Published in final edited form as:
Genet Epidemiol. 2010 July ; 34(5): 427–433. doi:10.1002/gepi.20495.

Using Cases to Strengthen Inference on the Association between Single Nucleotide Polymorphisms and a Secondary Phenotype in Genome-Wide Association Studies

Huilin Li, **Mitchell H. Gail**, **Sonja Berndt**, and **Nilanjan Chatterjee**
Division of Cancer Epidemiology and Genetics, National Cancer Institute, Executive Plaza South, Bethesda, MD, 20982

Abstract

Case-control genome-wide association studies provide a vast amount of genetic information that may be used to investigate secondary phenotypes. We study the situation in which the primary disease is rare and the secondary phenotype and genetic markers are dichotomous. An analysis of the association between a genetic marker and the secondary phenotype based on controls only is valid, whereas standard methods that also use cases result in biased estimates and highly inflated type I error if there is an interaction between the secondary phenotype and the genetic marker on the risk of the primary disease. Here we present an adaptively weighted method that combines the case and control data to study the association, while reducing to the controls only analysis if there is strong evidence of an interaction. The possibility of such an interaction and the misleading results for standard methods, but not for the adaptively weighted or controls only approaches, are illustrated by data from a case-control study of colorectal adenoma, in which the secondary phenotype is smoking. Simulations and asymptotic theory indicate that the adaptively weighted method can reduce the mean square error for estimation with a pre-specified SNP and increase the power to discover a new association in a genome-wide study, compared to an analysis of controls only. Further experience with genome-wide studies is needed to determine when methods that assume no interaction and gain precision and power, thereby can be recommended, and when methods such as the adaptively weighted or controls only approaches are needed to guard against the possibility of non-zero interactions.

Keywords

adaptively weighted; case-control study; genome-wide association study; maximum likelihood; secondary phenotype

1 INTRODUCTION

The genome-wide association study (GWAS) is a powerful tool to identify genetic associations with a disease. A GWAS may also provide information on secondary phenotypes that are measured for all subjects. Added value could be gained from a GWAS by studying the association between genes and the secondary phenotypes.

When the disease in a GWAS is rare, the controls could be regarded as a random sample from the general population, and they could be used for estimating the association with a

Address for Huilin Li(Corresponding author): Huilin Li, Ph.D., Biostatistic Branch, DCEG, NCI, NIH, 6120 Executive Boulevard, Room 8034, Rockville, MD 20852-7242, Tel: (301) 435-2510, lih5@mail.nih.gov.

secondary phenotype. However, if the disease is associated with both the SNP and the secondary phenotype, including cases may introduce bias in the estimation of association, even when the disease is rare.

Lin and Zeng [2009] studied a likelihood-based method that combines the cases and controls efficiently to analyze secondary phenotypes in GWASs. This method is comprehensive because it covers both quantitative and dichotomous secondary phenotypes and both rare and non-rare disease. However, the maximum likelihood estimates are only unbiased and statistically efficient under the assumption that there is no interaction between gene and secondary phenotype effects on disease risk. In their paper, they stated that all standard methods based on controls only, cases only, and the combination of cases and controls yield unbiased estimates if the disease is rare. However, this is not true if there is an interaction between gene and secondary phenotype effects on disease risk.

The purpose of this study was to develop a procedure that can use cases as well as controls to increase efficiency without introducing bias. We focused on the important special situation in which the original disease is rare and the secondary phenotype and genetic marker are dichotomous. We investigated the performance of standard methods and the maximum likelihood estimation (MLE) method discussed in Lin and Zeng [2009] while allowing for the interaction. We found that the MLE method is equivalent to the analysis of controls only and does not provide additional efficiency, if one includes an interaction in the model.

We proposed an adaptively weighted method that combines the case and control data to estimate the association with reduced mean square error, based on a balance between bias and variance. In the presence of strong interaction, this method reduces to the controls only analysis. Both simulated and real data examples suggest an advantage of the proposed adaptively weighted estimator in both estimation and gene discovery. This research also provides guidance on the validity of the various proposed approaches for analyzing secondary phenotypes.

The paper is organized as follows. In the Methods Section, we describe the study setting and data and present the different analytic methods, including our adaptively weighted method. In the Results Section, we illustrate the performance of different methods on colorectal adenoma case-control data, for which there is an interaction between gene and secondary phenotype (smoking) on disease risk. In simulations, we evaluate the properties of estimates and tests from the different methods. We also illustrate the promising power of the adaptively weighted method in large-scale SNP discovery GWAS. Conclusions are in the Discussion Section and technical details are in the Appendix .

2 METHODS

2.1 STUDY SETTING AND NOTATION

We consider the simple but important scenario of an unmatched case-control study with a rare disease, dichotomous genetic marker G, and dichotomous secondary phenotype X. Let $D = 1$ or 0 denote the diseased or non-diseased state for each individual. Let $G = 1$ or 0 according as an individual carries at least one SNP allele of interest or not. Let $X = 1$ or 0 denote whether the individual has or does not have the secondary trait. n_0 and n_1 are the number of controls and cases, respectively. The data can be represented as in Table 1.

Let $\mathbf{r_0} = (r_{000}, r_{001}, r_{010}, r_{011})$ and $\mathbf{r_1} = (r_{100}, r_{101}, r_{110}, r_{111})$ denote the case and control cell frequency vectors, respectively. Let $\mathbf{p_0} = (p_{000}, p_{001}, p_{010}, p_{011})$, where $p_{011} = 1-p_{000}-p_{001}-p_{010}$, and $\mathbf{p_1} = (p_{100}, p_{101}, p_{110}, p_{111})$, where $p_{111} = 1-p_{100}-p_{101}-p_{110}$,

denote the unknown true cell probabilities in the underlying case and control populations, respectively. The observed cell frequencies can be viewed as realizations from two independent multinomial distributions, namely, $\mathbf{r_0} \sim$ Multinomial(n_0,$\mathbf{p_0}$) and $\mathbf{r_1} \sim$ Multinomial(n_1, $\mathbf{p_1}$).

One can use the logistic regression model for the dichotomous secondary phenotype

$$\mathrm{P}(X=1|G)=\frac{\exp(\beta_0+\beta_1 G)}{1+\exp(\beta_0+\beta_1 G)}. \quad (1)$$

When the disease is rare, the dependency of D on G and X can be modeled as

$$\mathrm{P}(D=1|G,X) \doteq \exp(\mu+\delta_1 G+\delta_2 X+\delta_{12} GX). \quad (2)$$

We are interested in the inference regarding β_1, which represents the log odds ratio of G and X in the general population, namely $\exp(\beta_1) = \mathrm{OR}_{\mathrm{GX}}$.

2.2 STANDARD METHODS

1. THE CONTROL-ONLY ESTIMATOR—The controls can be regarded as a random sample of the general population if disease is rare, and therefore the odds ratio of G and X among the controls estimates the odds ratio in the population, namely $\mathrm{OR}_{\mathrm{GX}}=\mathrm{OR}_{\mathrm{GX}}^{D=0}$. The MLE of β_1 using controls only is given by

$$\widehat{\beta}_{1\mathrm{CO}}=\log\left(\frac{r_{000}r_{011}}{r_{001}r_{010}}\right).$$

This estimator is nearly unbiased if the disease is rare. From standard asymptotic theory, $\widehat{\sigma}^2_{\mathrm{CO}}=\widehat{\mathrm{Var}}\left(\widehat{\beta}_{1\mathrm{CO}}\right)=\sum_{g=0}^{1}\sum_{x=0}^{1}\left(1/r_{0gx}\right)$. The chi-square test of association is based on the Wald statistic $W_{\mathrm{CO}}=\widehat{\beta}^2_{1\mathrm{CO}}/\widehat{\sigma}^2_{\mathrm{CO}}$.

2. THE CASE-ONLY ESTIMATOR—The estimator of β_1 using cases only is given by

$$\widehat{\beta}_{1\mathrm{CA}}=\log\left(\frac{r_{100}r_{111}}{r_{101}r_{110}}\right).$$

This is not unbiased unless $\delta_{12} = 0$ in model (2), because the odds ratio of G and X among the cases is

$$\mathrm{OR}_{\mathrm{GX}}^{D=1}=\mathrm{OR}_{\mathrm{GX}}\exp(\delta_{12}). \quad (3)$$

The variance estimate of the case estimator is $\widehat{\sigma}^2_{\mathrm{CA}}=\widehat{\mathrm{Var}}\left(\widehat{\beta}_{1\mathrm{CA}}\right)=\sum_{g=0}^{1}\sum_{x=0}^{1}\left(1/r_{1gx}\right)$, which can be used in the Wald statistic $W_{\mathrm{CA}}=\widehat{\beta}^2_{1\mathrm{CA}}/\widehat{\sigma}^2_{\mathrm{CA}}$.

3. WEIGHTED COMBINATION OF INDEPENDENT CASE AND CONTROL ESTIMATORS—When there is no interaction between G and X in model (2), both cases

and controls can be used to estimate β_1. Combining the case and control estimates by inverse variance weighting leads to

$$\widehat{\beta}_{1W}=w_{cc}\widehat{\beta}_{1CO}+(1-w_{cc})\widehat{\beta}_{1CA},$$

where $w_{cc}=\widehat{\sigma}^2_{CA}/\left(\widehat{\sigma}^2_{CA}+\widehat{\sigma}^2_{CO}\right)$. By ignoring the variation in $\widehat{\sigma}^2_{CA}$ and $\widehat{\sigma}^2_{CO}$, we estimate $\widehat{\sigma}^2_{W}=\widehat{\mathrm{Var}}\left(\widehat{\beta}_{1W}\right)=\widehat{\sigma}^2_{CA}\widehat{\sigma}^2_{CO}/\left(\widehat{\sigma}^2_{CA}+\widehat{\sigma}^2_{CO}\right)$, which can be used in the Wald statistic $W_W=\widehat{\beta}^2_{1W}/\widehat{\sigma}^2_{W}$. Although $\widehat{\beta}_{1W}$ is more efficient than $\widehat{\beta}_{1CO}$, it is only unbiased when $\delta_{12}=0$.

2.3 MAXIMUM LIKELIHOOD ESTIMATION (MLE) METHOD

Lin and Zeng [2009] used the maximum likelihood estimation method to analyze secondary phenotype data based on the retrospective likelihood function,

$$\prod_{j=0}^{1}\prod_{i=1}^{n_j}\mathrm{P}\left(G_i,X_i|D=j\right)=\prod_{j=0}^{1}\prod_{i=1}^{n_j}\frac{\mathrm{P}\left(D=j|G_i,X_i\right)\mathrm{P}\left(X_i|G_i\right)\mathrm{P}\left(G_i\right)}{\mathrm{P}\left(D=j\right)}. \quad (4)$$

Using the rare disease assumption with a dichotomous secondary phenotype, we found the following results from this likelihood:

Result 1—Using the saturated disease model (2), the maximum likelihood estimator $\widehat{\beta}_1^{\mathrm{MLE}}=\widehat{\beta}_1^{\mathrm{CO}}$, and the maximum likelihood method does not use any information from the cases. (See Appendix for proof.)

Result 2—Lin and Zeng [2009] assumed $\delta_{12}=0$ in model (2). Under this assumption, and for a rare disease, $\widehat{\beta}_{1MLE}$ has only very slightly smaller asymptotic variance than $\widehat{\beta}_{1W}$. For example, for $\delta_{12}=0$ and $\beta_1=0.25$, when $n_1=n_0=1,000$, $\widehat{Var}\left(\widehat{\beta}_{1MLE}\right)=0.011134$ and $\widehat{Var}\left(\widehat{\beta}_{1W}\right)=0.011152$; when $n_1=n_0=10,000$, $\widehat{\mathrm{Var}}\left(\widehat{\beta}_{1MLE}\right)=0.0010852$ and $\widehat{\mathrm{Var}}\left(\widehat{\beta}_{1W}\right)=0.0010853$.

2.4 ADAPTIVELY WEIGHTED METHOD

To capture some of the efficiency of $\widehat{\beta}_{1W}$ or $\widehat{\beta}_{1MLE}$ while avoiding the bias in these estimates that results when $\delta_{12}\neq 0$, we proposed an estimator that down-weights the contribution from cases as the evidence against $\delta_{12}=0$ increases.

Motivated by an empirical Bayes shrinkage estimator for gene-environment interaction [Mukherjee and Chatterjee, 2008], we proposed the following estimate:

$$\widehat{\beta}_{1AW}=\frac{\widehat{\delta}^2_{12}}{\left(\widehat{\sigma}^2_{CO}+\widehat{\delta}^2_{12}\right)}\widehat{\beta}_{1CO}+\frac{\widehat{\sigma}^2_{CO}}{\left(\widehat{\sigma}^2_{CO}+\widehat{\delta}^2_{12}\right)}\widehat{\beta}_{1W}, \quad (5)$$

where $\widehat{\delta}_{12}=\left(\widehat{\beta}_{1CA}-\widehat{\beta}_{1CO}\right)$ estimates the interaction.

From (5), as $\widehat{\delta}_{12}^2 \to 0$, $\widehat{\beta}_{1AW} \to \widehat{\beta}_{1W}$, and as $\widehat{\delta}_{12}^2 \to \infty$, $\widehat{\beta}_{1AW} \to \widehat{\beta}_{1CO}$. Thus $\widehat{\beta}_{1AW}$ adaptively combines the weighted and control only estimators based on the interaction estimate $\widehat{\delta}_{12}^2$.

We estimated the variance of $\widehat{\beta}_{1AW}$ by rewriting (5) in term of $\widehat{\beta}_{1CO}$ and $\widehat{\beta}_{1CA}$ as

$$\widehat{\beta}_{1AW} = \widehat{\beta}_{1CO} + \frac{\widehat{\sigma}_{CO}^2 (1 - w_{cc}) \left(\widehat{\beta}_{1CA} - \widehat{\beta}_{1CO}\right)}{\widehat{\sigma}_{CO}^2 + \left(\widehat{\beta}_{1CA} - \widehat{\beta}_{1CO}\right)^2}. \tag{6}$$

By noting that $\widehat{\beta}_{1CO}$ and $\widehat{\beta}_{1CA}$ are independent and neglecting the variability of $\widehat{\sigma}_{CO}^2$ and $\widehat{\sigma}_{CA}^2$, we obtained the following variance estimate by Taylor expansion,

$$\widehat{\sigma}_{AW}^2 = \widehat{Var}\left(\widehat{\beta}_{1AW}\right) = \widehat{\sigma}_{CO}^2 \left[1 + \frac{\widehat{\sigma}_{CO}^2 (1 - w_{cc}) \left\{\left(\widehat{\beta}_{1CA} - \widehat{\beta}_{1CO}\right)^2 - \widehat{\sigma}_{CO}^2\right\}}{\left\{\widehat{\sigma}_{CO}^2 + \left(\widehat{\beta}_{1CA} - \widehat{\beta}_{1CO}\right)^2\right\}^2}\right]^2 + \widehat{\sigma}_{CA}^2 \left[\frac{\widehat{\sigma}_{CO}^2 (1 - w_{cc}) \left\{\left(\widehat{\beta}_{1CA} - \widehat{\beta}_{1CO}\right)^2 - \widehat{\sigma}_{CO}^2\right\}}{\left\{\widehat{\sigma}_{CO}^2 + \left(\widehat{\beta}_{1CA} - \widehat{\beta}_{1CO}\right)^2\right\}^2}\right]^2.$$

This estimator is used to construct Wald statistic $W_{AW} = \widehat{\beta}_{1AW}^2 / \widehat{\sigma}_{AW}^2$.

3 RESULTS

3.1 COLORECTAL CANCER, SMOKING AND NAT2

Colorectal adenoma is a precursor of colorectal cancer. Colorectal adenoma is positively associated with smoking and negatively associated with haplotypes of a gene NAT2 that promotes rapid acetylation of carcinogens [Moslehi and others, 2006]. Here we code NAT2 = 1 for haplotypes corresponding to rapid acetylation and NAT2 = 0 for other haplotypes. NAT2 is also important in the metabolism of smoking-related carcinogens. Moslehi and others [2006] analyzed a case-control study of advanced colorectal adenoma in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial at the National Cancer Institute. Table 2 shows the observed cell counts for this case-control study. From marginal tables constructed from Table 2, one sees that the odds ratios relating colorectal adenoma to NAT2 and smoking are 0.93 and 1.41 respectively, demonstrating the protective effect of the rapid acetylator phenotype that corresponds to NAT2=1. The odds ratios relating NAT2 to smoking are 1.85 in controls but 0.378 in cases, indicating a protective interaction between smoking and NAT2 on the risk of colorectal adenoma. Adjusting for age and sex, we found $\widehat{\delta}_{12} = -1.5875$ with p-value= 0.0032. Ignoring this interaction will introduce bias in analyzing the secondary phenotype, smoking. Estimates of the log odds ratio associating NAT2 with smoking are shown for the five methods in Section 2 in Table 3. The adaptively weighted estimate is positive and similar to the control only estimate, while the case only, weighted and MLE estimates are negative. This example is chosen to illustrate the possibility that interactions might exist. In this case, NAT2 rapid acetylation helps eliminate carcinogens from smoking, thereby reducing the risk of colorectal adenoma from smoking. Typically, there might be no or only weak interactions, but this example illustrates the need for methods that are not misleading when interactions are present.

3.2 SIMULATIONS FOR ESTIMATES OF EFFECTS OF A PRE-SELECTED SNP

We used Monte Carlo simulation to evaluate the performance of different estimators of the effects of a pre-selected SNP. We fixed the probabilities of carrying one or two alleles of

interest, G, and the secondary phenotype, X, as $P(G = 1) = 0.3$ and $P(X = 1) = 0.3$. We let $\beta_1 = 0$ and 0.25 under the null and alternative hypotheses, respectively. Fixing these three values, one can determine the distribution of G and X in the population by solving three equations. Under the rare disease assumption, this solution also approximates the control cell probability vector $\mathbf{p_0}$. For the disease-risk model (2), we set $\mu = -10$ to reflect the rare disease and let $\delta_1 = \delta_2 = 0$ and δ_{12} vary from −2 to 2. The case cell probability vector $\mathbf{p_1}$ can be determined by combining $\mathbf{p_0}$ with the disease risk model [Satten and Kupper, 1993]. For each set of simulation parameters, we generated 10, 000 datasets with 1, 000 cases and 1, 000 controls from two independent multinomial distributions corresponding to the case and control populations. We estimated β_1 by the adaptively weighted (AW) method, the controls only (CO) method, the cases only (CA) method, the weighted (W) method and the MLE method.

Figure 1 shows the relative biases (left panels), the coverage probabilities of 95% confidence intervals for β_1 (middle panels), and mean squared errors (MSE) for the AW and CO methods (right panels). Figure 2 displays the type I error for all methods and the power for testing $\mathbf{H_0} : \beta_1 = 0$ against $\mathbf{H_1} : \beta_1 = 0.25$ for the CO and AW methods. All calculations were based on $n_0 = n_1 = 1, 000$ and $\alpha = 0.05$.

The AW and CO methods perform very well. Their estimates are virtually unbiased and their confidence intervals have proper coverage probabilities (Figure 1). The AW method has smaller MSE near $\delta_{12} = 0$ than the CO method (Figure 1), but not for values of δ_{12} away from 0. The CO method maintains nominal size (Figure 2). The AW method maintains near nominal type I error, although there is a very slight increase in size above 0.05 near $|\delta_{12}| = 0.4$. The power of AW exceeds that of CO for $\delta_{12} > 0$, but is less than that of CO for $\delta_{12} < 0$ (Figure 2).

The CA, W and MLE methods, are seriously biased (Figure 1), have substantially subnominal coverage (Figure 1), and have above nominal Type I error (Figure 2) for $\delta_{12} \neq 0$.

3.3 IMPACT OF NON-ZERO δ_{12} ON GENE DISCOVERY

3.3.1 GENOME-WIDE SIZE AND POWER—In the previous section we studied estimation and hypothesis testing for a single pre-specified SNP. Here we investigate the size and power of different methods in large-scale SNP discovery GWAS.

By replacing the $\{\mathbf{r_0}, \mathbf{r_1}\}$ in the different methods with the expected cell counts $\{n_0\mathbf{p_0}, n_1\mathbf{p_1}\}$, we obtained the asymptotic mean and variance for each estimate and hence the non-centrality, λ, of the corresponding one degree freedom chi-squared distribution for the Wald statistic, denoted as $\chi^2_{1,\lambda}$. Under the null hypothesis, when $\delta_{12} = 0$, each method has $\lambda = 0$; when $\delta_{12} \neq 0$, every method has nonzero λ except the control only method. Under the alternative, all the λs are non-zero. We can compute the asymptotic size and power for testing one given SNP from the formula: $\mathrm{P}_{\chi^2_{1,\lambda}}\left[W > q_{\chi^2_{1,0},1-\alpha}\right]$ with the corresponding λ under the null and alternative hypothesis for each method respectively and $q_{\chi^2_{1,0},1-\alpha}$ being the quantile of the central chi-square distribution, .

Assuming there were $N = 500, 000$ independent SNP genetic markers, we controlled the experimentwise significance by setting $\alpha = 0.05/(5 \times 10^5) = 10^{-7}$. It may be reasonable to suppose that $\delta_{12} = 0$ for a large proportion of SNPs. We assume 99% of SNPs have $\delta_{12} = 0$, and 1% of δ_{12} are independently distributed as $N(0, (\log(2)/2)^2)$, which implies that about 95% of nonzero δ_{12} values reside in $[-\log(2), \log(2)]$. We evaluated the genome-wide type I

error and power of the various Wald tests averaged over the mixture distribution of δ_{12}. For each method we estimated the genome-wide type I error under the null from the formula

$$\int_{-\infty}^{\infty}\left\{1-\prod_{j=1}^{N}\mathrm{P}_{\chi^2_{1,\lambda}}\left(W_j<q_{\chi^2_{1,0},1-\alpha}|\delta_{12}^{(j)}\right)\right\}\mathrm{d}F\left(\delta_{12}^{(1)},\delta_{12}^{(2)},\cdots,\delta_{12}^{(N)}\right)$$
$$=1-\left\{0.99\mathrm{P}_{\chi^2_{1,\lambda}}\left(W<q_{\chi^2_{1,0},1-\alpha}|\delta_{12}=0\right)+0.01\int_{-\infty}^{\infty}P_{\chi^2_{1,\lambda}}\left(W<q_{\chi^2_{1,0},1-\alpha}|\delta_{12}\right)f(\delta_{12})\,\mathrm{d}\delta_{12}\right\}^N, \quad (7)$$

and the power under the alternative $\beta_1 = 0.25$ can be calculated from

$$0.99\mathrm{P}_{\chi^2_{1,\lambda}}\left(W\geq q_{\chi^2_{1,0},1-\alpha}|\delta_{12}=0\right)+0.01\int_{-\infty}^{\infty}P_{\chi^2_{1,\lambda}}\left(W\geq q_{\chi^2_{1,0},1-\alpha}|\delta_{12}\right)f(\delta_{12})\,\mathrm{d}\delta_{12}, \quad (8)$$

where $f(x)$ is the normal density N[0, {log(2)/2}^{2}]. The integrals in equations (7) and (8) are approximated by drawing 100,000 δ_{12} from N[0, {log(2)/2}^{2}] and averaging the corresponding values of $\mathrm{P}_{\chi^2_{1,\lambda}}\left(W<q_{\chi^2_{1,0},1-\alpha}|\delta_{12}\right)$ for equation (7) and of $\mathrm{P}_{\chi^2_{1,\lambda}}\left(W\geq q_{\chi^2_{1,0},1-\alpha}|\delta_{12}\right)$ for equation (8). The non-centrality λ depends on δ_{12} and are recomputed for each realization.

The genome-wide type I error and power for different methods are presented in Table 4 for numbers of cases and controls $n = 1000$, $n = 5000$ and $n = 10{,}000$. The genome-wide type I error is 1.0 for W_{CA}, W_W, and W_{MLE}, indicating that these tests should not be used even if only a small proportion of SNPs have $\delta_{12} \neq 0$. Both W_{CO} and W_{AW} statistics have near nominal genome-wide type I error. However, for sample size $n_1 = n_0 = 5,000$ or 10, 000, the power of W_{AW} greatly exceeds that for W_{CO}. For example, when $n_0 = n_1 = 10{,}000$, the power of W_{AW} is almost 98%, while that of W_{CO} is only 50%. Thus, substantial power gains can be achieved with W_{AW}. Unreported simulations from a mixture distribution with 20% non-zero δ_{12}s also show that W_{AW} has greater power than CO but a very slight excess in size (e.g. size= 0.053) was observed for W_{AW}.

3.3.2 GENOME-WIDE DETECTION PROBABILITY—Instead of setting fixed critical values for declaring an association statistically significant, one can rank p-values or chi-square statistics to select promising SNPs [Gail et. al., 2008]. In this section, we study the detection probability which is the probability that the test statistic for a specified disease-associated SNP will be among the top T chi-square values for all SNPs. We estimated the detection probability with 1, 000 simulated replicates. In each replicate, we generated data for $N = 500,000$ SNPs with the same parameters used in Section 3.2, except that 10 disease-associated SNPs had nonzero β_1, whereas the 499,990 remaining null SNPs had $\beta_i = 0$. All SNPs had interactions δ_{12} drawn from a mixture distribution in which 99% of SNPs have $\delta_{12} = 0$, and 1% of δ_{12} are independently distributed as $N(0,\sigma^2)$. We conducted four independent simulations for the combinations of $\beta_1 = \{0.25, 0.69\}$ and $\sigma^2 = \{(\log(2)/2)^2, (\log(2))^2\}$.

When the associated SNPs have small log odds ratios with the secondary phenotype, i.e. $\beta_1 = 0.25$ (upper portion of Table 5), all the methods have low detection probabilities for $T \leq 100$.. When $T \leq 100$ and the variability of δ_{12} is large, the W and MLE methods have detection probabilities near zero, because some of the SNPs with interactions have large chi-square values than the disease-associated SNPs. The CO and AW methods perform slightly better. When T is 10, 000, (2% of the total SNPs), the detection probabilities of all methods increase and the W and MLE methods outperform the CO and AW methods. When the

associated SNPs have a stronger association with the secondary phenotype, i.e. $\beta_1 = 0.69$ (lower portion of Table 5), and with $\sigma^2 = \{\log(2)/2\}^2$, W and MLE have smaller detection probabilities than CO and AW for $T = 10$, but for $T = 100$ or 10, 000, the detection probabilities of W and MLE exceed those of CO and AW. When the variability of δ_{12} is larger, with $\sigma^2 = \{\log(2)\}^2$, the W and MLE methods have smaller detection probabilities than CO and AW, especially for $T = 10$ and 100. The CA method always has the lowest detection probability. As expected, CO is not affected by increasing variability of δ_{12}; AW is also robust to increasing variability of δ_{12}.

4 DISCUSSION

For a rare disease and dichotomous secondary endpoint and genetic marker, we have investigated whether and how to use data from diseased subjects to study the association between a genetic marker and secondary phenotype. We considered both estimating and testing the null hypothesis of no association for a pre-specified SNP, and for discovering an association in a GWAS with either hypothesis testing approaches or approaches based on ranking the chi-square statistics. In the absence of an interaction δ_{12} in model (2), each of the five methods we considered leads to valid inference, and the W and MLE methods are particularly efficient. In the presence of interaction ($\delta_{12} \neq 0$), the CO method controls the type I error perfectly, and the AW has proper size for rare interactions (1%) and only modestly supra-nominal type I error for common interactions (20%). The CA, W and MLE methods do not control type I error and cannot be recommended if it is plausible that $\delta_{12} \neq 0$. The AW method has lower MSE for a pre-selected SNP and greater power than the CO method, which is achieved at the cost of a slight increase in type I error. We showed that the MLE method reduces to the CO method if the model allows for non-null δ_{12}.

Under the assumption of rare disease and dichotomous genetic marker and phenotype, the CO method is robust in that it maintains the unbiasedness and nominal type I error despite any interaction effect. The W and MLE methods fully utilize both controls and cases and are most efficient for estimation. When there is no interaction effect, both weighted and MLE are almost twice as efficient as the control only method in estimation and have around 70% more power than the control only method. We prefer the weighted method because it is nearly fully efficient and its estimator is non-iterative. Thus, there are no problems of convergence as can arise with the MLE method. However, even a small interaction effect causes large bias and highly inflated type I error for the CA, W, and MLE methods. The AW method strikes a balance between the robust CO method and the W method. It maintains the unbiasedness and near nominal type I error across most values of δ_{12}, although it has moderately inflated type I error when δ_{12} is not far from zero. If δ_{12} is near zero, estimates based on AW have smaller MSE than those from CO for a prespecified SNP. Under a mixture distribution for δ_{12} which was chosen to allow most δ_{12} values to be zero, the AW method achieved an important gain in power compared to CO. The detection probabilities of the W and MLE methods degrade when ranking SNPs in the presence of increasing variability of δ_{12}. However, CO and AW methods maintain their detection probabilities and are robust to increasing variable δ_{12}.

Jiang, Scott and Wild [2006] discussed methods for analyzing secondary phenotypes in case-control studies. Their fully non-parametric approach (SPML1) corresponds to MLE under our model (2) with δ_{12} included, which is equivalent to the CO method for inference on β_1. Assuming $\delta_{12} = 0$ corresponds to possibly misspecified parametric modeling (SPML2) in their notation. MLE under SPML2 was described by Lin and Zeng [2009], who also covered non rare diseases and both dichotomous and continuous secondary phenotypes. If $\delta_{12} \neq 0$, the MLE method of Lin and Zeng does not control the type I error, as indicated

by our results in Section 3.2 and 3.3 and in the discussion of model misspecification for SPML2 by Jiang et al. [2006].

In unreported analysis, we evaluated the performance of the various methods using prostate cancer data from the Cancer Genetic Markers of Susceptibility (CGEMS) study. We conducted a genome-wide scan on the association between the secondary phenotype, body mass index BMI (1, if BMI≤ 25; 0, else), and the 516,564 SNPs from 22 autosomal chromosomes. We estimated the distribution of δ_{12}, and found no evidence that the variance of $\widehat{\delta}_{12}$ across SNPs exceeded that which would be expected from the multinomial sampling error alone. Thus, we did not find evidence that $\delta_{12} \neq 0$ for some SNPs. Under such situation, W and MLE are two most efficient methods, and both identified SNP rs7575639 with a genome-wide significant $p < 10^{-7}$. The 20 SNPs with smallest p-values selected by the W and MLE methods were identical, with only slight differences in ranking. For 11 SNPs, spurious results resulted from convergence problems for MLE. Only careful scrutiny of the extreme values for these SNPs revealed the problem with MLE. For this reason, we recommend the numerically stable W method instead of MLE.

Kraft [2007] argued that it is unlikely for both the secondary phenotype and genetic marker to affect the original case-control disease risk, let alone for there to be an interaction. In terms of equation (2), he suggested that either δ_1 or δ_2 would usually be zero and implicitly that δ_{12} would be zero. If this is so, one could use the W method and gain precision and power thereby. More experience is needed with GWASs to see if the W or MLE methods yield many false positive results as a consequence of $\delta_{12} \neq 0$, or if their detection probabilities for ranking promising SNPs are degraded by the presence of interaction effects. Our work makes it clear that spurious positive findings may result from such an interaction, and that one can protect against such findings by using the control only or adaptively weighted approaches.

Acknowledgments

The authors thank Dr. Kai Yu for helpful suggestions. This work was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute. This research was partially supported by the NIH Gene Environment Initiative (GEI) program. This research utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, Maryland, USA (http://biowulf.nih.gov).

6 APPENDIX. DERIVATION OF OF β^1MLE USING DISEASE MODEL (2) WITH AND WITHOUT INTERACTION TERM

Start from the retrospective likelihood function,

$$L=\prod_{j=0}^{1}\prod_{i=1}^{n_j}\frac{\mathrm{P}(D=j|G_i,X_i)\,\mathrm{P}(G_i|X_i)\,\mathrm{P}(X_i)}{\mathrm{P}(D=j)}.$$

1. Using the saturated disease model $\mathrm{P}(D=1|G,X) \doteq \exp(\mu+\delta_1 G+\delta_2 X+\delta_{12}GX)$ and notations in Table 1, we obtained the log-likelihood function,

$$
\begin{aligned}
l = {} & (r_{100}+r_{110}+r_{000}+r_{010}) \log (P_0) \\
& - \log \left(1+e^{\beta_0}\right) \\
& + (r_{101}+r_{111}+r_{001}+r_{011}) \left(\log (1-P_0) - \log \left(1+e^{\beta_0+\beta_1}\right)\right) \\
& - (r_{100}+r_{110}+r_{101}+r_{111}) \log \left(\frac{1+e^{\delta_1+\beta_0}}{1+e^{\beta_0}} P_0 + \frac{e^{\delta_2}\left(1+e^{\delta_1+\delta_{12}+\beta_0+\beta_1}\right)}{1+e^{\beta_0+\beta_1}} (1-P_0) \right) \\
& + r_{110} (\delta_1+\beta_0) \\
& + r_{101} \delta_2 \\
& + r_{111} (\delta_1+\delta_2+\delta_{12}+\beta_1+\beta_0) \\
& + r_{010} \beta_0 \\
& + r_{011} (\beta_0+\beta_1), \qquad (9)
\end{aligned}
$$

where $P_0 = P(X = 0)$. In this log-likelihood function, there are six unknown parameters $\{\beta_0, \beta_1, P_0, \delta_1, \delta_2, \delta_{12}\}$. By differentiating (9) with respect to each parameter and setting the derivatives to zero, we obtain six equations. Solving them, we obtained the following analytic solutions: $\widehat{\beta}_0^{\mathrm{MLE}} = \log\left(\frac{r_{010}}{r_{000}}\right)$, $\widehat{\beta}_1^{\mathrm{MLE}} = \log\left(\frac{r_{011} r_{000}}{r_{001} r_{010}}\right)$, $\widehat{P}_0^{\mathrm{MLE}} = \frac{r_{000}+r_{010}}{r_{000}+r_{010}+r_{001}+r_{011}}$, $\widehat{\delta}_1^{\mathrm{MLE}} = \log\left(\frac{r_{110} r_{000}}{r_{100} r_{010}}\right)$, $\widehat{\delta}_2^{\mathrm{MLE}} = \log\left(\frac{r_{101} r_{000}}{r_{100} r_{001}}\right)$, and $\widehat{\delta}_3^{\mathrm{MLE}} = \log\left(\frac{r_{111} r_{100} r_{001} r_{010}}{r_{101} r_{110} r_{001} r_{000}}\right)$. Thus $\widehat{\beta}_1^{\mathrm{MLE}} = \widehat{\beta}_1^{\mathrm{CO}}$, proving the assertion in Section 2.3.

2. Using the reduced disease model $P(D=1|G, X) \doteq \exp(\mu+\delta_1 G+\delta_2 X)$ with $\delta_{12} = 0$, we have the following log-likelihood function:

$$
\begin{aligned}
l = {} & (r_{100}+r_{110}+r_{000}+r_{010}) \log (P_0) \\
& - \log \left(1+e^{\beta_0}\right) \\
& + (r_{101}+r_{111}+r_{001}+r_{011}) (\log (1-P_0) \\
& - \log \left(1+e^{\beta_0+\beta_1}\right) \\
& - (r_{100}+r_{110}+r_{101}+r_{111}) \log \left(\frac{1+e^{\delta_1+\beta_0}}{1+e^{\beta_0}} P_0 + \frac{e^{\delta_2}\left(1+e^{\delta_1+\beta_0+\beta_1}\right)}{1+e^{\beta_0+\beta_1}} (1-P_0) \right) \\
& + r_{110} (\delta_1+\beta_0) \\
& + r_{101} \delta_2 \\
& + r_{111} (\delta_1+\delta_2+\beta_1+\beta_0) \\
& + r_{010} \beta_0 \\
& + r_{011} (\beta_0+\beta_1). \qquad (10)
\end{aligned}
$$

In this loglikelihood function, there are five unknown parameters $\{\beta_0, \beta_1, P_0, \delta_1, \delta_2\}$. By differentiating (10) with respect to each parameter and setting the derivatives to zero, we obtain five equations. There are no explicit solutions for these parameters, except for P_0. Before solving them numerically, we simplified them to:

$$
e^{\beta_1} = \frac{1}{e^{\delta_1+\beta_0}} \frac{e^{\delta_1+\beta_0}(r_{100}-r_{111}) - (r_{110}+r_{111})}{e^{\delta_1+\beta_0}(-r_{101}-r_{100}) + (r_{110}-r_{101})} \qquad (11)
$$

$$e^{\delta_1} = \frac{(r_{110} - r_{000})\, e^{\beta_0} + r_{110} + r_{010}}{((r_{100} + r_{000})\, e^{\beta_0} + r_{100} - r_{010})\, e^{\beta_0}} \tag{12}$$

$$e^{\delta_1} = \frac{(r_{001} - r_{111})\, e^{\beta_0+\beta_1} - (r_{111} + r_{011})}{(r_{011} + r_{101})\, e^{\beta_0+\beta_1} - (r_{101} + r_{001})\, e^{2(\beta_0+\beta_1)}} \tag{13}$$

$$e^{\delta_2} = \frac{1+e^{\beta_0+\beta_1}}{1+e^{\beta_0}} \frac{1+e^{\beta_0+\delta_1}}{1+e^{\beta_0+\beta_1+\delta_1}} \frac{P_0}{1-P_0} \frac{r_{101}+r_{111}}{r_{100}+r_{110}} \tag{14}$$

$$\widehat{P}_0^{\mathrm{MLE}} = \frac{r_{000}+r_{010}}{r_{000}+r_{010}+r_{001}+r_{011}}. \tag{15}$$

Equations (11) - (13) depend on $\{\beta_0, \beta_1, \delta_1\}$ only. Using SAS PROC MODEL, we obtained $\left\{\widehat{\beta}_0^{\mathrm{MLE}}, \widehat{\beta}_1^{\mathrm{MLE}} \widehat{\beta}_1^{\mathrm{MLE}}\right\}$. Substituting these values and $\widehat{P}_0^{\mathrm{MLE}}$ from (15) into (14), we solved for $\widehat{\delta}_2^{\mathrm{MLE}}$. Using these solutions we can estimate the variance of $\widehat{\beta}_{1\mathrm{MLE}}$ from the observed matrix of second derivatives of the log likelihood and compare them with corresponding estimates from the log likelihood (9).

References

[1]. Jiang Y, Scott AJ, Wild CJ. Secondary analysis of case-control data. Stat Med 2006;25:1323–1339. [PubMed: 16220494]

[2]. Gail MH, Pfeiffer RM, Wheeler W, Pee D. Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. Biostatistics (Oxford, England) 2008;9(2):201–15.

[3]. Kraft P. Analyses of genome-wide association scans for additional outcomes. Epidemiology 2007;18:838. [PubMed: 18049198]

[4]. Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genet Epidemiol 2009;33:256–265. [PubMed: 19051285]

[5]. Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, Weissfield J, Hein DW, Hayes RB. Cigarette smoking n-acetyltransferase genes and the risk of advanced colorectal adenoma. Pharmacogenomics 2006;7:819–829. [PubMed: 16981843]

[6]. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of casecontrol studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 2008;64(3):685–94. [PubMed: 18162111]

[7]. Mukherjee B, Ahn J, Gruber S, Rennert G, Moreno V, Chatterjee N. Tests for gene-enviroment interaction from case-control data: a novel study of type I error, power and designs. Genet Epidemiol 2008;32:615–626. [PubMed: 18473390]

[8]. Satten GA, Kupper LL. Inferences about exposure-disease associations using probability-of-exposure information. J Am Sta Assoc 1993;88:200–208.

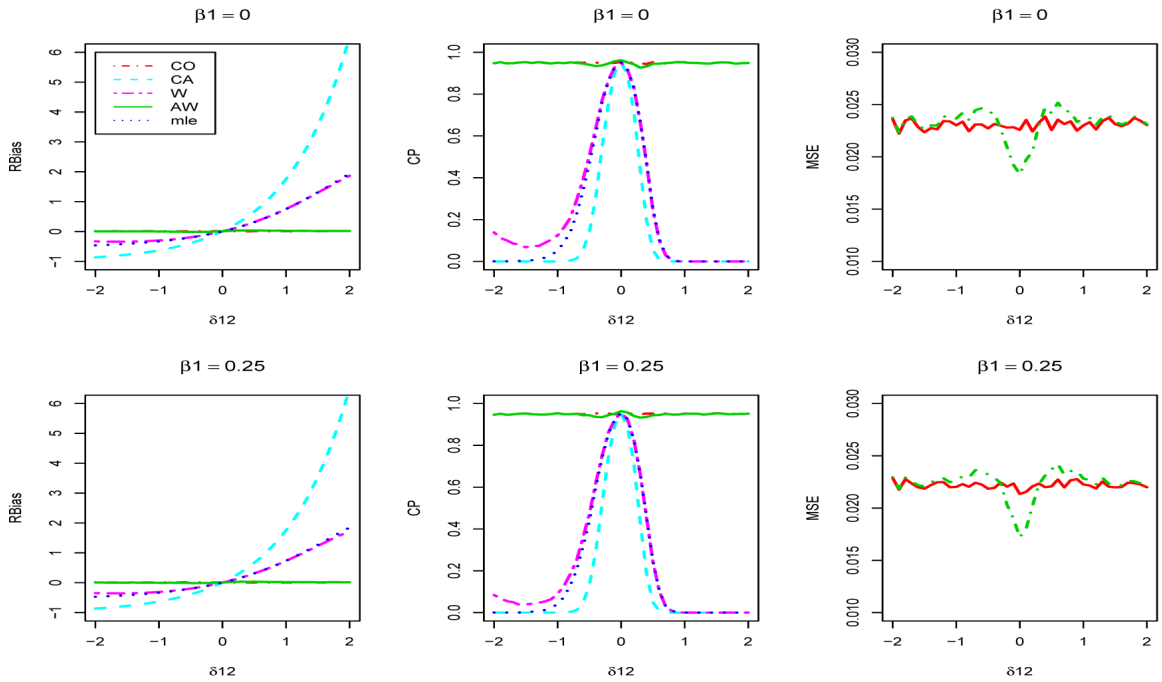

Figure 1.
Relative biases (RBias), coverage probabilities (CP) of 95% confidence intervals and mean squared errors (MSE) for different estimators for two values of β_1:(1) $\beta_1 = 0$; (2) $\beta_2 = 0.25$.

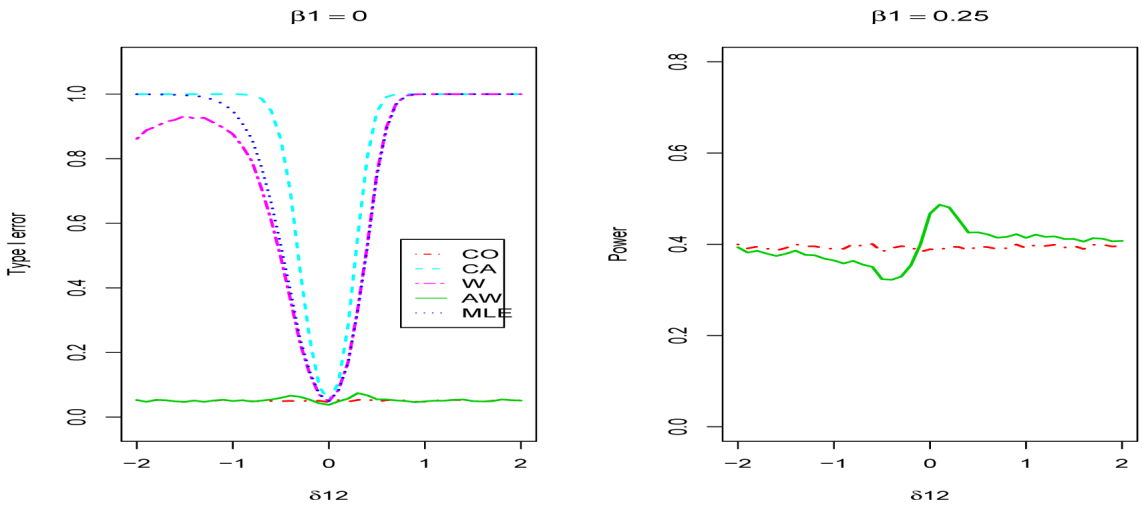

Figure 2.
Type I error rates and power of association tests at the 5% nominal significance level.

Table 1

Data for an unmatched case-control study with dichotomous genetic marker and secondary phenotype

	$G = 0$		$G = 1$		
	$X = 0$	$X = 1$	$X = 0$	$X = 1$	Total
$D = 0$	r_{000}	r_{001}	r_{010}	r_{011}	n_0
$D = 1$	r_{100}	r_{101}	r_{110}	r_{111}	n_1

Table 2

Data for studying NAT2 and smoking association in the colorectal adenoma case-control study*

	NAT2=0		NAT2=1		
	Smoking=0	**Smoking=1**	**Smoking=0**	**Smoking=1**	**Total**
D = 0	255	317	10	23	605
D = 1	199	380	18	13	610

*D = 1 denotes colorectal adenoma, and D = 0 denotes control; Smoking=1 denotes current smoker, and smoking=0 denotes former or never smoker; NAT2 = 1 denotes rapid acetylator haplotypes, and NAT2 = 0 denotes slow acetylator haplotypes.

Table 3

Odds ratio estimates associating NAT2 with smoking in the colorectal adenoma case-control study

	log(Odds Ratio)	s.d.
Control only	0.615	0.39
Case only	−0.972	0.27
Weighted	−0.207	0.27
Adaptively weighted	0.569	0.40
MLE	−0.172	0.26

Table 4

Genome-wide type I error and power for δ_{12} from a mixture distribution for various sample sizes with genome-wide significance 0.05

	$\beta_1 = 0$; Size			$\beta_1 = .25$; Power		
n0=n1	**1,000**	**5,000**	**10,000**	**1,000**	**5,000**	**10,000**
Control only, W_{CO}	0.049	0.049	0.049	0.000	0.060	0.504
Adaptively weighted, W_{AW}	0.049	0.049	0.049	0.002	0.500	0.982
Case only, W_{CA}	1.000	1.000	1.000	0.001	0.064	0.505
Weighted, W_W	1.000	1.000	1.000	0.002	0.504	0.984
MLE, W_{MLE}	1.000	1.000	1.000	0.002	0.504	0.984

Table 5

Associated SNP detection probability for various numbers of selected SNPs, *T*. $\delta_{12} = 0$ with probability 0.99 and with probability 0.01, δ_{12} is drawn from a normal distribution with mean 0 and variance σ^2

		$\sigma^2 = \{\log(2)/2\}^2 = 0.12$			$\sigma^2 = \{\log(2)\}2 = 0.48$		
	T	**10**	**100**	**10,000**	**10**	**100**	**10,000**
$\beta_1 = 0.25$	Control only(CO)	0.005	0.024	0.275	0.005	0.025	0.274
	Adaptively weighted(AW)	0.012	0.044	0.375	0.012	0.045	0.374
	Case only(CA)	0.000	0.001	0.249	0.000	0.001	0.231
	Weighted(W)	0.001	0.022	0.516	0.000	0.001	0.496
	MLE	0.001	0.019	0.512	0.000	0.001	0.491
$\beta_1 = 0.69$	Control only(CO)	0.62	0.85	0.99	0.62	0.85	0.99
	Adaptively weighted(AW)	0.73	0.90	0.99	0.73	0.90	0.99
	Case only(CA)	0.00	0.16	0.99	0.00	0.00	0.99
	Weighted(W)	0.52	0.99	1.00	0.00	0.09	1.00
	MLE	0.49	0.98	1.00	0.00	0.05	1.00