the prevalence values of primary disease and secondary phenotype are known. We performed simulations to investigate the sensitivity of our approach for misspecification of prevalence values. Our simulation results showed that the relatively small misspecifications of prevalence values (e.g., $\pm 1\%$) of primary disease and secondary phenotype do not have any major impact on the OR estimation. We further investigated the robustness of our approach to relatively larger misspecifications for prevalence values. We observed that the proposed approach is overall robust but could be slightly sensitive when the interaction effect was also very large. Even though the true prevalences of phenotypes are unknown, estimated prevalence values for many phenotypes can be obtained via a literature search or from nationwide survey reports (e.g., *Morbidity and Mortality Weekly* from the Centers for Disease Control). Also, for most of the common disorders, these prevalence values are estimated with very narrow confidence intervals. However, we have to admit that, if the estimates of prevalences deviate significantly from the true values, our approach might provide less accurate OR estimation. Also, our approach might not be appropriate if the prevalence values are not available for some traits (e.g., a specialized population).

In conclusion, the bias correction approach proposed in our original study is a robust approach, in both unmatched and frequency-matched study designs and with different primary disease prevalences (rare, less common and common), for studying the association

between the SNP and the secondary phenotype, even when there is an interactive effect of the SNP and secondary phenotype on the primary disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health grant 1R01CA131324 (SS) and by a faculty fellowship from The University of Texas MD Anderson Cancer Center Duncan Family Institute for Cancer Prevention and Risk Assessment (JW).

Appendix I

Estimation of E_{ki}

Given a sample with N_1 primary disease cases and N_0 controls, the expected numbers of individuals in the sample with secondary phenotype random variable $T = k$ and SNP random variable $X = i$ can be obtained using following equation:

$$E_{ki} = E(n_{ki} | N_0, N_1) = \sum_j (N_j \times \Pr(T=k, X=i | Y=j)) \times \frac{N_j}{N} = \sum_j \frac{N_j^2}{N} p_{kij},$$

where $i = 0, 1$, and 2 , and $k = 0, 1$. The conditional probability p_{kij} is written as

$$p_{kij} = \frac{\Pr(Y=j | T=k, X=i) \Pr(T=k | X=i) \Pr(X=i)}{\Pr(Y=j)} = \frac{p_{jki} p_{ki} p_i}{q_j},$$

for $i = 0, 1$, and 2 , and $j, k = 0, 1$.

Note that the conditional probabilities p_{kij} and p_{jki} were given as functions of regression coefficients, as shown in Equations (1) and (2), respectively. Note that the interactive effect of secondary phenotype and SNP is now embedded in Equation (2). The probabilities p_i , $i = 0, 1$, and 2 represent the genotype frequencies, and q_1 gives the prevalence of the primary disease (f_D) in the general population.

When the original case-control study of primary disease is frequency-matched on the secondary phenotype of interest, the expected numbers of individual E_{ki} need to be calculated differently:

$$\begin{aligned} E_{ki} &= E_{ki|1} + E_{ki|0} = E(n_{ki} | N_1) + E(n_{ki} | N_0) \\ &= (N_1 \times \Pr(T=k, X=i | Y=1)) \times \frac{N_1}{N} + (N_0 \times \Pr(X=i | T=k, Y=0) \times \Pr(T=k | Y=1)) \times \frac{N_0}{N}, \\ &= \frac{N_1^2}{N} p_{kij} + \frac{N_0^2}{N} h_{ik0} h_{k1} \end{aligned}$$

for $i = 0, 1$, and 2 , and $k = 0, 1$.

The conditional probabilities h_{ik0} h_{k1} can be written as

$$h_{i|k0}h_{k|1} = \frac{p_{0|ki}p_{k|i}p_i \times \sum_l p_{1|kl}p_{k|l}p_l}{f_D \times \sum_l p_{0|kl}p_{k|l}p_l}, \quad \text{for } i=0, 1, \text{ and } 2, \text{ and } k=0, 1.$$

Appendix II

Estimation of Standard Error $\tilde{s}$ for Additive Genetic Model of SNP

Denote random variable $\tilde{\alpha}_1^1 = \log(\overline{OR}_1)$, and $\tilde{\alpha}_1^2 = \log(\overline{OR}_2)$. The standard error of $\log(\overline{OR})$ can be written as:

$$\begin{aligned} \tilde{s} &= \sqrt{\text{var}\left(\frac{\tilde{\alpha}_1^1 + \tilde{\alpha}_1^2}{2}\right)} \\ &= \sqrt{\frac{1}{4}\left(\text{var}(\tilde{\alpha}_1^1) + \frac{1}{4}\text{var}(\tilde{\alpha}_1^2) + 2 \times \frac{1}{2} \times \text{cov}(\tilde{\alpha}_1^1, \tilde{\alpha}_1^2)\right)}, \\ &= \frac{1}{2} \sqrt{\tilde{s}_1^2 + \tilde{s}_2^2/4 + \tilde{c}_{12}} \end{aligned}$$

Where $\tilde{s}_1$ and $\tilde{s}_2$, are the standard errors of random variables $\tilde{\alpha}_1^1$ and $\tilde{\alpha}_1^2$, and $\tilde{c}_{12}$ is the covariance between them. Conditional on the sample with N_I primary disease cases and N_0 controls, the asymptotic standard errors $\tilde{s}_1$ and $\tilde{s}_2$ can be evaluated as below

$$\begin{aligned} \tilde{s}_1 &= \sqrt{\frac{1}{E_{11}} + \frac{1}{E_{00}} + \frac{1}{E_{01}} + \frac{1}{E_{10}}}, \text{ and} \\ \tilde{s}_2 &= \sqrt{\frac{1}{E_{12}} + \frac{1}{E_{00}} + \frac{1}{E_{02}} + \frac{1}{E_{10}}}. \end{aligned}$$

The covariance $\tilde{c}_{12}$ can be evaluated using the delta-method. Denote the total sample size as N . Let the probabilities of observing n_{ki} numbers of individuals in the sample be p_{ki} , $i = 0, 1$, and 2 , and $k = 0, 1$. Define $Z_{ki} = (n_{ki} - Np_{ki})/\sqrt{N}$ for $i = 0, 1$, and 2 , and $k = 0, 1$. As N becomes larger, Z_{ki} has an approximately normal distribution. Therefore, the covariance of $\tilde{\alpha}_1^1$ and $\tilde{\alpha}_1^2$ can be written as

$$\tilde{c}_{12} = \text{cov}\left(\frac{1}{\sqrt{N}}\left(\frac{Z_{11}}{p_{11}} + \frac{Z_{00}}{p_{00}} - \frac{Z_{10}}{p_{10}} - \frac{Z_{01}}{p_{01}}\right), \frac{1}{\sqrt{N}}\left(\frac{Z_{12}}{p_{12}} + \frac{Z_{00}}{p_{00}} - \frac{Z_{10}}{p_{10}} - \frac{Z_{02}}{p_{02}}\right)\right)$$

Given that $\text{cov}(Z_{rs}/p_{rs}, Z_{uv}/p_{uv}) = \begin{cases} (1 - p_{rs})/p_{rs}, & r=u \text{ and } s=v \\ -1, & \text{otherwise} \end{cases}$, we can derive that $\tilde{c}_{12} = 1/\sqrt{N}(1/p_{00} + 1/p_{11}) = 1/n_{00} + 1/n_{11} = 1/E_{00} + 1/E_{11}$. The expected numbers of individuals in the sample E_{ki} can be evaluated using the formulas shown in Appendix I.

References

- CDC. Cigarette smoking among adults --- United States, 2002. MMWR. 2004; 53:427–431. [PubMed: 15163928]
- CDC. Cigarette smoking among adults --- United States, 2004. MMWR. 2005; 54:1121–1124. [PubMed: 16280969]
- Cohen R, Martinez M. Early Release of Selected Estimates Based on Data From the January-March 2010 National Health Interview Survey. National Center for Health Statistics . 2010

- Efron, B.; Tibshirani, R.J. An introduction to the bootstrap. Chapman and Hall; New York: 1993.
- Erhardt JG, Kreichgauer HP, Meisner C, Bode JC, Bode C. Alcohol, cigarette smoking, dietary factors and the risk of colorectal adenomas and hyperplastic polyps--a case control study. *Eur J Nutr.* 2002; 41:35–43. [PubMed: 11990006]
- Giovannucci E. An updated review of the epidemiological evidence that cigarette smoking increases risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev.* 2001; 10:725–731. [PubMed: 11440957]
- Ji BT, Weissfeld JL, Chow WH, Huang WY, Schoen RE, Hayes RB. Tobacco smoking and colorectal hyperplastic and adenomatous polyps. *Cancer Epidemiol Biomarkers Prev.* 2006; 15:897–901. [PubMed: 16702367]
- Li H, Gail MH, 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. *Genet Epidemiol.* 2010; 34:427–433. [PubMed: 20583284]
- Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol.* 2009; 33:256–265. [PubMed: 19051285]
- Monsees GM, Tamimi RM, Kraft P. Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol.* 2009; 33:717–728. [PubMed: 19365863]
- Moorman JE, Zahran H, Truman BI, Molla MT. Current asthma prevalence - United States, 2006-2008. *MMWR Surveill Summ.* 2011; 60(Suppl):84–86. [PubMed: 21430629]
- Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, Weissfeld J, Hein DW, Hayes RB. Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics.* 2006; 7:819–829. [PubMed: 16981843]
- Richardson DB, Rzehak P, Klenk J, Weiland SK. Analyses of case-control data for additional outcomes. *Epidemiology.* 2007; 18:441–445. [PubMed: 17473707]
- Rothman, KJ.; Greenland, S. Modern epidemiology. Lippincott Williams & Wilkins; Philadelphia, PA: 1998.
- Wang J, Shete S. Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet Epidemiol.* 2011a; 35:190–200. [PubMed: 21308766]
- Wang J, Shete S. Power and type I error results for a bias-correction approach recently shown to provide accurate odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet Epidemiol.* 2011b; 35:739–743. [PubMed: 21769937]

Table 1

Dominant genetic model. Median ORs and 95% confidence intervals (CI) estimated using logistic regressions (without and with adjustment of primary disease, i.e., $\widehat{OR}$ and $\widehat{OR}_{adj}$), CO approach (i.e., OR_{CO}), AW method (i.e., OR_{AW}), IPW regression (i.e., OR_{IPW}), and the bias correction approach (i.e., $\widetilde{OR}_m$) for the study of SNP-secondary phenotype association. Data were simulated under null and alternative hypotheses ($OR=1$ and 1.5). All the results were based on 1000 replicates, each with 3000 cases and 3000 frequency-matched controls.

Parameter		Median OR (95% CI)				
f_D	β_{12}	$\widehat{OR}$	$\widehat{OR}_{adj}$	OR_{CO}	OR_{AW}	OR_{IPW}
True OR = 1.00						
5%	-1.0986	0.60(0.51-0.72)	0.60(0.51-0.72)	1.06(0.82-1.37)	1.05(0.81-1.36)	1.00(0.78-1.27)
5%	-0.4055	0.84(0.72-0.98)	0.84(0.72-0.98)	1.02(0.82-1.28)	1.00(0.80-1.26)	1.00(0.81-1.24)
5%	0.0000	0.99(0.86-1.15)	1.00(0.86-1.15)	0.99(0.81-1.22)	0.99(0.83-1.19)	0.99(0.82-1.21)
5%	0.4055	1.16(1.01-1.32)	1.16(1.01-1.33)	0.96(0.79-1.16)	0.97(0.80-1.17)	0.98(0.82-1.17)
5%	1.0986	1.37(1.21-1.55)	1.38(1.22-1.56)	0.87(0.74-1.03)	0.87(0.74-1.03)	0.91(0.78-1.06)
10%	-1.0986	0.64(0.54-0.76)	0.64(0.54-0.76)	1.12(0.86-1.45)	1.11(0.86-1.44)	1.00(0.79-1.25)
10%	-0.4055	0.85(0.73-0.99)	0.85(0.72-0.99)	1.04(0.83-1.30)	1.03(0.82-1.29)	1.00(0.82-1.23)
10%	0.0000	0.99(0.85-1.15)	0.99(0.85-1.15)	0.99(0.81-1.22)	0.99(0.82-1.19)	0.99(0.82-1.19)
10%	0.4055	1.11(0.96-1.27)	1.10(0.96-1.27)	0.91(0.76-1.10)	0.92(0.76-1.12)	0.95(0.80-1.13)
10%	1.0986	1.22(1.08-1.39)	1.23(1.08-1.39)	0.77(0.65-0.91)	0.77(0.65-0.91)	0.84(0.72-0.98)
30%	-1.0986	0.79(0.66-0.94)	0.79(0.66-0.94)	1.37(1.05-1.80)	1.36(1.04-1.79)	0.98(0.81-1.18)
30%	-0.4055	0.93(0.79-1.10)	0.93(0.79-1.10)	1.14(0.90-1.45)	1.12(0.88-1.43)	1.01(0.85-1.21)
30%	0.0000	0.98(0.84-1.14)	0.98(0.84-1.14)	0.98(0.79-1.21)	0.98(0.81-1.18)	0.98(0.83-1.16)
30%	0.4055	0.97(0.84-1.13)	0.97(0.84-1.13)	0.81(0.66-0.98)	0.81(0.66-1.00)	0.90(0.77-1.05)
30%	1.0986	0.89(0.78-1.01)	0.89(0.78-1.01)	0.54(0.45-0.65)	0.54(0.45-0.65)	0.72(0.63-0.84)
True OR = 1.50						
5%	-1.0986	0.89(0.75-1.08)	0.89(0.74-1.07)	1.61(1.21-2.12)	1.59(1.19-2.11)	1.50(1.15-1.95)
5%	-0.4055	1.26(1.07-1.48)	1.26(1.06-1.48)	1.56(1.23-1.97)	1.53(1.19-1.95)	1.52(1.22-1.90)
5%	0.0000	1.49(1.27-1.74)	1.49(1.27-1.74)	1.49(1.20-1.84)	1.49(1.23-1.80)	1.49(1.21-1.83)
5%	0.4055	1.72(1.49-1.99)	1.73(1.49-1.99)	1.44(1.18-1.74)	1.45(1.19-1.77)	1.46(1.21-1.76)
5%	1.0986	2.02(1.77-2.30)	2.03(1.79-2.32)	1.31(1.11-1.55)	1.31(1.11-1.55)	1.36(1.16-1.60)
10%	-1.0986	0.94(0.78-1.13)	0.94(0.78-1.13)	1.68(1.26-2.23)	1.66(1.24-2.21)	1.48(1.15-1.89)
10%	-0.4055	1.28(1.09-1.51)	1.28(1.09-1.52)	1.58(1.25-2.01)	1.55(1.21-2.00)	1.52(1.23-1.88)
						1.51(1.20-1.92)

Parameter		Median OR (95% CI)				
f_D	β_{12}	$\hat{OR}$	$\hat{OR}_{adj}$	OR_{CO}	OR_{AW}	OR_{PW}
						$\sim OR_m$
10%	0.0000	1.48(1.27-1.73)	1.48(1.27-1.73)	1.47(1.19-1.83)	1.48(1.21-1.80)	1.48(1.21-1.80)
10%	0.4055	1.65(1.43-1.91)	1.66(1.43-1.92)	1.37(1.13-1.67)	1.38(1.13-1.69)	1.42(1.18-1.70)
10%	1.0986	1.80(1.58-2.05)	1.81(1.58-2.06)	1.16(0.98-1.37)	1.16(0.98-1.38)	1.26(1.07-1.47)
30%	-1.0986	1.16(0.96-1.41)	1.16(0.96-1.41)	2.09(1.55-2.82)	2.07(1.53-2.80)	1.45(1.18-1.78)
30%	-0.4055	1.40(1.18-1.66)	1.40(1.18-1.66)	1.72(1.34-2.21)	1.68(1.30-2.18)	1.52(1.27-1.83)
30%	0.0000	1.45(1.24-1.71)	1.45(1.24-1.71)	1.45(1.16-1.83)	1.45(1.18-1.79)	1.46(1.22-1.73)
30%	0.4055	1.45(1.25-1.69)	1.46(1.25-1.69)	1.21(0.98-1.48)	1.22(0.99-1.51)	1.34(1.14-1.58)
30%	1.0986	1.29(1.13-1.49)	1.30(1.13-1.49)	0.81(0.67-0.97)	0.81(0.68-0.97)	1.06(0.91-1.23)
						1.51(1.24-1.83)

Table 2

Additive genetic model. Median ORs and 95% confidence intervals (CI) estimated using logistic regressions (without and with adjustment of primary disease, i.e., $\widehat{OR}$ and $\widehat{OR}_{adj}$), CO approach (i.e., OR_{CO}), AW method (i.e., OR_{AW}), IPW regression (i.e., OR_{IPW}), and the bias correction approach (i.e., $\widetilde{OR}_m$) for the study of SNP-secondary phenotype association. Data were simulated under null and alternative hypotheses ($OR=1$ and 1.5). All the results were based on 1000 replicates, each with 3000 cases and 3000 frequency-matched controls.

Parameter		Median OR (95% CI)				
f_D	β_{12}	$\widehat{OR}$	$\widehat{OR}_{adj}$	OR_{CO}	OR_{AW}^*	OR_{IPW}
True OR = 1.00						
5%	-1.0986	0.63(0.55-0.72)	0.63(0.55-0.72)	1.04(0.87-1.26)	-	0.99(0.83-1.18)
5%	-0.4055	0.84(0.75-0.94)	0.84(0.75-0.94)	1.02(0.87-1.20)	-	1.00(0.86-1.16)
5%	0.0000	0.99(0.90-1.10)	0.99(0.90-1.10)	0.99(0.86-1.14)	-	0.99(0.87-1.13)
5%	0.4055	1.16(1.06-1.27)	1.16(1.06-1.27)	0.95(0.83-1.08)	-	0.97(0.86-1.09)
5%	1.0986	1.35(1.26-1.46)	1.37(1.27-1.48)	0.81(0.73-0.91)	-	0.86(0.78-0.95)
10%	-1.0986	0.66(0.57-0.75)	0.66(0.57-0.75)	1.10(0.91-1.32)	-	0.99(0.84-1.17)
10%	-0.4055	0.85(0.76-0.96)	0.85(0.76-0.95)	1.04(0.89-1.22)	-	1.00(0.87-1.16)
10%	0.0000	0.99(0.89-1.09)	0.99(0.89-1.09)	0.99(0.85-1.14)	-	0.99(0.87-1.12)
10%	0.4055	1.11(1.01-1.21)	1.11(1.01-1.21)	0.90(0.79-1.03)	-	0.94(0.84-1.06)
10%	1.0986	1.18(1.09-1.27)	1.18(1.09-1.28)	0.70(0.62-0.79)	-	0.78(0.71-0.87)
30%	-1.0986	0.79(0.69-0.90)	0.79(0.69-0.90)	1.31(1.09-1.57)	-	0.96(0.84-1.11)
30%	-0.4055	0.93(0.83-1.05)	0.93(0.83-1.05)	1.14(0.97-1.33)	-	1.01(0.89-1.14)
30%	0.0000	0.98(0.88-1.08)	0.97(0.88-1.08)	0.98(0.84-1.14)	-	0.98(0.87-1.10)
30%	0.4055	0.97(0.88-1.07)	0.97(0.88-1.07)	0.79(0.68-0.91)	-	0.90(0.81-1.00)
30%	1.0986	0.87(0.79-0.96)	0.87(0.79-0.95)	0.50(0.43-0.58)	-	0.71(0.65-0.79)
True OR = 1.50						
5%	-1.0986	0.92(0.81-1.06)	0.92(0.81-1.06)	1.57(1.30-1.90)	-	1.49(1.24-1.78)
5%	-0.4055	1.25(1.12-1.40)	1.25(1.12-1.40)	1.53(1.31-1.80)	-	1.50(1.29-1.75)
5%	0.0000	1.49(1.35-1.65)	1.50(1.35-1.65)	1.50(1.30-1.72)	-	1.50(1.31-1.71)
5%	0.4055	1.73(1.58-1.88)	1.74(1.59-1.90)	1.43(1.26-1.62)	-	1.46(1.30-1.64)
5%	1.0986	1.94(1.80-2.10)	2.00(1.85-2.16)	1.21(1.09-1.35)	-	1.27(1.15-1.41)
10%	-1.0986	0.96(0.84-1.11)	0.96(0.84-1.11)	1.64(1.35-1.98)	-	1.48(1.24-1.75)
10%	-0.4055	1.27(1.14-1.43)	1.27(1.14-1.42)	1.56(1.33-1.83)	-	1.50(1.29-1.73)
						1.49(1.21-1.83)

Parameter		Median OR (95% CI)				
f_D	β_{12}	$\hat{OR}$	$\hat{OR}_{adj}$	OR_{CO}	OR_{AW}^*	$\widetilde{OR}_m$
10%	0.0000	1.48(1.34-1.64)	1.49(1.35-1.64)	1.48(1.29-1.71)	-	1.48(1.31-1.68)
10%	0.4055	1.64(1.50-1.79)	1.65(1.51-1.80)	1.35(1.19-1.53)	-	1.40(1.25-1.57)
10%	1.0986	1.68(1.56-1.82)	1.71(1.58-1.86)	1.04(0.93-1.17)	-	1.15(1.04-1.27)
30%	-1.0986	1.14(1.00-1.31)	1.14(1.00-1.31)	1.94(1.60-2.36)	-	1.41(1.22-1.63)
30%	-0.4055	1.39(1.24-1.55)	1.39(1.24-1.56)	1.70(1.45-2.00)	-	1.50(1.33-1.70)
30%	0.0000	1.47(1.32-1.63)	1.47(1.32-1.63)	1.46(1.26-1.70)	-	1.47(1.31-1.64)
30%	0.4055	1.43(1.30-1.58)	1.44(1.31-1.59)	1.18(1.02-1.35)	-	1.33(1.20-1.47)
30%	1.0986	1.24(1.13-1.36)	1.24(1.14-1.36)	0.74(0.64-0.84)	-	1.02(0.93-1.12)

* AW method does not allow additive genetic modeling

Table 3

Recessive genetic model. Median ORs and 95% confidence intervals (CI) estimated using logistic regressions (without and with adjustment of primary disease, i.e., $\widehat{OR}$ and $\widehat{OR}_{adj}$), CO approach (i.e., OR_{CO}), AW method (i.e., OR_{AW}), IPW regression (i.e., OR_{IPW}), and bias correction approach (i.e., $\widetilde{OR}_m$) for the study of SNP-secondary phenotype association. Data were simulated under null and alternative hypotheses ($OR=1$ and 1.5). All the results were based on 1000 replicates, each with 3000 cases and 3000 frequency-matched controls.

Parameter		Median OR (95% CI)				
f_D	β_{12}	$\widehat{OR}$	$\widehat{OR}_{adj}$	OR_{CO}	OR_{AW}	OR_{IPW}
True OR = 1.00						
5%	-1.0986	0.66(0.53-0.82)	0.66(0.53-0.82)	1.06(0.80-1.40)	1.05(0.80-1.39)	1.01(0.78-1.32)
5%	-0.4055	0.84(0.69-1.03)	0.84(0.69-1.03)	1.02(0.78-1.35)	1.00(0.76-1.33)	1.00(0.77-1.30)
5%	0.0000	0.99(0.82-1.19)	0.99(0.82-1.19)	0.99(0.76-1.30)	0.99(0.78-1.25)	0.99(0.77-1.28)
5%	0.4055	1.19(1.01-1.42)	1.19(1.00-1.42)	0.95(0.73-1.24)	0.98(0.74-1.29)	0.98(0.76-1.25)
5%	1.0986	1.62(1.40-1.89)	1.63(1.40-1.90)	0.85(0.66-1.11)	0.86(0.67-1.12)	0.93(0.73-1.17)
10%	-1.0986	0.69(0.56-0.86)	0.69(0.56-0.86)	1.12(0.84-1.47)	1.11(0.84-1.46)	1.02(0.79-1.32)
10%	-0.4055	0.86(0.71-1.05)	0.86(0.71-1.05)	1.04(0.80-1.37)	1.02(0.77-1.35)	1.01(0.78-1.29)
10%	0.0000	0.99(0.82-1.20)	0.99(0.82-1.20)	0.99(0.75-1.30)	0.99(0.78-1.26)	0.99(0.77-1.26)
10%	0.4055	1.13(0.95-1.35)	1.13(0.95-1.35)	0.89(0.68-1.17)	0.92(0.69-1.22)	0.94(0.74-1.19)
10%	1.0986	1.43(1.22-1.68)	1.43(1.22-1.68)	0.74(0.56-0.97)	0.75(0.56-0.99)	0.87(0.70-1.09)
30%	-1.0986	0.87(0.70-1.07)	0.87(0.70-1.07)	1.40(1.06-1.83)	1.39(1.05-1.82)	1.06(0.85-1.33)
30%	-0.4055	0.94(0.77-1.15)	0.94(0.76-1.15)	1.14(0.87-1.52)	1.12(0.84-1.51)	1.02(0.82-1.27)
30%	0.0000	0.98(0.80-1.20)	0.98(0.80-1.19)	0.97(0.73-1.30)	0.97(0.76-1.26)	0.98(0.79-1.22)
30%	0.4055	1.00(0.83-1.22)	1.00(0.83-1.22)	0.80(0.59-1.09)	0.83(0.60-1.14)	0.93(0.75-1.14)
30%	1.0986	1.02(0.84-1.24)	1.02(0.85-1.24)	0.52(0.37-0.73)	0.53(0.37-0.75)	0.82(0.67-1.01)
True OR = 1.50						
5%	-1.0986	0.97(0.80-1.18)	0.97(0.80-1.18)	1.57(1.22-2.03)	1.56(1.21-2.02)	1.50(1.17-1.92)
5%	-0.4055	1.26(1.05-1.51)	1.26(1.05-1.51)	1.53(1.19-1.97)	1.50(1.16-1.95)	1.50(1.18-1.91)
5%	0.0000	1.49(1.26-1.77)	1.49(1.26-1.77)	1.49(1.17-1.91)	1.49(1.20-1.86)	1.49(1.18-1.88)
5%	0.4055	1.78(1.52-2.09)	1.78(1.52-2.09)	1.43(1.13-1.82)	1.46(1.14-1.87)	1.46(1.17-1.83)
5%	1.0986	2.39(2.07-2.75)	2.40(2.08-2.77)	1.28(1.02-1.61)	1.29(1.02-1.63)	1.38(1.12-1.71)
10%	-1.0986	1.04(0.85-1.26)	1.04(0.85-1.26)	1.69(1.31-2.19)	1.68(1.30-2.18)	1.54(1.21-1.96)
10%	-0.4055	1.28(1.07-1.54)	1.28(1.07-1.54)	1.56(1.22-2.01)	1.54(1.19-1.99)	1.50(1.20-1.89)
True OR = 1.50						
5%	-1.0986	0.97(0.80-1.18)	0.97(0.80-1.18)	1.57(1.22-2.03)	1.56(1.21-2.02)	1.50(1.17-1.92)
5%	-0.4055	1.26(1.05-1.51)	1.26(1.05-1.51)	1.53(1.19-1.97)	1.50(1.16-1.95)	1.50(1.18-1.91)
5%	0.0000	1.49(1.26-1.77)	1.49(1.26-1.77)	1.49(1.17-1.91)	1.49(1.20-1.86)	1.49(1.18-1.88)
5%	0.4055	1.78(1.52-2.09)	1.78(1.52-2.09)	1.43(1.13-1.82)	1.46(1.14-1.87)	1.46(1.17-1.83)
5%	1.0986	2.39(2.07-2.75)	2.40(2.08-2.77)	1.28(1.02-1.61)	1.29(1.02-1.63)	1.38(1.12-1.71)
10%	-1.0986	1.04(0.85-1.26)	1.04(0.85-1.26)	1.69(1.31-2.19)	1.68(1.30-2.18)	1.54(1.21-1.96)
10%	-0.4055	1.28(1.07-1.54)	1.28(1.07-1.54)	1.56(1.22-2.01)	1.54(1.19-1.99)	1.50(1.20-1.89)

Parameter		Median OR (95% CI)					
f_D	β_{12}	$\hat{OR}$	$\hat{OR}_{adj}$	OR_{Co}	OR_{AW}	OR_{PW}	$\widetilde{OR}_m$
10%	0.0000	1.48(1.25-1.76)	1.48(1.25-1.76)	1.49(1.16-1.91)	1.49(1.19-1.86)	1.49(1.19-1.86)	1.51(1.19-1.92)
10%	0.4055	1.70(1.45-2.00)	1.71(1.45-2.00)	1.35(1.06-1.73)	1.38(1.07-1.78)	1.42(1.15-1.76)	1.49(1.19-1.87)
10%	1.0986	2.10(1.81-2.43)	2.10(1.81-2.44)	1.11(0.87-1.43)	1.13(0.88-1.44)	1.29(1.06-1.59)	1.51(1.22-1.85)
30%	-1.0986	1.28(1.05-1.56)	1.27(1.05-1.56)	2.10(1.63-2.71)	2.09(1.62-2.70)	1.57(1.27-1.94)	1.48(1.12-1.95)
30%	-0.4055	1.42(1.18-1.71)	1.42(1.18-1.71)	1.73(1.34-2.25)	1.70(1.30-2.23)	1.54(1.26-1.88)	1.51(1.16-1.97)
30%	0.0000	1.47(1.22-1.76)	1.47(1.23-1.76)	1.46(1.12-1.91)	1.46(1.16-1.86)	1.46(1.21-1.78)	1.50(1.16-1.94)
30%	0.4055	1.51(1.27-1.80)	1.51(1.27-1.81)	1.21(0.92-1.59)	1.24(0.93-1.65)	1.40(1.15-1.69)	1.52(1.18-1.95)
30%	1.0986	1.51(1.27-1.79)	1.51(1.27-1.80)	0.78(0.57-1.05)	0.79(0.58-1.07)	1.21(1.00-1.46)	1.51(1.19-1.93)

Table 4

Estimated type I error probabilities at nominal significance level $\alpha=0.05$. Different approaches were evaluated, including logistic regressions (without and with adjustment of primary disease), the CO approach, AW method, IPW regression, and bias correction approach for the study of SNP-secondary phenotype association. All the results were based on 1000 replicates, each with 3000 cases and 3000 frequency-matched controls.

Genetic Model	f_D	β_{12}	Type I Error ($\alpha=0.05$)					Bias Correction Approach		
			Logistic (without adjustment)	Logistic (with adjustment)	CO	AW*	IPW	Wald Test		Bootstrapping
Dominant Model	5%	-1.0986	1.000	1.000	0.082	0.073	0.057	0.077	0.055	
	5%	-0.4055	0.601	0.603	0.051	0.044	0.043	0.036	0.040	
	5%	0	0.053	0.053	0.062	0.043	0.061	0.040	0.060	
	5%	0.4055	0.537	0.539	0.080	0.079	0.074	0.023	0.046	
	5%	1.0986	1.000	1.000	0.389	0.381	0.238	0.020	0.054	
	10%	-1.0986	0.999	0.999	0.135	0.122	0.061	0.066	0.050	
	10%	-0.4055	0.533	0.532	0.068	0.057	0.050	0.037	0.049	
	10%	0	0.054	0.050	0.052	0.043	0.054	0.025	0.046	
	10%	0.4055	0.289	0.280	0.154	0.141	0.107	0.016	0.049	
	10%	1.0986	0.885	0.885	0.855	0.845	0.600	0.025	0.066	
	30%	-1.0986	0.754	0.753	0.650	0.618	0.061	0.041	0.050	
	30%	-0.4055	0.127	0.126	0.184	0.147	0.045	0.010	0.048	
Additive Model	30%	0	0.062	0.063	0.066	0.055	0.066	0.005	0.047	
	30%	0.4055	0.074	0.077	0.564	0.498	0.267	0.010	0.055	
	30%	1.0986	0.425	0.425	1.000	1.000	0.989	0.030	0.044	
	5%	-1.0986	1.000	1.000	0.062	-	0.047	0.035	0.048	
	5%	-0.4055	0.877	0.876	0.070	-	0.056	0.017	0.049	
	5%	0	0.057	0.056	0.061	-	0.062	0.019	0.055	
	5%	0.4055	0.916	0.916	0.113	-	0.067	0.013	0.038	
	5%	1.0986	1.000	1.000	0.948	-	0.814	0.017	0.062	
	10%	-1.0986	1.000	1.000	0.170	-	0.072	0.035	0.054	
	10%	-0.4055	0.801	0.802	0.079	-	0.056	0.008	0.050	
	10%	0	0.056	0.055	0.057	-	0.056	0.014	0.054	
	10%	0.4055	0.580	0.591	0.332	-	0.191	0.007	0.051	

Genetic Model	f_D	β_{12}	Type I Error ($\alpha=0.05$)				Bias Correction Approach		
			Logistic (without adjustment)	Logistic (with adjustment)	CO	AW*	IPW	Wald Test	Bootstrapping
Recessive Model	10%	1.0986	0.970	0.982	1.000	-	0.996	0.006	0.049
	30%	-1.0986	0.940	0.929	0.810	-	0.107	0.019	0.056
	30%	-0.4055	0.238	0.250	0.345	-	0.050	0.007	0.049
	30%	0	0.074	0.075	0.057	-	0.067	0.000	0.047
	30%	0.4055	0.102	0.088	0.902	-	0.520	0.006	0.043
	30%	1.0986	0.798	0.828	1.000	-	1.000	0.014	0.040
	5%	-1.0986	0.970	0.970	0.073	0.069	0.047	0.031	0.047
	5%	-0.4055	0.388	0.392	0.061	0.070	0.055	0.040	0.050
	5%	0	0.042	0.039	0.042	0.031	0.042	0.038	0.047
	5%	0.4055	0.525	0.523	0.057	0.055	0.045	0.043	0.047
	5%	1.0986	1.000	1.000	0.209	0.184	0.093	0.059	0.043
	10%	-1.0986	0.934	0.933	0.126	0.116	0.053	0.025	0.058
	10%	-0.4055	0.318	0.323	0.069	0.068	0.051	0.022	0.048
	10%	0	0.059	0.058	0.044	0.045	0.053	0.030	0.053
	10%	0.4055	0.274	0.275	0.099	0.070	0.062	0.032	0.050
	10%	1.0986	0.986	0.990	0.585	0.538	0.217	0.035	0.053
	30%	-1.0986	0.248	0.256	0.656	0.631	0.084	0.010	0.050
	30%	-0.4055	0.083	0.084	0.188	0.150	0.057	0.008	0.044
	30%	0	0.068	0.067	0.058	0.065	0.067	0.011	0.064
	30%	0.4055	0.046	0.040	0.294	0.190	0.104	0.006	0.040
	30%	1.0986	0.059	0.056	0.978	0.966	0.473	0.014	0.041

* AW method does not allow additive genetic modeling

Table 5

Sensitivity analysis results for the bias correction approach (i.e., $\widetilde{OR}_m$). Dominant genetic model was assumed. All the results were based on 1000 replicates, each with 3000 cases and 3000 frequency-matched controls. Data were simulated under null and alternative hypotheses (OR=1 and 1.5). Δ was fixed as 1%.

Parameter		$\widetilde{OR}_m$											
f_D	β_{12}	f_D	$f_{D+\Delta}$	$f_{D-\Delta}$	$f_{D+\Delta}$	$f_{D-\Delta}$	$f_{D+\Delta}$	$f_{D-\Delta}$	f_D	$f_{D+\Delta}$	$f_{D-\Delta}$	$f_{D+\Delta}$	$f_{D-\Delta}$
		f_T	$f_{T+\Delta}$	$f_{T-\Delta}$	$f_{T+\Delta}$	$f_{T-\Delta}$	$f_{T+\Delta}$	$f_{T-\Delta}$	f_T	$f_{T+\Delta}$	$f_{T-\Delta}$	$f_{T+\Delta}$	$f_{T-\Delta}$
True OR = 1.00													
5%	-1.0986	1.00	0.99	0.99	1.01	1.01	1.00	1.00	1.00	0.99	0.99	1.01	1.01
5%	-0.4055	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00
5%	0.0000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5%	0.4055	1.00	1.01	1.01	0.99	0.99	1.00	1.00	1.00	1.01	1.01	0.99	0.99
5%	1.0986	1.00	1.02	1.03	0.97	0.98	1.00	1.00	1.00	1.03	1.03	0.98	0.98
10%	-1.0986	1.00	0.99	0.99	1.01	1.01	1.00	1.00	1.00	0.99	0.99	1.01	1.01
10%	-0.4055	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00
10%	0.0000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10%	0.4055	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99
10%	1.0986	1.00	1.02	1.03	0.97	0.98	1.00	1.00	1.00	1.02	1.02	0.98	0.98
30%	-1.0986	0.99	0.98	0.98	1.00	1.00	0.99	0.99	0.99	0.98	0.98	1.00	1.00
30%	-0.4055	1.00	1.00	1.00	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.01	1.01
30%	0.0000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30%	0.4055	1.00	1.01	1.01	0.99	1.00	1.00	1.00	1.00	1.01	1.01	0.99	0.99
30%	1.0986	1.00	1.01	1.02	0.98	0.99	1.00	1.00	1.01	1.02	1.02	0.99	0.99
True OR = 1.50													
5%	-1.0986	1.51	1.49	1.49	1.53	1.53	1.51	1.51	1.51	1.49	1.49	1.53	1.53
5%	-0.4055	1.52	1.51	1.51	1.53	1.53	1.52	1.52	1.52	1.51	1.51	1.53	1.53
5%	0.0000	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
5%	0.4055	1.50	1.52	1.52	1.49	1.49	1.50	1.50	1.50	1.52	1.52	1.49	1.49
5%	1.0986	1.50	1.54	1.55	1.46	1.47	1.50	1.51	1.51	1.54	1.54	1.46	1.46
10%	-1.0986	1.49	1.48	1.48	1.51	1.51	1.49	1.49	1.49	1.48	1.48	1.51	1.51
10%	-0.4055	1.51	1.51	1.51	1.52	1.52	1.51	1.51	1.51	1.51	1.51	1.52	1.52

Parameter		$\widetilde{OR}_m$											
f_D	β_{12}	f_D	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	f_D	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$
		f_T	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	f_T	$f_T+\Delta$	f_T	f_T
10%	0.0000	1.49	1.50	1.50	1.50	1.49	1.49	1.49	1.49	1.49	1.50	1.50	1.49
10%	0.4055	1.50	1.51	1.51	1.51	1.48	1.49	1.49	1.50	1.50	1.51	1.51	1.49
10%	1.0986	1.50	1.53	1.54	1.54	1.46	1.48	1.48	1.50	1.51	1.54	1.54	1.47
30%	-1.0986	1.50	1.48	1.49	1.49	1.52	1.52	1.52	1.50	1.50	1.48	1.48	1.52
30%	-0.4055	1.51	1.51	1.51	1.51	1.52	1.52	1.52	1.51	1.51	1.51	1.51	1.52
30%	0.0000	1.50	1.50	1.50	1.50	1.49	1.49	1.49	1.50	1.50	1.50	1.50	1.49
30%	0.4055	1.50	1.51	1.51	1.51	1.49	1.49	1.49	1.50	1.50	1.51	1.51	1.49
30%	1.0986	1.51	1.52	1.54	1.54	1.48	1.49	1.49	1.50	1.51	1.53	1.53	1.49

Table 6

Sensitivity analysis results for the bias correction approach (i.e., $\widetilde{OR}_m$). Additive genetic model was assumed. All the results were based on 1000 replicates, each with 3000 cases and 3000 frequency-matched controls. Data were simulated under null and alternative hypotheses (OR=1 and 1.5). Δ was fixed as 1%.

Parameter		$\widetilde{OR}_m$											
f_D	β_{12}	f_D	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$
		f_T	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$
True OR = 1.00													
5%	-1.0986	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
5%	-0.4055	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
5%	0.0000	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99
5%	0.4055	1.00	1.01	1.01	0.99	0.99	1.00	1.00	1.00	1.00	1.01	1.01	0.99
5%	1.0986	1.00	1.03	1.04	0.97	0.97	1.00	1.01	1.01	1.01	1.03	1.03	0.97
10%	-1.0986	1.00	0.99	0.99	1.01	1.01	1.00	1.00	1.00	1.00	0.99	1.01	1.01
10%	-0.4055	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
10%	0.0000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10%	0.4055	1.00	1.01	1.01	0.99	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.00
10%	1.0986	1.00	1.02	1.03	0.97	0.97	1.00	1.01	1.01	1.01	1.02	1.02	0.98
30%	-1.0986	1.00	0.99	0.99	1.01	1.01	1.00	1.00	1.00	1.00	0.99	1.01	1.01
30%	-0.4055	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
30%	0.0000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30%	0.4055	1.00	1.01	1.01	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.00
30%	1.0986	1.00	1.01	1.02	0.99	1.00	1.00	1.00	1.01	1.01	1.02	1.02	0.99
True OR = 1.50													
5%	-1.0986	1.51	1.49	1.49	1.52	1.53	1.51	1.51	1.51	1.51	1.49	1.52	1.52
5%	-0.4055	1.51	1.50	1.50	1.51	1.51	1.51	1.51	1.51	1.51	1.50	1.51	1.51
5%	0.0000	1.51	1.51	1.51	1.50	1.50	1.51	1.51	1.51	1.51	1.51	1.51	1.50
5%	0.4055	1.51	1.52	1.52	1.49	1.50	1.51	1.51	1.51	1.51	1.52	1.52	1.49
5%	1.0986	1.50	1.53	1.55	1.45	1.46	1.49	1.51	1.51	1.51	1.54	1.54	1.46
10%	-1.0986	1.51	1.49	1.49	1.52	1.52	1.51	1.51	1.51	1.51	1.49	1.52	1.52
10%	-0.4055	1.49	1.48	1.48	1.49	1.49	1.49	1.49	1.49	1.49	1.48	1.48	1.49

Parameter		$\widetilde{OR}_m$											
f_D	β_{12}	f_D	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	f_D	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$
		f_T	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	f_T	$f_T+\Delta$	f_T	f_T
10%	0.0000	1.50	1.51	1.51	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.51	1.50
10%	0.4055	1.50	1.51	1.51	1.49	1.49	1.49	1.50	1.50	1.50	1.50	1.51	1.49
10%	1.0986	1.50	1.53	1.55	1.46	1.48	1.48	1.49	1.51	1.51	1.54	1.54	1.47
30%	-1.0986	1.50	1.48	1.49	1.52	1.52	1.52	1.50	1.50	1.50	1.49	1.51	1.52
30%	-0.4055	1.50	1.49	1.49	1.51	1.51	1.51	1.50	1.50	1.50	1.49	1.51	1.51
30%	0.0000	1.51	1.51	1.51	1.51	1.51	1.51	1.51	1.51	1.51	1.51	1.51	1.51
30%	0.4055	1.50	1.51	1.51	1.49	1.50	1.50	1.50	1.50	1.51	1.51	1.51	1.50
30%	1.0986	1.50	1.51	1.53	1.48	1.49	1.49	1.50	1.51	1.51	1.52	1.52	1.49

Table 7

Sensitivity analysis results for the bias correction approach (i.e., $\widetilde{OR}_m$). Recessive genetic model was assumed. All the results were based on 1000 replicates, each with 3000 cases and 3000 frequency-matched controls. Data were simulated under null and alternative hypotheses (OR=1 and 1.5). Δ was fixed as 1%.

Parameter		$\widetilde{OR}_m$											
f_D	β_{12}	f_D	$f_{D+\Delta}$	$f_{D-\Delta}$	f_T	$f_{T+\Delta}$	$f_{T-\Delta}$	f_D	$f_{D+\Delta}$	$f_{D-\Delta}$	f_T	$f_{T+\Delta}$	$f_{T-\Delta}$
True OR = 1.00													
5%	-1.0986	1.00	0.99	0.99	1.01	1.01	1.01	1.00	1.00	1.00	0.99	0.99	1.01
5%	-0.4055	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00
5%	0.0000	1.00	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99
5%	0.4055	1.00	1.01	1.01	0.99	0.99	0.99	1.00	1.00	1.00	1.01	1.01	0.99
5%	1.0986	1.00	1.03	1.04	0.97	0.98	1.00	1.01	1.01	1.01	1.03	1.03	0.98
10%	-1.0986	0.99	0.98	0.98	1.00	1.00	1.00	0.99	0.99	0.99	0.98	0.98	1.00
10%	-0.4055	1.00	1.00	1.00	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.01
10%	0.0000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10%	0.4055	0.98	0.99	0.99	0.97	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.98
10%	1.0986	1.00	1.02	1.03	0.98	0.98	1.00	1.00	1.00	1.00	1.03	1.03	0.98
30%	-1.0986	1.00	0.99	0.99	1.01	1.01	1.01	1.00	1.00	1.00	0.99	1.01	1.01
30%	-0.4055	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30%	0.0000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30%	0.4055	1.00	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.00
30%	1.0986	1.00	1.01	1.02	0.99	0.99	1.00	1.00	1.01	1.01	1.02	0.99	0.99
True OR = 1.50													
5%	-1.0986	1.48	1.46	1.46	1.49	1.49	1.49	1.48	1.48	1.48	1.46	1.49	1.49
5%	-0.4055	1.50	1.49	1.49	1.50	1.50	1.50	1.50	1.50	1.50	1.49	1.50	1.50
5%	0.0000	1.50	1.51	1.51	1.50	1.50	1.50	1.50	1.50	1.50	1.51	1.50	1.50
5%	0.4055	1.50	1.52	1.52	1.49	1.49	1.49	1.50	1.50	1.50	1.52	1.49	1.49
5%	1.0986	1.51	1.55	1.56	1.46	1.47	1.51	1.51	1.52	1.52	1.56	1.47	1.47
10%	-1.0986	1.49	1.48	1.48	1.51	1.51	1.51	1.49	1.49	1.49	1.48	1.51	1.51
10%	-0.4055	1.49	1.49	1.49	1.50	1.50	1.50	1.49	1.49	1.49	1.49	1.49	1.50

Parameter		$\widetilde{OR}_m$											
f_D	β_{12}	f_D	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$	f_D	$f_D+\Delta$	$f_D+\Delta$	$f_D+\Delta$
		f_T	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	$f_T+\Delta$	f_T	$f_T+\Delta$	f_T	f_T
10%	0.0000	1.51	1.51	1.51	1.50	1.50	1.50	1.51	1.51	1.51	1.51	1.51	1.50
10%	0.4055	1.49	1.50	1.51	1.48	1.48	1.48	1.49	1.49	1.49	1.50	1.50	1.48
10%	1.0986	1.51	1.54	1.55	1.46	1.47	1.47	1.50	1.51	1.51	1.54	1.54	1.47
30%	-1.0986	1.48	1.47	1.46	1.50	1.50	1.50	1.48	1.48	1.48	1.46	1.46	1.50
30%	-0.4055	1.51	1.51	1.51	1.52	1.52	1.52	1.51	1.51	1.51	1.51	1.51	1.52
30%	0.0000	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
30%	0.4055	1.52	1.52	1.53	1.51	1.51	1.51	1.52	1.52	1.52	1.53	1.53	1.51
30%	1.0986	1.51	1.53	1.54	1.49	1.49	1.50	1.51	1.51	1.52	1.53	1.53	1.49

Table 8

Results from the analysis of the association between gene *NAT2* and smoking status based on colorectal adenoma case-control data.

Approach	Odds ratio	95% Confidence interval	p value
Logistic (without adjustment)	0.8375	(0.5040, 1.3916)	0.4936
Logistic (with adjustment)	0.8418	(0.5057, 1.4014)	0.5078
CO	1.8502	(0.8648, 3.9582)	0.1128
AW	1.7664	(0.8093, 3.8551)	0.1531
IPW	1.7212	(0.8403, 3.5211)	0.1378
Bias correction approach	1.7377	(0.8406, 3.5922) *	0.1359 *
		(0.4128, 2.0942) **	0.1738 **

* Obtained based on the Wald test

** Obtained based on the bootstrapping approach