Published in final edited form as:

Stat Med. 2016 October 15; 35(23): 4226–4237. doi:10.1002/sim.6976.

Robust analysis of secondary phenotypes in case-control genetic association studies

Chuanhua Xing^{a,*}, Janice M. McCarthy^b, Josée Dupuis^{a,c}, L. Adrienne Cupples^{a,c}, James B. Meigs^{d,e}, Xihong Lin^f, and Andrew S. Allen^{b,g,*}

^aDepartment of Biostatistics, Boston University School of Public Health, Boston MA 02118, U.S.A

^bDepartment of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, U.S.A

^cNational Heart, Lung and Blood Institute's Framingham Heart Study, Framingham, MA 01702, U.S.A

^dGeneral Medicine Division, Massachusetts General Hospital, Boston, MA 02114, U.S.A

^eDepartment of Medicine, Harvard Medical School, Boston, MA 02115, U.S.A

^fDepartment of Biostatistics, Harvard University, Cambridge, MA 01238, U.S.A

^gCenter for Human Genome Variation, Duke University, Durham, NC 27710, U.S.A

Abstract

The case-control study is a common design for assessing the association between genetic exposures and a disease phenotype. Though association with a given (case-control) phenotype is always of primary interest, there is often considerable interest in assessing relationships between genetic exposures and other (secondary) phenotypes. However, the case-control sample represents a biased sample from the general population. As a result, if this sampling framework is not correctly taken into account, analyses estimating the effect of exposures on secondary phenotypes can be biased leading to incorrect inference. In this paper we address this problem and propose a general approach for estimating and testing the population effect of a genetic variant on a secondary phenotype. Our approach is based on inverse probability weighted (IPW) estimating equations, where the weights depend on genotype and the secondary phenotype. We show that, though slightly less efficient than a full likelihood-based analysis when the likelihood is correctly specified, it is substantially more robust to model misspecification, and can out-perform likelihood-based analysis, both in terms of validity and power, when the model is misspecified. We illustrate our approach with an application to a case-control study extracted from the Framingham Heart Study.

Keywords

case-control; estimating equation; inverse-probability-weighted; secondary phenotype; genetic association study

*Correspondence to: andrew.s.allen@duke.edu.

1. Introduction

A case-control study is an economical approach for assessing association between genetic exposures and disease. It is comprised of two separate samples, a sample of individuals with disease (cases) and a sample of individuals that are disease-free (controls). Because cases and controls are sampled separately, cases can be oversampled so that the case sample is made up of a similar number of individuals as the control sample. This balance minimizes the number of exposures that need to be assessed for a given level of statistical power. When a disease is rare or when exposures are expensive to collect, the case-control study can represent a considerable costs savings relative to studying a cohort prospectively. For these reasons, genome-wide and whole exome association studies often utilize a case-control design.

Though genotyping and sequencing costs continue to fall, measuring an individual's genetic exposures, either genome-wide or exome-wide, still represents a significant expense, and the total investment in a genome-wide association study is immense. Thus, there is considerable interest in making the most out of these datasets. Most of these studies will measure other phenotypes, either because they are readily available, or because they are thought to be related to the underlying disease process. Studying the genetic influences of these 'secondary' phenotypes may be of interest in itself or may help our understanding of the biological pathways involved in the disease process. Examples involving secondary phenotypes are numerous and include analyses of body mass index [1, 2], lipid levels [3, 4, 5], and human height [6, 7, 8, 9].

The case-control sample does not constitute a random sample from the general population. As a result, any population association between genetic variants and secondary phenotypes can be distorted in the case-control sample. Thus analysis methods that ignore the case-control design can give biased estimates of the population effect that a genetic variant has on a secondary phenotype.

To address this problem, Richardson, et al. [10] proposed an approach based on a weighted logistic regression of a binary secondary phenotype on binary exposure. Monsees et al. [11] extended this basic approach to more general secondary phenotypes and ordinal or categorical exposures (genotypes). The weights used in both [10] and [11] were the inverse of the known sampling fractions of the cases and controls from the general population. Thus, Monsees, et al. [11] refer to their approach as inverse-probability-of-sampling-weighted (*IPSW*) regression. As the sampling probabilities are assumed known, both manuscripts advocate using this approach in situations where the sampling is explicitly controlled by design, such as when a case-control study is nested within a prospective cohort [11]. However, VanderWeele and Vansteelandt [12] showed that if cases and controls are randomly sampled from a population with known disease prevalence, λ , then the *IPSW* weights can be expressed as a function of λ and the ratio of cases to controls in the sample. Lin and Zeng [13] proposed a statistical method based on a retrospective likelihood of the secondary trait and genotype conditional on case-control status. Their method is a likelihood-based model, thus it is fully efficient when the likelihood is correctly specified. However, it may perform poorly under model misspecification. While robustness under

misspecification has been addressed by Wei, et al. [14], their approach requires a bootstrap procedure to estimate the variance of the estimator for the parameter of interest. Because of the computational expense of such a procedure, that method is not appropriate for use in genome-wide studies.

Here we propose a general framework for robust inference on the population effect of a genetic variant on a secondary phenotype, based on inverse-probability weighted (IPW) estimating equations. Though, our approach is also based on weighting, it differs fundamentally from the sampling-probability approach proposed by Richardson et al. [10] and Monsees et al. [11]. Both [10] and [11] weight by the inverse of a subject's probability of being included in the case-control sample and this probability is assumed to be either known by design, or computed based on a known prevalence [12]. In both cases, the weight is assumed to be a fixed constant, and in the case where the weights are computed based on a known prevalence, the constant only differs based on case-control status. Our approach inversely weights by a subject's probability of being a case (or a control), conditional on their genetic and phenotypic information. This probability is a population quantity, but we show that it can be estimated from a case-control sample either when the disease is rare or when the population prevalence is known. These two cases are complementary, as the population prevalence is often unknown when the disease is rare, but can often be estimated with reasonable precision when the disease is common. We show that this approach gives consistent estimates of the parameters of interest. Our approach performs better than *IPSW* in simulations, having both smaller mean-squared error and higher power. Though slightly less efficient than a full likelihood-based analysis when the likelihood is correctly specified, it is also substantially more robust to model misspecification. Thus, it may outperform a likelihood-based analysis, both in terms of validity and power, when the model is misspecified.

2. Methods

Consider a case-control study in which n_1 cases and n_0 controls have been randomly sampled. Let D be an indicator of case-control status such that $D = 1$ indicates a case and $D = 0$ indicates a control. Let Y denote the secondary phenotype of interest and let G denote genotype. For simplicity, we assume G is biallelic and is coded by the number of minor alleles in the genotype, i.e., $G = 0, 1$, or 2 . We encode a hypothesized genetic model through the function $\eta(G)$. For example, a recessive model is given by $\eta(G) = I(G = 2)$, where I is an indicator function. A dominant model can be specified by $\eta(G) = I(G > 0)$ and an additive model by $\eta(G) = G$. It is also possible for $\eta(G)$ to be a vector, as in the codominant model $\eta(G) = \{I(G > 0), I(G = 2)\}^T$. Let $X = \{1, \eta(G)^T\}^T$ be a design vector. Note that although X could include additional covariates, we omit this case from the presentation for simplicity.

We wish to make inference on the population parameter which indexes the conditional expectation of Y given X in the general population. Specifically, we assume that $E(Y|X)$ in the general population can be specified up to an unknown, finite parameter β through the function $\mu(X, \beta)$, i.e., we assume that $E(Y|X) = \mu(X, \beta)$. For example, if Y is continuous, we may model the conditional expectation using a linear model by assuming $\mu(X, \beta) = \beta^T X$, where β is comprised of an intercept and a parameter vector τ indexing $\eta(G)$. (For technical

reasons, we require that $\mu(X, \beta)$ is twice-differentiable with continuous second derivative with respect to the parameter β .) Our goal is to make inference on β while minimizing assumptions on the distributions of Y and/or X .

Let $n = n_0 + n_1$. If $\{(X_i, Y_i); i = 1, \dots, n\}$ represented a random sample from the population of interest, then β could be estimated as the root to the estimating equation given by

$$U_\beta = \sum_{i=1}^n U_\beta(Y_i, X_i) = \sum_{i=1}^n D^T(X_i, \beta) V^{-1}(X_i) [Y_i - \mu(X_i, \beta)], \quad (1)$$

where $D(X, \beta) = \frac{\partial \mu(X, \beta)}{\partial \beta^T}$ and $V(X)$ is a “working” model for $\text{Var}(Y|X)$. Note that under random sampling, $n^{-1}U_\beta$ is a consistent estimator of $E_*[U_\beta(Y, X)]$ where the ‘*’ is used to emphasize that this expectation is taken with respect to the *true* distribution that generated X and Y in the *population*. This and the fact that $E_*[U_\beta(Y, X)] = 0$, when evaluated at the true population β , implies that the root to equation (1), is a consistent estimator of the true population β . Incorrect specification of $V(X)$ can affect the efficiency, but not the consistency, of the estimator [15].

The subjects in a case-control study do not comprise a random sample, but rather represent two potentially biased samples. Let $\{(X_i, Y_i); i = 1, \dots, n_0\}$ denote controls and $\{(X_i, Y_i); i = n_0 + 1, \dots, n_0 + n_1 = n\}$ denote cases. Let U_β^d be the estimating equations obtained by restricting (1) to the sample with $D = d$, i.e.,

$$U_\beta^0 = \sum_{i=1}^{n_0} U_\beta(Y_i, X_i) \quad (2)$$

and

$$U_\beta^1 = \sum_{i=n_0+1}^n U_\beta(Y_i, X_i). \quad (3)$$

When D is related to X and/or Y , the distribution of (X, Y) obtained under case-control sampling can be distorted from the distribution of (X, Y) in the population. In this case, $n^{-1}U_\beta$ will no longer be a consistent estimator of $E_*[U_\beta(Y, X)]$ and (1) will lead to biased estimates of β . Similarly, neither $n_0^{-1}U_\beta^0$ or $n_1^{-1}U_\beta^1$ is guaranteed to converge to $E_*[U_\beta(Y, X)]$ but instead to $E[U_\beta(Y, X)|D=0]$ and $E[U_\beta(Y, X)|D=1]$, respectively. Thus inference based on (2) or (3) can also lead to bias.

2.1. Inverse-probability weighted estimating equations

Our approach is to weight the components of (2) and (3) so that each sample (case and control) can be used to consistently estimate the population parameter β . In the Appendix section of Supplementary Materials, we show that

$$E \left[\frac{U_{\beta}(Y, X)}{1 - p_{XY}} | D=0 \right] = 0 \iff E_*[U_{\beta}(Y, X)] = 0 \quad (4)$$

and

$$E \left[\frac{U_{\beta}(Y, X)}{p_{XY}} | D=1 \right] = 0 \iff E_*[U_{\beta}(Y, X)] = 0, \quad (5)$$

where p_{XY} denotes the probability of being a case given X and Y , in the population. Thus, our approach is to base inference on the following estimating equations,

$$\tilde{U}_{\beta}^0 = \sum_{i=1}^{n_0} \frac{U_{\beta}(Y_i, X_i)}{1 - p_{X_i Y_i}} \quad (6)$$

and

$$\tilde{U}_{\beta}^1 = \sum_{i=n_0+1}^n \frac{U_{\beta}(Y_i, X_i)}{p_{X_i Y_i}}. \quad (7)$$

Note that $n_0^{-1} \tilde{U}_{\beta}^0$ and $n_1^{-1} \tilde{U}_{\beta}^1$ are consistent estimators for $E \left[\frac{1}{1 - p_{XY}} U_{\beta}(Y, X) | D=0 \right]$ and

$E \left[\frac{1}{p_{XY}} U_{\beta}(Y, X) | D=1 \right]$, respectively. Therefore, relations (4) and (5) imply that the roots to equations (6) and (7) are each consistent estimators of the true population β .

2.2. Estimating p_{XY} with case-control data

The probability, p_{XY} , found in equations (6) and (7) is that of being a case given X and Y in the *population*. We assume that this probability can be modeled via logistic regression, i.e.,

$$p_{XY} \equiv \Pr(D=1 | X, Y) = \frac{e^{\gamma_0 + \gamma_1 \eta(G) + \gamma_2 Y}}{1 + e^{\gamma_0 + \gamma_1 \eta(G) + \gamma_2 Y}}. \quad (8)$$

Though γ_1 and γ_2 are readily estimated from case-control data, the intercept γ_0 is not. However, there are two cases where either the intercept can be estimated or where the intercept plays no role: 1) when the population prevalence is known, and 2) when the disease is rare in the population. We note that these two cases are complementary, as the population prevalence will often be unknown when the disease is rare, but reasonably accurate estimates will often be available when the disease is common.

2.2.1. Estimating p_{XY} with known population prevalence—In a classic paper, Prentice and Pyke [16] showed that logistic regression gives valid inference of population odds ratios when applied to case-control data. Thus γ_1 and γ_2 in equation (8) can be easily estimated using logistic regression applied to the case-control sample. However, the artificial distribution of cases and controls found in a case-control study will cause the case-control logistic model's intercept to differ from the true population intercept γ_0 , though the two intercepts are related. Let γ_0^* be the intercept implied by applying the logistic regression model to case-control data. Let λ denote the true population disease prevalence. It is not hard to show (see for example [17]) that γ_0^* is related to γ_0 through,

$$\gamma_0 = \gamma_0^* + \log\left(\frac{n_0}{n_1}\right) + \log\left(\frac{\lambda}{1-\lambda}\right). \quad (9)$$

This suggests, that when λ is known, we can estimate p_{XY} by

$$\hat{p}_{XY} = \frac{e^{\hat{\gamma}_0^* + \log\left(\frac{n_0}{n_1}\right) + \log\left(\frac{\lambda}{1-\lambda}\right) + \hat{\gamma}_1 \eta(G) + \hat{\gamma}_2 Y}}{1 + e^{\hat{\gamma}_0^* + \log\left(\frac{n_0}{n_1}\right) + \log\left(\frac{\lambda}{1-\lambda}\right) + \hat{\gamma}_1 \eta(G) + \hat{\gamma}_2 Y}}, \quad (10)$$

where $\hat{\gamma}_0^*$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ are the parameter estimates obtained by applying logistic regression to the case-control sample.

2.2.2. Estimating p_{XY} under a rare disease assumption—When the disease is rare in the population and assuming the logistic model (8) we have that $Pr(D=1|X, Y) \approx e^{\gamma_0 + \gamma_1 \eta(G) + \gamma_2 Y}$ and $Pr(D=0|X, Y) \approx 1$. In this case, equations (6) and (7) become

$$\tilde{U}_\beta^0 = \sum_{i=1}^{n_0} \frac{U_\beta(Y_i, X_i)}{1 - p_{X_i Y_i}} \approx \sum_{i=1}^{n_0} U_\beta(Y_i, X_i) \quad (11)$$

$$\tilde{U}_\beta^1 = \sum_{i=n_0+1}^n \frac{U_\beta(Y_i, X_i)}{p_{X_i Y_i}} \approx \sum_{i=n_0+1}^n \frac{U_\beta(Y_i, X_i)}{e^{\gamma_0 + \gamma_1 \eta(G_i) + \gamma_2 Y_i}} = e^{-\gamma_0} \sum_{i=n_0+1}^n \frac{U_\beta(Y_i, X_i)}{e^{\gamma_1 \eta(G_i) + \gamma_2 Y_i}}. \quad (12)$$

Thus we see that though we still need to estimate γ_1 and γ_2 (which is possible with case-control data), the intercept factors out of $\tilde{U}_\beta^1$ and therefore will not affect inference on β .

2.3. Combining estimators and variance estimation

Equations (6) and (7) yield two separate estimators of the same parameter β . Here we discuss an approach for combining these estimators. Let $\hat{\beta}_0$ be the estimator derived from controls (via equation 6) and $\hat{\beta}_1$ be the estimator derived from cases (via equation 7). Most often we are interested in the genetic effect associated with $\eta(G)$, which we will denote by τ , i.e., $\beta = (\alpha, \tau^T)^T$ where α is an intercept parameter and τ is the parameter that indexes the genetic effect $\eta(G)$. In order to simplify the presentation, we assume τ is a scalar, but we note that all results can be easily extended to the case where τ is a vector. Estimators of τ will be denoted by $\hat{\tau}_0$ for controls (via equation 6) and $\hat{\tau}_1$ for cases (via equation 7).

Our approach is to derive a combined estimator of the genetic effect by taking a weighted combination of the case and control estimators, i.e., $\hat{\tau} = a_0 \hat{\tau}_0 + a_1 \hat{\tau}_1$, where the weights, a_0, a_1 , are chosen to minimize the variance of $\hat{\tau}$. In the Appendix section of Supplementary Materials, we show that the weights that minimize the variance of $\hat{\tau}$ are given by

$$a^T \equiv (a_0, a_1) = \frac{\mathbf{1}^T V_{01}^{-1}}{\mathbf{1}^T V_{01}^{-1} \mathbf{1}} \quad (13)$$

where $\mathbf{1}^T = (1, 1)$ and V_{01} is the variance-covariance matrix of $\hat{\tau}_0$ and $\hat{\tau}_1$.

Even though $\hat{\tau}_0$ and $\hat{\tau}_1$ are derived from estimating equations applied to separate samples, they remain correlated since they both depend on common parameters involved in specifying p_{XY} . Thus, in estimating $\text{Var}(\hat{\tau}_0, \hat{\tau}_1)$, we must take this correlation into account. To do so, we consider joint estimation of β and γ (recall that γ is the parameter vector involved in specifying the model for p_{XY}). Specifically, consider:

$$U_{\theta,i} \equiv \begin{pmatrix} U_\gamma(D_i, X_i, Y_i) \\ I(D_i=0) \frac{U_\beta(Y_i, X_i)}{1-p_{X_i Y_i}} \\ I(D_i=1) \frac{U_\beta(Y_i, X_i)}{p_{X_i Y_i}} \end{pmatrix},$$

where $U_\gamma(D_i, X_i, Y_i)$ denotes the i th individual's contribution to the score function implied by (8). We will make use of $U_{i,\theta}$ to derive our variance estimate.

First note that although the theory of M-estimators and estimating equations is usually developed under the assumption that the sample represents a number observations of independent and identically distributed (iid) random variables, the assumption of identical distribution is not necessary. In fact, the results change only slightly and proof requires only replacement of the usual central limit theorem and law of large numbers with an application of the Lindeberg-Feller central limit theorem and a law of large numbers for non-iid random

variables. We sketch the argument here, leaving a detailed proof to the Appendix section of Supplementary Materials. Let $\hat{\theta} \equiv (\hat{\gamma}^T, \hat{\beta}_0^T, \hat{\beta}_1^T)^T$ be the root of $\sum_{i=1}^n U_{\theta,i} = 0$. Denote the ‘true’ value of the parameter by θ_0 . By use of a Taylor series expansion, we see that to first order and $n(\hat{\theta} - \theta_0) \approx A_n(\theta_0)^{-1} B_n(\theta_0) [A_n(\theta_0)^{-1}]^T$, where $A_n(\theta_0) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial U_{\theta,i}}{\partial \theta^T} |_{\theta=\theta_0}$ and $B_n(\theta_0) = \frac{1}{n} \sum_{i=1}^n U_{\theta_0,i} U_{\theta_0,i}^T$. Let $A(\theta_0), B(\theta_0)$ denote the limiting values of $A_n(\theta_0), B_n(\theta_0)$, respectively. Then it follows that $n(\hat{\theta} - \theta_0) \sim \text{AMN}(0, V(\theta_0))$, where $V(\theta_0) = A(\theta_0)^{-1} B(\theta_0) [A(\theta_0)^{-1}]^T$. A closed form expression for $V(\theta_0)$ may be difficult to obtain, nevertheless, we see that by construction, $V_n(\theta_0)$ is a consistent estimator of $V(\theta_0)$. Because $\hat{\theta}$ is a consistent estimator of θ_0 , we can consistently estimate $V(\theta_0)$ with $V_n(\hat{\theta})$.

To estimate the variance-covariance matrix of $\hat{\tau}_0$ and $\hat{\tau}_1$, we extract the sub-matrix of V_n with elements corresponding to the variance-covariance estimates of $\hat{\tau}_0$ and $\hat{\tau}_1$. We denote this sub-matrix as $\hat{V}_{01}$ and use it to calculate the weights (plugging $\hat{V}_{01}$ into equation 13), as well as obtaining an estimate of the variance of $\hat{\tau}$, which is given by $(\mathbf{1}^T \hat{V}_{01}^{-1} \mathbf{1})^{-1}$.

3. Results

3.1. Simulation studies

Using simulation, we compared the performance of the approach detailed above with four different approaches to analyzing secondary phenotypes in case-control studies: 1) the retrospective likelihood model implemented in the program Secondary Phenotype REGression (SPREG) [13]; 2) the naive, unweighted regression model specified by regressing Y on X (*NAIVE*); 3) the conditional regression model specified by regressing Y on X and D (*COND*); and 4) the *IPSW* approach of [10] and [11]. We note that our approach requires an estimate of p_{XY} and that the approach to estimating p_{XY} depends on whether or not the disease is rare and whether the population prevalence is known. Throughout, we use the approach to estimating p_{XY} that is reflective of the simulation structure. Hence, we will use the rare disease approximation (section 2.2.2), when the disease is rare (we will denote this method as *IPW_R*), and we will assume the prevalence is known (section 2.2.1), when the disease is common (which we will denote as *IPW_K*). We note that *SPREG* also utilizes a different analysis model depending on whether the disease is rare or the prevalence is known. As with *IPW*, we use the *SPREG* analysis model that best reflects the simulation structure being analyzed, which we denote by *SPREG_K*, if prevalence is assumed known, or *SPREG_R*, when a rare disease assumption is made. Since *IPSW* assumes that the sampling weights are known, we present results for *IPSW* only when the disease is common and the prevalence is known, since the sampling weights may be reasonably estimated in this case. We use *IPSW_K* to emphasize the fact that the prevalence is assumed known by this approach.

The data were simulated prospectively, through the following steps:

1. Generate X . We generate biallelic genotypes, G , assuming Hardy-Weinberg equilibrium and a minor allele frequency of 0.3. The design vector, X , is given

by $X^T = (1, \eta(G))$, where $\eta(G)$ gives the number of minor alleles composing the genotype.

2. Generate $Y|X$. Given X , we assume that the distribution of Y can be specified through the linear model

$$Y = \beta^T X + \epsilon,$$

where $\beta^T = (\alpha, \tau)$ and $\epsilon \sim N(0,1)$ or $\epsilon \sim (\chi_2^2 - 2)/2$. The different choices of ϵ are intended to allow the methods to be evaluated when the data both follow, and depart from, a normal distribution. We set that $\tau = 0$ under the null hypothesis and $\tau = -0.12$ under the alternative. The intercept α was set to one throughout.

3. Generate $D|Y, X$. Given X and Y , we assume that the distribution of D is specified via equation (8). We let $\gamma_1 \in \{0, \log(1.2), \log(1.4)\}$ and $\gamma_2 \in \{0, \log(2)/2, \log(2)\}$. Given these parameters we chose γ_0 to calibrate our simulated population disease prevalence. We conducted a large number of simulations representing a spectrum of disease prevalences. For illustration, we focus on two

values of γ_0 . We use $\gamma_0 = \log\left(\frac{0.001}{1 - 0.001}\right) \approx -6.9$ to illustrate a study with a rare disease (results in prevalences between 0.001 and 0.0033 depending on other

parameters) and $\gamma_0 = \log\left(\frac{0.1}{1 - 0.1}\right) \approx -2.2$ to illustrate a study where the disease is common (results in prevalences between 0.1 and 0.24).

4. Repeat steps 1-3 until we obtain a sample of n_1 cases and n_0 controls.

Throughout, we used $n_1 = n_0 = 1000$ and 10,000 simulated datasets for each combination of model parameters. For the *IPW* analyses, we assume $\mu(X, \beta) = \beta^T X$ and a constant working variance function.

3.1.1. Results when $\epsilon \sim N(0,1)$ —Table 1 summarizes the performance of the various methods for estimating τ when the conditional distribution of the secondary phenotype is normal, i.e., $\epsilon \sim N(0,1)$. We fixed $\gamma_2 = \log(2)/2$ in each of these comparisons and present

results for when the disease is rare, i.e., $\gamma_0 = \log\left(\frac{0.001}{1 - 0.001}\right)$ and for when the disease is

common, i.e., $\gamma_0 = \log\left(\frac{0.1}{1 - 0.1}\right)$. Type I error was computed under the null ($\tau = 0$) and power under the alternative ($\tau = -0.12$). We see that when the disease is rare, both the percent bias and the MSE are similar among *IPW_R*, *COND*, and *SPREG_R*, while *NAÏVE* shows significantly greater bias and MSE. Similarly, when the disease is common, *IPW_K* and *SPREG_K* outperform *NAÏVE* and *COND* which both show larger bias. All methods except *NAÏVE* appropriately control type I error at the nominal rate, and *NAÏVE* shows serious inflation when the disease is rare. As expected, *SPREG_R* and *SPREG_K* are most powerful in these simulations, but *IPW_R*, *IPW_K* and *COND* are competitive. The fact that *IPW_R* and *IPW_K* are competitive with *SPREG_R* and *SPREG_K* is noteworthy, as the likelihood model is correctly specified, suggesting that *SPREG* should be optimal.

Comparing the two inverse-probability approaches when the disease is common shows that while both appropriately control the type I error rate and show little bias, IPW_K has lower MSE and higher power than $IPSW_K$.

3.1.2. Results when $\epsilon \sim (\chi^2_2 - 2)/2$ —Table 2 summarizes the performance of the various methods for estimating τ when the conditional distribution of the secondary phenotype is chi-square, i.e., $\epsilon \sim (\chi^2_2 - 2)/2$. We fixed $\gamma_2 = \log(2)/2$ in each of these comparisons and

present results for when the disease is rare, i.e., $\gamma_0 = \log\left(\frac{0.001}{1 - 0.001}\right)$ and for when the

disease is common, i.e., $\gamma_0 = \log\left(\frac{0.1}{1 - 0.1}\right)$. As above, type I error was computed under the null ($\tau = 0$) and power under the alternative ($\tau = -0.12$). Percent bias is small for all models with the exception of *NAÏVE*, when the disease is rare, and *COND*, when the disease is common. Both of these methods show increasing bias with increasing γ_1 . IPW_R , IPW_K , $IPSW_K$, $SPREG_R$, and $SPREG_K$ all successfully control the type I error rate at the nominal level (0.01). *NAÏVE* show significant inflation when the disease is rare and *COND* is inflated when the disease is common. The newly proposed inverse probability weighted estimators (IPW_R and IPW_K) have lower MSE and higher power throughout. Again, comparing the two inverse-probability approaches when the disease is common shows that while both appropriately control the type I error rate and show little bias, IPW_K has lower MSE and higher power than $IPSW_K$.

3.1.3. Summary of simulation studies—We observed similar results across a large number of simulations, representing the spectrum of disease prevalences from rare to common disease. We also varied γ_2 between 0 and $\log(2)$ ($\gamma_2 = \log(2)/2$ in the above simulations). The patterns observed were entirely similar to those described above. When the conditional distribution of the secondary phenotype is normal, $SPREG_R$, $SPREG_K$, IPW_R and IPW_K show very similar performance and their performance is as good or better than any other method. Since, the *SPREG* approach represents a correctly specified likelihood-based analysis, the fact that the *IPW* approaches perform similarly is worth noting. However, when the conditional distribution of the secondary phenotype is non-normal, the *IPW* approaches perform much better than all other models under consideration. This improvement is especially pronounced when the disease is rare. Our new *IPW* approach, performed better than the previously proposed inverse-probability weighted approach, *IPSW*, showing smaller MSE and higher power throughout.

3.2. An example involving fasting blood glucose and BMI derived from the Framingham Heart Study

To illustrate our approach, we extracted a case-control sample from the Framingham Heart Study. We used obesity ($BMI > 30$) as our case definition and sampled 243 cases and 243 controls from the greater Framingham cohort. Fasting blood glucose (FBG) was also measured on these individuals and for the purpose of this illustration was considered the secondary phenotype. Because FBG is known to be highly variable and deregulated in diabetics, we excluded those from our sample. Using the Offspring and Gen 3 cohorts, we randomly selected individuals from separate pedigrees. Excluding diabetics, our initial

dataset consisted of 1114 individuals, where 243 individuals had BMI > 30 and therefore were classified as cases, and 871 individuals with BMI ≤ 30 were classified as controls. We refer to this sample of 1114 as our unrelated cohort. We then randomly selected 243 controls from those who were not obese to form our case-control study. As FBG and BMI are known to be related to one another ([18],[19]), genetic analyses of FBG will need to account for the case-control sampling to avoid bias.

Genotyping was done using Affymetrix 500K SNP arrays supplemented with the MIPS 50K array. Quality control filters included call rate (97%), Mendelian errors (< 1000) and average heterozygosity within 5 SD of the mean (between 25.758% and 29.958%). Analysis was restricted to SNPs with minor allele frequency ≥ 1%, call rate ≥ 95% and HWE $p > 10^{-6}$. We fixed the prevalence at that observed in the unrelated cohort data ($243/1114 \approx 0.2181$).

To demonstrate that our model effectively corrects for bias in the case-control sample, we compared our IPW_K analysis of the secondary phenotype for the case-control sample to a linear regression analysis (of genotype on the secondary phenotype) using the whole, unrelated cohort. If the IPW_K analysis successfully corrects for bias induced by the case control sampling, parameter estimates between these two analyses should be similar. As most of the genome will not be associated with obesity, and hence, would not induce a biased relationship between genotypes and the secondary phenotype under case-control sampling, we focused this analysis on the 100 SNPs that have the strongest association (smallest p-value) with obesity (i.e., case/control status). We performed the same analysis comparing $IPSW$, $NAïVE$ and $COND$ to a linear regression using the whole, unrelated cohort.

To make the above comparisons, we used a linear regression model to estimate slope that compares the beta estimates from each of the two approaches being compared. A slope of 1 would indicate, that on average, the two estimates are the same. Figure 1 displays the results of these comparisons. The figure indicates bias in the $COND$, $IPSW$ and $NAïVE$ approaches, but suggests bias correction in the IPW_K case. This is quantified in table 3, where we see that the IPW_K comparison has a slope of 0.99 (CI: [0.76, 1.21]) and the slopes for the $COND$, $IPSW$ and $NAïVE$ are 0.78 CI: [0.52, 1.05], 1.26 (CI: [1.02, 1.50]) and 1.32 (CI: [1.12, 1.52]), respectively.

4. Discussion

There has been a great deal of interest in estimating genetic effects of secondary phenotypes in case-controls studies. Because of this, several methods have been proposed to deal with the fact that the oversampling of cases in these studies can lead to bias. Our inverse-probability weighted estimating equation approach differs from these previous approaches and gives genetic researchers a flexible and robust approach to analyzing these data. Unlike [10] and [11], we do not require that the sampling weights be known in advance, but show how they can be estimated from the data in two complementary situations: when the disease is rare or when the population prevalence is known. Though these two cases will represent the majority of situations (common diseases will most often have good prevalence information), it is possible that a common disease will not have a good prevalence estimate.

Theoretically, the approach of [13] should handle this situation. In practice, however, *SPREG* is extremely unstable in the absence of a good prevalence estimate. As an alternative, we suggest that a sensitivity analysis could be conducted in which our *IPW* approach is used to compute estimates using several plausible prevalence values.

Though we have developed the presentation considering a continuous secondary phenotype, the modeling framework presented here is quite flexible, allowing different data types to be modeled through different mean functions. For example, when the secondary phenotype is binary, a logistic mean function may be appropriate. It is also possible to specify a working variance model to improve efficiency or to model multiple correlated secondary phenotypes via generalized estimating functions. As there are many good generalized linear model texts (see for example [21]) that highlight good choices for mean and variance functions for different data types, we will not enumerate all possibilities here, but note that correct specification of the mean function is critical to the performance of our approach (and others as well). As such, we recommend standard model diagnostics be employed to help diagnose misspecification. Misspecification of the variance function is not as critical, as it will only affect the power and not the validity of the approach. Also note that though we chose a simple logistic model for p_{XY} , more complicated models are possible, including those that model interactions between X and Y . Again, correct specification is key and standard model diagnostics should be employed.

Our approach is computationally efficient. We ran the programs on a Linux-based machine with 64-bit processing unit, CPU speed ~ 3.4 Ghz, and memory > 3.9 G. When we performed the analysis using 1000 cases and 1000 controls with 10,000 replicates, the average time to run our model was around 90 minutes. The average runtime for *SPREG* is approximately 10 times longer than our *IPW* approach. *SPREG*'s performance is particularly sluggish in analyses where the null hypothesis is true. As this is the case at most genomic sites in association studies, the approximate 10-fold decrease in computational time offered by our approach is likely conservative.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JRB, Elliott KS, Lango H, Rayner NW, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*. 2007; 316(5826):889–894. [PubMed: 17434869]
2. Loos RJF, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature genetics*. 2008; 40(6):768–775. [PubMed: 18454148]
3. Saxena, Richa, Voight, BF., Lyssenko, V., Burt, NP., de Bakker, PIW., Chen, H., Roix, JJ., Kathiresan, S., Hirschhorn, JN., Daly, MJ., et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007; 316(5829):1331–1336. [PubMed: 17463246]

4. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics*. 2008; 40(2):161–169. [PubMed: 18193043]
5. Kathiresan S, Melander O, Guiducci C, Surti A, Burt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*. 2008; 40(2):189–197. [PubMed: 18193044]
6. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JRB, Stevens S, Hall AS, et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*. 2008; 40(5):575–583. [PubMed: 18391952]
7. Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JRB, Elliott KS, Hackett R, Guiducci C, Shields B, et al. A common variant of HMG2 is associated with adult and childhood height in the general population. *Nature genetics*. 2007; 39(10):1245–1250. [PubMed: 17767157]
8. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, et al. Many sequence variants affecting diversity of adult human height. *Nature genetics*. 2008; 40(5):609–615. [PubMed: 18391951]
9. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*. 2014; 46(11):1173–1186. [PubMed: 25282103]
10. Richardson DB, Rzehak P, Klenk J, Weiland SK. Analyses of case-control data for additional outcomes. *Epidemiology*. 2007; 18:441–445. [PubMed: 17473707]
11. Monsees GM, Tamimi RM, Kraft P. Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology*. 2009; 33:717–728. [PubMed: 19365863]
12. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*. 2010; 172:1339–1348. [PubMed: 21036955]
13. Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*. 2009; 33:256–265. [PubMed: 19051285]
14. Wei J, Carroll RJ, Müller UU, Van Keilegom I, Chatterjee N. Robust estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2013; 75(1):185–206. [PubMed: 23637568]
15. Tsiatis, AT. *Semiparametric Theory and Missing Data*. New York: Springer; 2006.
16. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
17. Roeder K, Carroll RJ, Lindsay BG. A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*. 1996; 91:722–732.
18. Siegel D, Meier J, Maas C, Lopez J, Swislocki AL. The effect of body mass index on fasting blood glucose after initiation of thiazide therapy in hypertensive patients. *American Journal of Hypertension*. 2008; 21(4):438–442. [PubMed: 18246056]
19. Libby A, Meier J, Lopez J, Swislocki AL, Siegel D. The effect of body mass index on fasting blood glucose and development of diabetes mellitus after initiation of extended-release niacin. *Metabolic Syndrome Related Disorders*. 2010; 8(1):79–84. [PubMed: 19943800]
20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*. 2007; 81:000–000.
21. McCullagh, P., Nelder, JA. *Generalized linear models*. CRC press; 1989. p. 37

Figure 1.

Comparison of parameter estimates obtained from the four different methods applied to case-control samples to those obtained from the analysis of the entire unrelated cohort extracted from the Framingham Heart Study. Solid lines are regression lines comparing estimates. Dashed lines are diagonals with slope = 1.

If this image has been provided by or is owned by a third party, as indicated in the caption line, then further permission may be needed before any further use. Please contact Wiley's permissions department on permissions@wiley.com or use the rightslink service by clicking on the 'request permissions' link accompanying this article. Wiley or author owned images may be used for non-commercial purposes, subject to proper citation of the article, author, and publisher.

Table 1

% Bias, MSE, Type I error, and Power when $\epsilon \sim N(0,1)$

	γ_1	Rare disease				Common disease			
		IPW_R	$SPREG_R$	NAIVE	COND	IPW_K	$SPREG_K$	NAIVE	COND
% Bias	0	0.3333	0.0833	3.0833	0.0833	0.5000	0.4167	1.4167	1.5833
	log(1.2)	0.4167	0.8333	10.3333	0.1667	0.4167	0.1667	4.6667	5.0833
	log(1.4)	0.0833	1.1667	21.0833	0.0000	0.0833	0.6667	8.4167	12.000
MSE	0	0.0013	0.0012	0.0012	0.0012	0.0013	0.0011	0.0012	0.0018
	log(1.2)	0.0013	0.0012	0.0013	0.0012	0.0013	0.0012	0.0012	0.0017
	log(1.4)	0.0013	0.0011	0.0018	0.0011	0.0013	0.0012	0.0013	0.0017
Type I error	0	0.0106	0.0100	0.0121	0.0106	0.0112	0.0130	0.0096	0.0115
	log(1.2)	0.0107	0.0140	0.0439	0.0096	0.0113	0.0120	0.0109	0.0131
	log(1.4)	0.0093	0.0100	0.1676	0.0090	0.0128	0.0070	0.0141	0.0208
Power	0	0.7570	0.8170	0.8267	0.8152	0.7671	0.8190	0.8170	0.7998
	log(1.2)	0.7717	0.8240	0.7079	0.8276	0.7828	0.8240	0.7773	0.8727
	log(1.4)	0.7825	0.8560	0.5868	0.8404	0.7918	0.8370	0.7460	0.9210

Table 2

% Bias, MSE, Type I error, and Power when $\epsilon \sim \chi^2_{22-2}/2$

	γ_1	Rare disease				Common disease			
		IPW_R	$SPREG_R$	NAIVE	COND	IPW_K	$SPREG_K$	NAIVE	COND
% Bias	0	0.2500	0.0833	4.0833	0.5833	1.0833	1.0833	1.0000	2.4167
	$\log(1.2)$	0.3333	0.5833	15.3333	0.0000	0.4167	2.4167	2.0833	9.2500
	$\log(1.4)$	0.7500	1.3333	31.4167	0.6667	0.9167	2.1667	3.0833	20.333
MSE	0	0.0010	0.0020	0.0021	0.0020	0.0011	0.0018	0.0015	0.0015
	$\log(1.2)$	0.0010	0.0020	0.0024	0.0020	0.0011	0.0019	0.0015	0.0014
	$\log(1.4)$	0.0010	0.0018	0.0035	0.0020	0.0011	0.0018	0.0014	0.0020
Type I Error	0	0.0093	0.0060	0.0103	0.0108	0.0112	0.0100	0.0115	0.0118
	$\log(1.2)$	0.0103	0.0150	0.0200	0.0113	0.0101	0.0090	0.0092	0.0183
	$\log(1.4)$	0.0117	0.0130	0.0501	0.0102	0.0121	0.0100	0.0100	0.0393
Power	0	0.8867	0.5490	0.5748	0.5474	0.8633	0.5710	0.6957	0.6725
	$\log(1.2)$	0.8967	0.5530	0.3890	0.5666	0.8661	0.6090	0.6874	0.8128
	$\log(1.4)$	0.9041	0.5810	0.2518	0.5755	0.8703	0.6370	0.6943	0.8949

Table 3
Summary of the slopes and the standard errors of model fitting $\hat{\beta}$ s for each of the four methods

	Slope	Standard Error	95% CI
<i>IPW_K</i>	0.99	0.1139	[0.7607, 1.2163]
<i>COND</i>	0.78	0.1323	[0.5197, 1.0488]
<i>IPSW</i>	1.26	0.1205	[1.0185, 1.5005]
<i>NAÏVE</i>	1.32	0.1012	[1.1171, 1.5220]