American Journal of Epidemiology

Vol. 179, No. 10
DOI: 10.1093/aje/kwu039
Advance Access publication:
April 9, 2014

Practice of Epidemiology

Robust Estimation for Secondary Trait Association in Case-Control Genetic Studies

Jean de Dieu Tapsoba, Charles Kooperberg, Alexander Reiner, Ching-Yun Wang, and James Y. Dai*

* Correspondence to Dr. James Y. Dai, M2-C200, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA, 98109 (e-mail: jdai@fhcrc.org).

Initially submitted September 11, 2013; accepted for publication February 12, 2014.

Secondary trait genetic association provides insight into the genetic architecture of disease etiology but requires caution in estimation. Ignoring case-control sampling may introduce bias into secondary trait association. In this paper, we compare the efficiency and robustness of various inverse probability weighted (IPW) estimators and maximum likelihood (ML) estimators. ML methods have been proposed but require correct modeling of both the secondary and the primary trait associations for valid inference. We show that ML methods using a misspecified primary trait model can severely inflate the type I error. IPW estimators are typically less efficient than ML estimators but are robust against model misspecification. When the secondary trait is available for the entire cohort, the IPW estimator with selection probabilities estimated nonparametrically and the augmented IPW estimator improve efficiency over the simple IPW estimator. We conclude that in large genetic association studies with complex sampling schemes, IPW-based estimators offer flexibility and robustness, and therefore are a viable option for analysis.

case-control sampling; design consistency; inverse probability weighting; maximum likelihood

Abbreviations: AIPW, augmented inverse probability weighted; EIPW, efficient augmented inverse probability weighted; GARNET, Genomics and Randomized Trial Network; IPW, inverse probability weighted; ML, maximum likelihood; SPML, semiparametric maximum likelihood.

Contemporary case-control genetic association studies are often nested within large cohorts. In addition to the primary case-control status that drives sample selection for genotyping, the cohorts also typically have extensive measures of covariates, disease risk factors, plasma biomarkers, and other intermediate phenotypes, herein referred to as "secondary traits." After interrogation of primary case-control genetic association, interest often arises in assessing genetic association with secondary traits, exploiting genotypic data already collected to further dissect the genetic architecture of disease etiology. Over the past decade, there has been a proliferation of genome-wide genotyping studies addressing secondary trait analysis of common variants. Reported secondary trait associations include height, body mass index (weight (kg)/height $(m)^2$), and lipid levels (1, 2), often through meta-analysis of multiple studies. A timely and careful evaluation of both theoretical issues and practical considerations related to secondary trait genetic analysis is of great importance.

When the case-control status is associated with the secondary trait, an association analysis of the secondary trait is complicated by the case-control sampling scheme for genotyping. If a genetic variant is associated with the disease status, standard regression analysis that ignores the sampling scheme will lead to spurious secondary trait association (3, 4). To correct for the case-control sampling, a number of statistical methods have been proposed for secondary trait association, ranging from a naïve analysis restricted to controls only, to inverse probability weighted (IPW) estimation (3, 5), to maximum likelihood (ML) estimation (4). Our discussion herein focuses on the comparison of IPW and ML in their robustness and efficiency. Such a comparison has been discussed considerably in classical case-control association analyses (6–9). We next provide a brief summary of some perspectives.

The development of case-control methodology is one of the most important contributions statisticians have made to epidemiology (10). For primary case-control association

parameters in popular logistic regression models, semiparametric maximum likelihood (SPML) estimators can be conveniently obtained by fitting prospective likelihood to the case-control data, ignoring retrospective outcome-dependent sampling (11, 12). In SPML formulation, the distribution of missing covariates, essentially nuisance parameters relative to regression coefficients, is left completely nonparametric (12). When sampling probabilities are available, however, survey statisticians often suggest the use of inverse probability weighted (IPW) estimators (6–8), even though IPW estimators are typically less efficient than SPML estimators (13, 14). To improve the efficiency of the simple IPW estimator, in which the case-control sampling probabilities are used, a general class of semiparametric estimators based on augmented IPW estimating equations has been proposed (15), which may use kernel smoothing methods for estimating sampling probabilities, as well as adding an augmentation term (16, 17). The appeal of various IPW-based estimators stems from their robustness against model misspecification; that is, even if the regression model is wrong, IPW estimators still converge to well-defined coefficients, namely the large-sample limit of the solution of estimating equations one would have obtained had the data for the entire cohort been available. On the contrary, under model misspecification, SPML estimators may differ substantially from such coefficients (7, 8).

Recently, this view has been modified to some extent for classical case-control data when one is interested in predicting individual risk (9), in that SPML usually predicts better for a majority of individuals in the study except for those at high risk. Selection of analytical methods should therefore depend on the goal of a study. We quote from Scott and Wild (9, p. 217) to summarize this perspective: "A prescriptive approach that says that we should always use one or other approach seems wrong: the method should be tailored to the particular application."

For secondary trait association in genome-wide association studies, the trade-off between efficiency, robustness, and practicality among the aforementioned 2 approaches needs to be carefully evaluated. There are several reasons for this necessity. First, ML or SPML in this setting involves an additional nuisance model that regresses the case-control status on the genetic variant and the secondary trait, which is very likely to be misspecified. Second, the primary goal of genetic association studies is to test whether there is a genetic association (so that properly controlling for false positive findings is imperative) and less commonly to predict individual risk. Third, secondary trait data may be available for everyone in the cohort, whereas genotyping data are available only for a case-control sample. This scenario has not been investigated and compared between IPW-based methods and the ML method. Fourth, secondary traits often come from a complex sampling scheme rising simply because of convenience. Application of the ML method can sometimes be computationally prohibitive, whereas IPW-based estimators remain viable in complex sampling.

Our motivational example comes from the Women's Health Initiative Study, one of the largest and farthest-reaching studies of women's health ever undertaken in the United States, harboring several large-scale case-control genetic studies, including the Genomics and Randomized Trial Network (GARNET) Study, to identify genetic risk alleles for myocardial infarction, stroke, venous thrombotic disease, and type 2 diabetes (18). After primary analyses, investigators were interested in genetic associations with blood pressure, which was measured yearly; approximately 1 million single nucleotide polymorphisms were genotyped in case-control samples based on the 4 different but slightly overlapping diseases and a shared control sample. The implementation of ML methods for this sampling scheme and the longitudinal secondary trait is difficult, whereas the simple IPW method coupled with generalized estimating equations for repeated measures is easy to apply. Moreover, blood pressure measurement is cheap and available for all participants in the cohort. It is of interest to investigate how to leverage secondary traits that are always observed as opposed to genotypes, which are available only in case-control samples.

In this paper we compare the ML-based methods and the various IPW-based methods in efficiency and robustness for secondary trait genetic association. We consider 2 representative scenarios of practical importance for assessment in simulations. The first is a classical setting in which only the case-control status is observed for everyone in the cohort, and all other variables of interest including the secondary trait and the genotype are measured only for a case-control sample. We derive the most efficient IPW estimator and compare it with the ML estimator and the simple IPW estimator. The second scenario is motivated by the GARNET Study, in which a continuous secondary trait is always observed together with the case-control status, and genotypes are collected only in a case-control sample. In this setting, we explore several IPW-derived methods including augmented IPW estimators, using kernel estimators of selection probabilities to leverage the always-observed secondary trait and investigate potential efficiency gain.

METHODS

Consider a case-control study nested in a cohort of n subjects. All participants in the cohort were ascertained for a dichotomous clinical endpoint D, with $D = 1$ coded for disease (case) and $D = 0$ for no disease (control). For assessment of genetic association with D, a case-control sample was drawn from the cohort. Let R be the indicator variable for whether a participant was included in the case-control sample. Let G denote the genetic variant, coded as 0, 1, or 2 depending on the number of variant alleles. Suppose there are also a secondary trait variable Y and a vector of confounding variables V to be adjusted for (e.g., top principal components from a genome-wide genetic data set and age). Typically for rare diseases, all cases and a small proportion of controls are included for genotyping. Suppose the interest is in assessing the secondary trait association in the following model:

$$\mathbb{E}(Y|G,V) = g(\beta_0 + \beta_1 G + \beta_2 V) = g(\boldsymbol{\beta}^T \chi), \qquad (1)$$

where β_1 is the genetic association of interest, $\chi = (1,G,V)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, and g is the expit function when Y is dichotomous or the identity function when Y is a quantitative trait.

Depending on whether the secondary trait Y is continuous or dichotomous and on whether Y is available for the entire

cohort or the case-control sample, there are several common scenarios for assessing secondary trait association. For conciseness of exposition, we select the following 2 representative scenarios for discussing methods:

- In scenario 1, the secondary trait is dichotomized and available only for the case-control sample. The data for a participant in the cohort are (D,RY,RG,RV), where (Y,G,V) are missing at random, so that $\Pr(R=1|D,Y,G,V)=\Pr(R=1|D)$. This is the standard scenario discussed by Monsees et al. (3), Lin and Zeng (4), and Jiang et al. (5).
- In scenario 2, the secondary trait is continuous and available for the entire cohort. The data vector for a participant in the cohort is (D,Y,RG,RV), where (G,V) are missing at random in that $\Pr(R=1|D,Y,G,V)=\Pr(R=1|D)$. This is the scenario motivated by the GARNET Study for genetic association with blood pressure. Secondary traits are often cheap to measure, and thus readily available for every participant.

In scenario 1, we showcase the robustness of IPW methods relative to ML methods, whereas in scenario 2 we explore potential efficiency gain when exploiting the information in the always-observed continuous secondary trait. Other study scenarios, such as a continuous secondary trait available only for case-control samples or a dichotomous secondary trait available for everyone, present similar settings for methodology treatment and for comparison of robustness and efficiency. We thus briefly discuss these alternative scenarios after the main method presentation for scenario 1 and below.

Naïve complete-case estimator

For either logistic regression or linear regression in the secondary trait association (equation 1), the estimating function for a subject is written as $U=\chi^T\{Y-g(\boldsymbol{\beta}^T\chi)\}$, where U is the estimating function. The naïve complete-case estimator solves the estimating equation, $\sum_i^n R_iU_i=0$. The fundamental problem of the complete-case estimator is that this estimating equation generally does not have 0 expectation (i.e., $\mathbb{E}(RU)\neq 0$) if the sampling process R is related to U. Exceptions do exist, however. It is useful to list the conditions under which the complete-case estimator remains unbiased. Denote by $\perp$ the stochastic independence of 2 random variables. It has been shown that any 1 of the following 3 conditions is sufficient to guarantee the consistency of the complete-case estimator (4):

- $D\perp Y|\chi$. Violation of this condition is the very reason for being concerned about secondary trait association, which can be tested a priori. When the secondary trait is in the etiological pathway to the disease outcome, or is a phenotype after the disease onset, this condition could be violated.
- $D\perp\chi|Y$, if g is the expit function. For a dichotomous secondary trait, the estimation of β_1 will not be distorted by ignoring the sampling if there is no genetic association with the primary disease status (3, 4). The impact of biased sampling is merely shifting the intercept in equation 1 because of the multiplicative risk model.
- $D\perp\chi|Y$ and $Y\perp\chi$, if g is the identity link. This condition says that if there is no genetic association for either the primary trait or the secondary trait, then $\beta_1=0$ is still consistently estimated, and the type I error is properly controlled, even if there is correlation between D and Y. In other words, the complete-case estimator provides a valid global test for no genetic association with either a primary or secondary trait.

In a genome-wide association study, the majority of genetic variants are null, for which the complete-case estimator does not introduce bias. Correction for the biased sampling is needed for those variants that are indeed associated with the primary trait.

ML estimation

ML estimation, and SPML estimation in particular, is well developed for case-control and, more generally, 2-phase sampling studies (12–14). The key element of SPML estimation is that the distribution for missing covariates, indexed by nuisance parameters, is left nonparametric. For secondary trait association, several forms of ML estimation have been developed with various nuisance models (4, 5). Let $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2)$ denote regression coefficients of the inference model (equation 1). The likelihood for the data arising in scenario 1 can be derived as

$$\{\Pr_\alpha(D|Y,G,V)\Pr_{\boldsymbol{\beta}}(Y|G,V)\Pr_f(G,V)\}^{R=1}\times\left\{\int_{G,V,Y}\Pr_\alpha(D|Y,G,V)\Pr_{\boldsymbol{\beta}}(Y|G,V)\Pr_f(G,V)dgdvdy\right\}^{R=0},$$

and the likelihood for the data arising in scenario 2 can be derived as

$$\{\Pr_\alpha(D|Y,G,V)\Pr_{\boldsymbol{\beta}}(Y|G,V)\Pr_f(G,V)\}^{R=1}\times\left\{\int_{G,V}\Pr_\alpha(D|Y,G,Y)\Pr_{\boldsymbol{\beta}}(Y|G,V)\Pr_f(G,V)dgdv\right\}^{R=0},$$

where α is the parameter indexing the distribution $\Pr(D|Y,G,V)$, and f is the distribution of the covariate vector (G,V). Both α and f are nuisance parameters relative to $\boldsymbol{\beta}$, among which α is typically formulated in a parametric logistic model (4, 5), and f is typically left nonparametric as in the classical SPML estimation for 2-phase sampling studies. The inference model can be logistic or linear regression in either likelihood; therefore, the formulation above is applicable beyond the 2 scenarios we provided.

The validity of SPML estimation hinges on correct modeling of both the inference model $\Pr_{\boldsymbol{\beta}}(Y|G,V)$ and the nuisance model $\Pr_\alpha(D|Y,G,V)$. The former could be difficult for a continuous Y because it requires the entire distribution of Y given G and V, whereas the latter could also be challenging if the secondary trait Y lies on the complex etiological pathway from genetic mutation G to disease outcome Y, so that Y could mediate the genetic association, modify the genetic association, or modify the association of V and D. When the secondary trait is collected through complex sampling, correct specification of $\Pr_\alpha(D|Y,G,V)$ may not be possible (e.g., in the aforementioned GARNET Study, there are 4 case groups and 1 shared control group). This dependence on a parametric nuisance model in secondary trait association is

in sharp contrast to the classical SPML estimation developed in the case-control association, in which the only nuisance model is the nonparametric distribution of missing covariates. Sometimes $\Pr_\alpha(D|Y,G,V)$ can be estimated nonparametrically (e.g., when Y, G, and V are all discrete). However, as we will show in the simulation study under scenario 1, SPML estimation yields nearly the same efficiency as the simple IPW method when the nuisance disease risk model is close to being saturated.

Computation of SPML estimation for secondary trait association resembles that of SPML estimation for primary trait association, treating α in $\Pr(D|Y,G,V)$ as additional regression parameters (13, 14, 19). Potentially high-dimensional, nonparametric covariate distribution can be eliminated through the profile likelihood approach (4, 19). Alternatively, as we implemented in our simulation for scenario 1, one can use the expectation-maximization algorithm to simultaneously estimate regression parameters and nonparametric point masses posited on each observed covariate value (20). The variance of estimated parameters was computed by numerical differentiation of the information matrix for the observed data.

Various IPW-based methods

In contrast, the validity of IPW-based estimators depends only on the correct specification of sampling probabilities. Denote by π_i the sampling probability for the ith subject. The simple IPW estimator solves

$$\sum_i^n \frac{R_i}{\pi_i} U_i = 0. \tag{2}$$

Even if $g(\chi\boldsymbol{\beta}^T)$ in the estimating function is not correctly specified, the solution of equation 2 still converges to a well-defined quantity, namely the solution of $\mathrm{E}(U) = 0$. This is the parameter one would have estimated had data from the entire cohort been observed. From the perspective that all models are to some extent misspecified, such a parameter is of practical use as an interpretable cohort-based estimand for association. This is the so-called "design-consistency" property advocated by survey statisticians (7, 8), which is not shared by ML estimators. Note that design consistency does not mean unbiasedness in large samples, because parameter estimates from a misspecified model are not directly interpretable. In simple language, it means that the IPW estimator approximates the inference drawn from the full cohort if the sampling probabilities are correctly specified. Because the sampling is controlled by investigators in case-control genetic studies nested in a cohort, π_i is almost always known. The simple IPW estimator is, thus, always consistent in this regard, though its efficiency can be substantially inferior to a SPML estimator when the models are indeed correctly specified (13, 14).

Strategies to improve the efficiency of the simple IPW estimator while preserving their design-consistency property have been proposed in the statistical literature (15, 16). Let W denote the vector of variables that are always observed (i.e., $W = D$ in scenario 1, and $W = (D,Y)^T$ in scenario 2). One way to improve efficiency of the simple IPW estimator is to replace the known sampling probabilities in equation 2 with the estimated sampling probabilities given W, denoted by $\hat{\pi}(W_i)$

$$\sum_i^n \frac{R_i}{\hat{\pi}(W_i)} U_i = 0. \tag{3}$$

This is easily accomplished when W is discrete. When W contains continuous variables, for example Y is continuous, it is convenient to estimate $\pi(W_i)$ consistently using the nonparametric Nadaraya-Watson kernel smoother (21, 22), given by

$$\hat{\pi}(W_i) = \frac{\sum_{j=1}^n R_j K_h(W_i - W_j)}{\sum_{j=1}^n K_h(W_i - W_j)}, \tag{4}$$

where $K_h(\cdot)$ is a kernel function with bandwidth h. With the proper choice of h, the asymptotic behavior of the estimator solving equation 3 was presented by Wang et al. (16).

More generally, a class of semiparametric estimators based on augmented inverse probability weighted (AIPW) estimating equations was proposed by Robins et al. (15). The optimal estimator in this class attains the semiparametric variance bound, in our notation solving

$$\sum_i^n \frac{R_i}{\pi(W_i)} h_{\mathrm{eff}}(\mathcal{X}_i)\{Y_i - g(\boldsymbol{\beta}^T\mathcal{X}_i)\} + \left(1 - \frac{R_i}{\pi(W_i)}\right)\mathbb{E}[h_{\mathrm{eff}}(\mathcal{X}_i)\{Y_i - g(\boldsymbol{\beta}^T\mathcal{X}_i)\}|W_i] = 0,$$

where $h_{\mathrm{eff}}(\chi)$ is the unique solution to the functional equation shown in proposition 4.2 in the article by Robins et al. (15). Generally speaking, $h_{\mathrm{eff}}(\chi)$ is difficult to estimate unless W is discrete. It requires modeling the true data-generating distribution including, in our case, the primary trait association model. In scenario 1, where D is the only variable observed for everyone, the most efficient augmented inverse probability weighted (EIPW) estimator in this class can be derived following section 5.2 of the article by Robins et al. (15). When $\pi(Y_i = 1) = 1$, we show in the Appendix that the derivation of the EIPW estimator is further simplified. For scenario 2, in which Y is continuous and available for everyone, computation of the most efficient AIPW estimator is difficult. One AIPW estimator for scenario 2 that does not require extensive computation (17), is given by

$$\sum_i^n \frac{R_i}{\pi(W_i)} U_i + \left(1 - \frac{R_i}{\pi(W_i)}\right)\mathbb{E}(U_i|W_i) = 0, \tag{5}$$

where $\pi(W_i)$ is estimated by equation 4, and $\mathrm{E}(U_i|W_i)$ is estimated by

$$\frac{\sum_{j=1}^n R_j U_j K_h(W_i - W_j)}{\sum_{j=1}^n R_j K_h(W_i - W_j)}. \tag{6}$$

Note that all of these IPW-based estimators preserve the design-consistency property. Solving for the IPW estimator in equations 3 and 4 can be implemented by any standard regression packages allowing individual weights. Solving for the AIPW and the EIPW estimators involves the Newton-Raphson algorithm, using the multiroot function in R statistical software (R Foundation for Statistical Computing,

Vienna, Austria), for example. Their variances can be estimated by the robust sandwich method (15–17), with the empirical variance of estimating functions in the center and the inverse of the gradient of estimating functions in the 2 sides.

Other study scenarios

Other sampling scenarios beyond scenarios 1 and 2 require minor modification for maximum likelihood methods. The scenario in which a continuous secondary trait is available only in the case-control sample can be treated similarly to scenario 1, with a linear inference model for $\Pr(Y|G,V)$. Numerical integration of (Y,G,V) may be needed to evaluate the likelihood of a participant who wasn't included in the case-control sample. A scenario in which a dichotomized secondary trait is always observed can be treated similarly to scenario 2, with a logistic inference model for $\Pr(Y|G,V)$. IPW-based methods are particularly simple in these 2 scenarios. Because the always-observed variables are discrete, the AIPW estimator degenerates to the simple IPW estimator with estimated weights (15, 17), which means no additional efficiency can be gained by adding the augmented term to equation 3. Similar to scenario 1, the optimal IPW-based estimator can be obtained by computing $h_{\text{eff}}(\chi)$ explicitly following section 5.2 of the article by Robins et al. (15).

RESULTS

We simulated a secondary trait study nested in a cohort with 10,000 subjects according to either scenario 1 or scenario 2. In scenario 1, we assume that both the secondary trait Y and the disease outcome D are dichotomized, generated from the following models, respectively:

$$\text{logit}\{\mathbb{E}(Y|G,V)\} = -2 + \beta_1 G + V,$$

$$\text{model 1}: \quad \text{logit}\{\mathbb{E}(D|Y,G,V)\} = -4.5 + \log(2)Y + \log(1.5)G + 0.5YG + V, \tag{7}$$

$$\text{model 2}: \quad \text{logit}\{\mathbb{E}(D|Y,G,V)\} = -4.5 + \log(2)Y + \log(1.5)G + V. \tag{8}$$

Robustness against model misspecification was assessed when 1 of the model terms was omitted: either the interaction YX in model 1 or the confounding variable V in model 2. The genetic variant G is in Hardy-Weinberg equilibrium with minor allele frequency 0.3, coded as 0, 1, or 2. A continuous confounding variable V was generated by the normal distribution $N(0.5G, 1)$, so that G and V are correlated. The disease status D was observed for every subject, but (Y,G,V) were observed only in a case-control sample. The cases were sampled with probability 1, and then the same number of controls was randomly selected. For each model 7 and 8 and for different values of β_1 (0 or log (1.5)), 1,000 data sets were generated with approximately 500 cases and 500 controls. In Table 1, we compare 4 estimators in terms of their finite-sample properties: the naïve complete-case estimator, the simple IPW estimator, the EIPW estimator, and the SPML estimator.

In Table 1, where there is an interaction between Y and G on disease risk in model 1, the complete-case estimator is severely biased because the case-control status is strongly correlated with the genetic variant and the secondary trait. The IPW estimator does not depend on specification of nuisance

Table 1. Finite-Sample Properties of Various Estimators for Secondary Trait Association in Scenario 1[a], When the Primary Trait Association is Generated by Model 1[b]

β_1	Property	CC Estimator	IPW Estimator	Correct Model		Omitting *YG*	
				EIPW Estimator	SPML Estimator	EIPW Estimator	SPML Estimator
0	Bias	0.310	0.010	0.008	0.011	0.011	0.244
	Var	0.012	0.025	0.025	0.023	0.025	0.012
	$\widehat{\text{Var}}$	0.012	0.028	0.028	0.024	0.027	0.012
	95% CP[c]	0.186	0.950	0.954	0.950	0.954	0.396
	Type I error	0.814	0.050	0.046	0.050	0.046	0.604
Log(1.5)	Bias	0.303	0.003	0.002	0.005	0.003	0.223
	Var	0.012	0.023	0.023	0.020	0.022	0.012
	$\widehat{\text{Var}}$	0.011	0.023	0.023	0.020	0.023	0.012
	95% CP[c]	0.194	0.946	0.950	0.958	0.948	0.458

Abbreviations: CC, complete-case; CP, coverage probability; EIPW, efficient inverse probability weighted; IPW, inverse probability weighted; SPML, semiparametric maximum likelihood; Var, sample variance of the estimator; $\widehat{\text{Var}}$, mean of estimated variances.

[a] In scenario 1, the secondary trait is dichotomized and available only for the case-control sample. The data for a participant in the cohort are (D,RY,RG,RV), where (Y,G,V) are missing at random, so that $\Pr(R=1|D,Y,G,V) = \Pr(R=1|D)$. This is the standard scenario discussed by Monsees et al. (3), Lin and Zeng (4), and Jiang et al. (5).

[b] Model 1: $\text{logit}\{E(D|Y,G,V)\} = -4.5 + \log(2)Y + \log(1.5)G + 0.5YG + V$.

[c] Coverage probability of 95% confidence interval.

models and is always consistent. The EIPW estimator uses model 1 to compute the efficient score, as we show in the Appendix, but it is robust against model specification, as the semiparametric theory dictates. Interestingly, there is little improvement in efficiency from the IPW estimator to the EIPW estimator. This is possibly because $h_{\text{eff}}(G)$ in the optimal estimating equation is close to a linear combination of G when G is coded (0,1,2). If G is binary, then any function $h(G)$ is also a linear combination of G, resulting in equivalent estimating functions. When the primary trait model is correctly specified, the SPML estimator is consistent, the variance estimates reflect true variation, and the 95% confidence intervals show proper coverage. For hypothesis testing at α level 0.05, the empirical type I errors for IPW, EIPW, and SPML estimation are all approximately 0.05. When the primary trait model is misspecified, however, the SPML estimator shows a sizable bias relative to its standard deviation. The type I error of SPML estimation increases to 60.4%, a 12-fold inflation, which is quite alarming.

In Table 2, the impact of omitting V in model 2 to SPML estimation is less severe, yet the type I error is still 4 times as much as it should be. The bias of the complete-case estimator is less severe in this parameter setting. In either misspecified model, IPW and EIPW methods remain consistent in estimation and preserve proper control of type I error. Note that the efficiency comparisons among IPW, EIPW, and SPML methods under correct specification of models 1 and 2 are quite different. When the true model contains the interaction between G and Y, SPML estimation is slightly more efficient than IPW or EIPW estimation, whereas under model 2, there is more than 50% efficiency gain from IPW estimation to SPML estimation. This is because equation 7 is close to a saturated nonparametric model. Similar phenomena were observed previously (5). In results not presented, we found that if there is no V in model 1, and G is dichotomous, IPW, EIPW, and SPML methods are identical.

In Figure 1, we investigate the sensitivity of SPML estimation in hypothesis testing across a gradient of model misspecification. Clearly omitting the interaction term YG in equation 7 yields severe inflation of false positives, even if the size of the interaction is small. Test validity is less sensitive when we omit a confounding variable V. One would need a log (2) effect size from the continuous V to double the type I error.

We next simulated scenario 2, in which both D and Y are observed for everyone in the cohort, but (G,V) are observed only in the case-control sample. We let Y be continuous, generated either from a Gaussian distribution or a standardized T distribution with 6 degrees of freedom as follows:

$$\begin{aligned}\text{model 3}: \quad & Y = 0.5 + \beta_1 G + V + \epsilon, \epsilon \sim \mathcal{N}(0,1)\\ \text{logit}\{\mathbb{E}(D|Y,G,V)\} &= -2.8 + \log(2)Y + \log(2)G\\ &\quad -0.5V \ + \log(1.5)YG, \end{aligned} \tag{9}$$

$$\begin{aligned}\text{model 4}: \quad & Y = 0.5 + \beta_1 G + V + \epsilon, \quad \epsilon \sim \text{standardized } T(6)\\ \text{logit}\{\mathbb{E}(D|Y,G,V)\} &= -2.8 + \log(2)Y + \log(2)G - 0.5V. \end{aligned} \tag{10}$$

The distribution of (G,V) was generated similarly as in scenario 1, and 1,000 data sets were generated with approximately 500 cases and 500 controls. In Tables 3 and 4, we compared the following 6 estimators on the basis of finite-sample properties: the complete-case estimator, the simple IPW estimator, the kernel-assisted IPW estimator (denoted IPW_K) in which selection probabilities were estimated by

Table 2. Finite-Sample Properties of Various Estimators for Secondary Trait Association in Scenario 1[a], When the Primary Trait Association is Generated by Model 2[b]

β_1	Property	CC Estimator	IPW Estimator	Correct Model		Omitting V	
				EIPW Estimator	SPML Estimator	EIPW Estimator	SPML Estimator
0	Bias	0.033	−0.001	0.000	0.001	−0.002	−0.148
	Var	0.015	0.037	0.037	0.015	0.037	0.017
	$\widehat{\text{Var}}$	0.015	0.037	0.036	0.015	0.037	0.015
	95% CP[c]	0.950	0.948	0.946	0.954	0.950	0.784
	Type I error	0.050	0.052	0.054	0.046	0.050	0.216
Log(1.5)	Bias	0.039	0.000	0.001	0.003	−0.001	−0.137
	Var	0.014	0.033	0.033	0.015	0.033	0.016
	$\widehat{\text{Var}}$	0.014	0.031	0.031	0.014	0.031	0.015
	95% CP[c]	0.930	0.954	0.960	0.950	0.956	0.776

Abbreviations: CC, complete-case; CP, coverage probability; EIPW, efficient inverse probability weighted; IPW, inverse probability weighted; SPML, semiparametric maximum likelihood; Var, sample variance of the estimator; $\widehat{\text{Var}}$, mean of estimated variances.

[a] In scenario 1, the secondary trait is dichotomized and available only for the case-control sample. The data for a participant in the cohort are (D,RY,RG,RV), where (Y,G,V) are missing at random, so that $\Pr(R=1|D,Y,G,V) = \Pr(R=1|D)$. This is the standard scenario discussed by Monsees et al. (3), Lin and Zeng (4), and Jiang et al. (5).

[b] Model 2: $\text{logit}\{E(D|Y,G,V)\} = -4.5 + \log(2)Y + \log(1.5)G + V$.

[c] Coverage probability of 95% confidence interval.

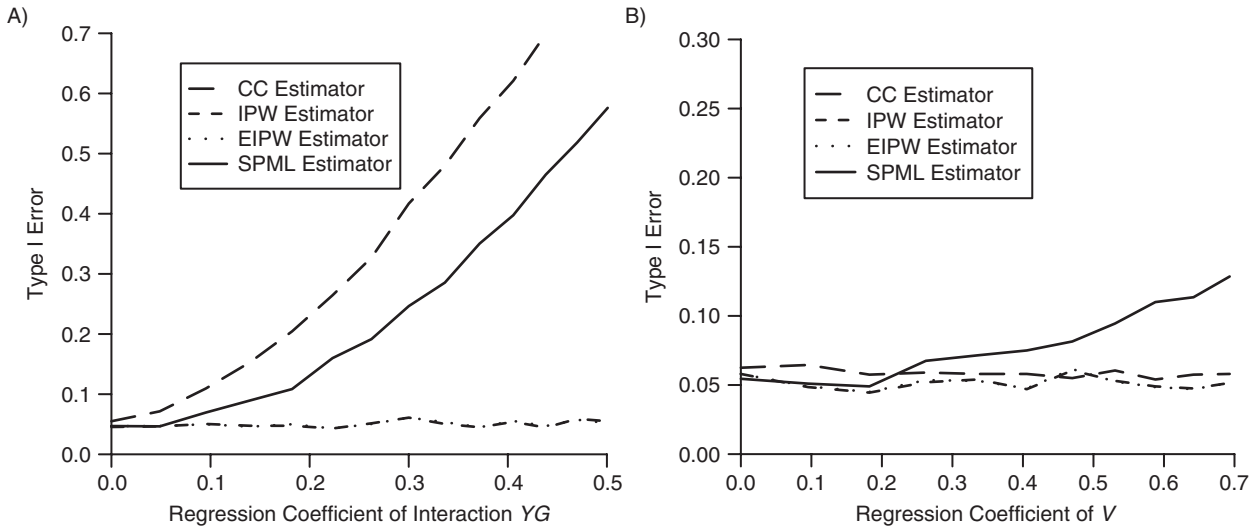

Figure 1. Sensitivity analysis to assess the impact of model misspecification on the type I error of testing secondary trait association. A) Type I error rate for a range of regression coefficients for interaction *YG* in model 1. B) Type I error rate for a range of regression coefficients for *V* in model 2. CC, complete-case; EIPW, efficient inverse probability weighted; IPW, inverse probability weighted; SPML, semiparametric maximum likelihood.

equation 4, the augmented IPW estimator solving equation 5, and the ML estimator under correct or wrong specification. We used the Gaussian kernel with the bandwidth $4\sigma_{Y|D}n^{-1/3}$ (23), in which $\sigma_{Y|D}$ is the standard deviation of Y given D. In computing the ML estimator, we used the Gauss-Hermite quadrature method for integration of V. Robustness against model misspecification was assessed when either the interaction YG in model 3 was omitted, or the distribution of Y, $T(6)$ in model 4 was misspecified to be Gaussian.

Under model 3 and when the interaction term was omitted, the bias of the ML estimator is quite substantial, resulting in a

Table 3. Finite-Sample Properties of Various Estimators for Secondary Trait Association in Scenario 2[a], When the Primary Trait Association is Generated by Model 3[b]

β_1	Property	CC Estimator	IPW Estimator	IPW_K Estimator[c]	AIPW Estimator	ML Estimator	
						Correct Model	Omitting *YG*
0	Bias	−0.1381	−0.0011	−0.0013	−0.0011	0.0027	−0.1279
	Var	0.0024	0.0039	0.0038	0.0040	0.0024	0.0041
	$\widehat{\text{Var}}$	0.0026	0.0041	0.0039	0.0039	0.0034	0.0025
	95% CP[d]	0.2280	0.9600	0.9540	0.9560	0.9820	0.3080
	Type I error	0.7720	0.0400	0.0460	0.0440	0.0180	0.6920
−log(2)	Bias	−0.1468	−0.0011	0.0047	0.0035	0.0023	−0.0908
	Var	0.0023	0.0040	0.0034	0.0034	0.0021	0.0032
	$\widehat{\text{Var}}$	0.0025	0.0041	0.0033	0.0032	0.0027	0.0016
	95% CP[d]	0.1800	0.9520	0.9580	0.9500	0.9780	0.4100

Abbreviations: AIPW, augmented inverse-probability weighted; CC, complete-case; CP, coverage probability; IPW, inverse probability weighted; ML, maximum likelihood; Var, sample variance of the estimator; $\widehat{\text{Var}}$, mean of estimated variances.

[a] In scenario 2, the secondary trait is continuous and available for the entire cohort. The data vector for a participant in the cohort is (D,Y,RG,RV), where (G,V) are missing at random in that $\Pr(R=1|D,Y,G,V)=\Pr(R=1|D)$. This is the scenario motivated by the GARNET Study for genetic association with blood pressure (18). Secondary traits are often cheap to measure, and thus readily available for every participant.

[b] Model 3: $\epsilon \sim \mathcal{N}(0,1)$; $\text{Logit}\{E(D|Y,G,V)\} = -2.8 + \log(2)Y + \log(1.5)G - 0.5V + \log(1.5)YG$.

[c] Inverse probability weighted estimator with selection probabilities estimated by kernel smoothers.

[d] Coverage probability of 95% confidence interval.

Table 4. Finite-Sample Properties of Various Estimators for Secondary Trait Association in Scenario 2[a], When the Primary Trait Association is Generated by Model 4[b]

β_1	Property	CC Estimator	IPW Estimator	IPW_K Estimator[c]	AIPW Estimator	ML Estimator	
						Correct Model ϵ	Wrong Model ϵ
0	Bias	0.0347	−0.0014	−0.0006	−0.0009	−0.0064	−0.0067
	Var	0.0026	0.0038	0.0036	0.0038	0.0018	0.0017
	$\widehat{\text{Var}}$	0.0027	0.0036	0.0032	0.0032	0.0020	0.0021
	95% CP[d]	0.9160	0.9360	0.9460	0.9400	0.9620	0.9660
	Type I error	0.0840	0.0640	0.0540	0.0600	0.0380	0.0340
−log(2)	Bias	0.0131	−0.0013	0.0080	0.0030	0.0004	0.0591
	Var	0.0031	0.0041	0.0034	0.0034	0.0014	0.0019
	$\widehat{\text{Var}}$	0.0030	0.0041	0.0030	0.0030	0.0016	0.0017
	95% CP[d]	0.9520	0.9580	0.9440	0.9460	0.9620	0.7220

Abbreviations: AIPW, augmented inverse-probability weighted; CC, complete-case; CP, coverage probability; IPW, inverse probability weighted; ML, maximum likelihood; Var, sample variance of the estimator; $\widehat{\text{Var}}$, mean of estimated variances.

[a] In scenario 2, the secondary trait is continuous and available for the entire cohort. The data vector for a participant in the cohort is (D,Y,RG,RV), where (G,V) are missing at random in that $\Pr(R=1|D,Y,G,V)=\Pr(R=1|D)$. This is the scenario motivated by the GARNET Study for genetic association with blood pressure (18). Secondary traits are often cheap to measure, and thus readily available for every participant.

[b] Model 4: $\epsilon \sim$ standardized $T(6)$; $\text{Logit}\{E(D|Y,G,V)\} = -2.8 + \log(2)Y + \log(1.5)G - 0.5V$.

[c] Inverse probability weighted estimator with selection probabilities estimated by kernel smoothers.

[d] Coverage probability of 95% confidence interval.

very inflated type I error (of 0.692) and dismal performance in the coverage probabilities in Table 3. For model 4 in Table 4, misspecification of $T(6)$ to a Gaussian distribution does not cause much bias under the null, but the bias under the alternative hypothesis is still sizable, leading to poor coverage probability (of 0.722). All IPW-based estimators were consistent. These observations are consistent with results in Table 1. Interestingly, in the settings in which there is secondary trait association, the variances of the IPW_K and AIPW estimators decrease by approximately 15%–20% relative to the variance of the simple IPW estimator, demonstrating the advantage of leveraging secondary trait data that are available for everyone. It is also interesting to observe that the IPW_K and AIPW estimators yield nearly identical performances, consistent with the theoretical results reported by Wang and Wang (17) that the 2 estimators are asymptotically equivalent. In other simulation settings not shown, the AIPW estimator can be slightly advantageous over the IPW_K estimator in finite sample performance.

DISCUSSION

In the context of case-control studies for secondary trait genetic association, we compared the efficiency and robustness of ML estimators and various IPW-based estimators. The new twist of the long-standing IPW-ML comparison is that ML estimation requires a nuisance case-control risk model. We showed in simulations that, when the nuisance risk model is incorrectly specified, ML or SPML estimation can be severely biased and can, thus, sometimes produce a drastic inflation of type I error. To increase the robustness of likelihood-based methods, one may consider a nonparametric model for the nuisance disease risk model, but that may yield nearly the same efficiency as IPW-based estimators (Table 1). On the other hand, IPW-based methods are robust and easy to implement, offering a competitive alternative approach.

Along with secondary traits, always-observed data often include additional demographic factors and other disease risk predictors. When there are high-dimensional always-observed data, some of which are categorical, nonparametric kernel smoothing approaches can be problematic to implement. With careful model fitting, one could consider parametric logistic regression for estimating sampling probabilities, thereby further improving the efficiency of estimation.

For genome-wide association studies, sample sizes are usually large, possibly assembled through meta-analysis. The trade-off between bias and efficiency may tilt toward reducing bias and properly controlling false positives, particularly when the secondary trait is a quantitative trait with an irregular distribution. We show in simulation that slight misspecification of the density function of the secondary trait could also introduce bias. Furthermore, the availability of secondary trait data often depends on a complex outcome-dependent sampling process. The ML estimator can be computationally difficult, if not impossible, for a complex sampling scheme and high-dimensional adjusting covariates, whereas IPW-based methods can be implemented for virtually any sampling scheme.

When the secondary trait is available for the entire cohort, we show that the IPW estimator with selection probabilities estimated by kernel smoothers and the AIPW estimator perform similarly, both yielding a 15%–20% efficiency gain

over the simple IPW estimator. These methods exploit the extra information in the secondary trait and remain robust against model misspecification, and thus should be used whenever applicable. In particular, the kernel-assisted IPW estimator is much more applicable in genome-wide studies, because sampling probabilities can be estimated once for all genetic variants.

ACKNOWLEDGMENTS

Author affiliations: Public Health Science Division, Fred Hutchinson Cancer Research Center, Seattle, Washington (Jean de Dieu Tapsoba, Charles Kooperberg, Ching-Yun Wang, James Y. Dai); Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Research Center, Seattle, Washington (James Y. Dai); Department of Biostatistics, University of Washington, Seattle, Washington (Charles Kooperberg, Ching-Yun Wang, James Y. Dai); and Department of Epidemiology, University of Washington, Seattle, Washington (Alexander Reiner).

This work was supported by the National Institutes of Health (grants P01 CA53996, R01 HL114901, R01 HG00 6164, and R01 ES017030).

Conflict of interest: none declared.

REFERENCES

1. Lettre G, Jackson AU, Gieger C, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet*. 2008;40(5):584–591.
2. Loos RJ, Lindgren CM, Li S, et al. Common variants near *MC4R* are associated with fat mass, weight and risk of obesity. *Nat Genet*. 2008;40(6):768–775.
3. Monsees GM, Tamimi RM, Kraft P. Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol*. 2009;33(8):717–728.
4. Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol*. 2009; 33(3):256–265.
5. Jiang Y, Scott AJ, Wild CJ. Secondary analysis of case-control data. *Stat Med*. 2006;25(8):1323–1339.
6. Godambe VP, Thompson ME. Parameters of superpopulation and survey population: their relationships and estimation. *Int Stat Rev*. 1986;54(2):127–138.
7. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice based sampling. *J R Stat Soc Series B Methodol*. 1986; 48(2):170–182.
8. Xie Y, Manski CF. The logit model and response-based samples. *Sociol Methods Res*. 1989;17(3):283–302.
9. Scott AJ, Wild CJ. On the robustness of weighted methods for fitting models to case-control data. *J R Stat Soc Series B Methodol*. 2000;64(2):207–219.
10. Breslow NE. Statistics in epidemiology: the case-control study. *J Am Stat Assoc*. 1996;91(433):14–28.
11. Anderson JA. Separate sample logistic discrimination. *Biometrika*. 1972;59(1):19–35.
12. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1986;66(3):403–411.
13. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J R Stat Soc Series B Methodol*. 1997;59(2):447–461.
14. Lawless JF, Kalbfleisch JD, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. *J R Stat Soc Series B Stat Methodol*. 1999;61(2):413–438.
15. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846–866.
16. Wang CY, Wang S, Zhao LP, et al. Weighted semiparametric estimation in regression analysis with missing covariate data. *J Am Stat Assoc*. 1997;92(438):512–525.
17. Wang S, Wang CY. A note on kernel assisted estimators in missing covariate regression. *Stat Probab Lett*. 2001;55(4): 439–449.
18. National Human Genome Research Institute. Genomics and Randomized Trials Network (GARNET). https://www.genome.gov/27541119. Accessed September 1, 2013.
19. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. *Biometrika*. 1997;84(1):57–71.
20. Lawless JF. Likelihood and pseudo likelihood estimation based on response-biased observation. Proceedings of the Georgia Symposium on Estimation Functions. Hayward, CA: Institute of Mathematical Statistics; 1997:43–56.
21. Nadaraya EA. On estimating regression. *Theory Probab Appl*. 1964;9(1):141–142.
22. Watson GS. Smooth regression analysis. *Sankhya Ser A*. 1964; 26(4):359–372.
23. Qi L, Wang CY, Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. *J Am Stat Assoc*. 2005;100(472):1025–1263.

APPENDIX

Translating proposition 4.2 from the article by Robins et al. (15) to our notation, the optimal $h_{\text{eff}}(\chi)$ is the unique solution of the functional equation

$$h_{\text{eff}}(\mathcal{X}) = \{\partial g(\mathcal{X};\beta)/\partial\beta\}t(\mathcal{X}) + \mathbb{E}\left[\left\{\frac{1-\pi}{\pi}\phi_{\text{eff}}(W)\right\}\varepsilon|\mathcal{X}\right]t(\mathcal{X}), \qquad (11)$$

where $\varepsilon = Y - g(\boldsymbol{\beta}^T\chi)$, $t(\chi) = \{\mathbb{E}(\varepsilon^2/\pi|\chi)\}^{-1}$, and $\phi_{\text{eff}}(W) = \mathbb{E}\{h_{\text{eff}}(\chi)\varepsilon|W\}$. In scenario 1, the always-observed variable W is D, Y is binary, and $\varepsilon = Y - g(\chi;\beta)$. When the sampling fraction for cases is 1, $(1-\pi/\pi)\phi_{\text{eff}}(D=1) = 0$, and so we need only to compute $\phi_{\text{eff}}(D=0)$. Multiply equation 11 by ε, take conditional expectation given D, and we have

$$\mathbb{E}[\partial g(\mathcal{X};\beta)/\partial\beta\}t(\mathcal{X})\varepsilon|D] = \phi_{\text{eff}} - \mathbb{E}\left\{\mathbb{E}\left[\left\{\frac{1-\pi}{\pi}\phi_{\text{eff}}(W)\right\}\varepsilon|\mathcal{X}\right]t(\mathcal{X})|D\right\}.$$

Let $t(\mathcal{X}) = \{\sum_{D,Y}\Pr(D|Y,\mathcal{X})\Pr(Y|\mathcal{X})\varepsilon^2/\pi\}^{-1}$, and let $m_0 = \mathbb{E}[\partial g(\mathcal{X};\beta)/\partial\beta\}t(\mathcal{X})\varepsilon|D=0]$, computed by empirical average. Let $l = \{\sum_{D,Y}\Pr(D|Y,\mathcal{X})\Pr(Y|\mathcal{X})\varepsilon(1-\pi)/\pi\}$. Then, $\phi_{\text{eff}}(D=0) = m_0/[1-\mathbb{E}\{t(\mathcal{X})\varepsilon l|D=0\}]$, and hence $h_{\text{eff}}(\chi)$ is obtained by equation 11. Note that the risk model for the primary trait $\Pr(D|Y,\chi)$ is used in the computation.