Communications for Statistical Applications and Methods
2019, Vol. 26, No. 2, 103–129

https://doi.org/10.29220/CSAM.2019.26.2.103
Print ISSN 2287-7843 / Online ISSN 2383-4757

A review of analysis methods for secondary outcomes in case-control studies

Elizabeth D. Schifano[1,a]

[a]Department of Statistics, University of Connecticut, USA

Abstract

The main goal of a case-control study is to learn the association between various risk factors and a primary outcome (e.g., disease status). Particularly recently, it is also quite common to perform secondary analyses of the case-control data in order to understand certain associations between the risk factors of the primary outcome. It has been repeatedly documented with case-control data, association studies of the risk factors that ignore the case-control sampling scheme can produce highly biased estimates of the population effects. In this article, we review the issues of the naive secondary analyses that do not account for the biased sampling scheme, and also the various methods that have been proposed to account for the case-control ascertainment. We additionally compare the results of many of the discussed methods in an example examining the association of a particular genetic variant with smoking behavior, where the data were obtained from a lung cancer case-control study.

Keywords: ascertainment, data reuse, retrospective study, sampling bias, selection bias

1. Introduction

Retrospective case-control studies are important epidemiological designs for analyzing relatively rare diseases (e.g., Prentice and Pyke, 1979; Breslow and Day, 1980; Schlesselman, 1981; Rothman, 1986), with the primary purpose of elucidating the relationship between disease occurrence and various exposures or risk factors (covariates). Case-control designs typically provide tremendous savings both in terms of time and money over prospective studies, but can still remain quite costly. To take full advantage of the collected information, researchers often additionally perform secondary analyses of the case-control data in order to understand certain associations between the disease risk factors. Particularly with the recent explosion of genome-wide association studies (GWAS) that often adopt a case-control design, it is becoming more common for researchers to not only study genetic associations with the disease outcome on which the case-control sampling was based, but also examine genetic associations with additional clinical traits (i.e., those serving as risk factors or covariates in the regression model for disease) measured on the same set of subjects. More generally, in this article we refer to the binary disease status (case/control outcome) as the *primary outcome*, the additional collected traits as *secondary outcomes*, and the predictor variables of interest as *exposures*.

Because disease-affected cases are over-sampled relative to the disease-free controls, the case-control sample does not represent a random sample of the general population. Despite this, it has been well-established that performing standard (prospective) logistic regression on case-control data produces correct maximum likelihood estimates of odds ratios of disease risk, and that only the disease rate (intercept) is generally biased (Prentice and Pyke, 1979). This result, however, only applies

[1] Department of Statistics, University of Connecticut, 215 Glenbrook Road, Storrs, CT, 06269, USA.
E-mail: elizabeth.schifano@uconn.edu

Published 31 March 2019 / journal homepage: http://csam.or.kr

to the primary disease outcome defining the case-control status, and not to secondary outcomes in general. There have consequently been a number of methods discussed and proposed to analyze secondary outcomes, which can be broadly grouped into the following categories: (a) naive methods that do not explicitly account for the sampling design; (b) bias correction methods using weighting schemes; and (c) joint modeling of the primary and secondary outcomes, explicitly accounting for the sampling design. We describe methods in each of these categories in detail in this review. Prior to 2009, most methods were developed only for binary secondary outcomes, with focus usually on binary exposure variables. The emergence of case-control GWAS sparked renewed interest in secondary outcome analyses, and spurred on the development of methods for both continuous and binary secondary outcomes with more general types of exposures.

The structure for the remainder of this article is as follows. Section 2 is broken up into several subsections, first establishing notation, and then discussing methods in each of the three categories described above. In Section 3, we compare results across many of the methods in an example examining the association of a particular genetic variant with smoking behavior, where the data were obtained from a lung cancer case-control study. The review concludes with a brief discussion.

2. Methods for analyzing secondary outcomes

2.1. Notation

Let D denote the disease status or primary outcome, and suppose there are n_0 disease-free controls with $D = 0$ and n_1 affected cases with $D = 1$. Let Y denote the secondary outcome, which may be binary or continuous. In the situation of a single exposure variable interest, let X be this exposure variable (e.g., minor allele count of a genetic variant in GWAS), and $\mathbf{Z}$ be the vector of additional covariates in the secondary model, with effects on the secondary outcome given by β_X and $\boldsymbol{\beta}_Z$, respectively. The goal of a secondary outcome analysis is then to examine the association between X and Y, possibly adjusting for additional covariates $\mathbf{Z}$. In some works, all regression coefficients of the secondary outcome model are of interest. For notational convenience, we use $\mathbf{X}$ to denote the $q \times 1$ dimensional covariate vector with effects on the secondary outcome given by $\boldsymbol{\beta}$.

2.2. Naive methods

In this work, naive methods for analyzing secondary outcomes refer to any standard (linear or logistic) regression techniques applied to the case-control data that do not explicitly account for the biased sampling scheme. These types of methods have also been referred to as *ad hoc* methods (Jiang *et al.*, 2006), and include (i) analyzing the combined sample of cases and controls without including case-control status as a covariate, (ii) analyzing the combined sample of cases and controls including case-control status as a covariate, and (iii) performing analyses within the case and/or control groups separately. As many authors have demonstrated (e.g., Nagelkerke *et al.*, 1995; Jiang *et al.*, 2006; Richardson *et al.*, 2007; Monsees *et al.*, 2009; Lin and Zeng, 2009; Wang and Shete, 2011a), these standard regression methods applied to case-control data may only safely be used in restrictive settings, and can otherwise bias inference.

Nagelkerke *et al.* (1995) discussed the problem for a binary secondary outcome Y in the context of logistic regression, and provided conditions in which the biased case-control sampling could be ignored. They show that if the secondary outcome Y and the case-control variable D are conditionally independent given the covariates $\mathbf{X}$ (i.e., D depends on Y only through $\mathbf{X}$), then ordinary logistic regression will yield valid estimates. By 'valid', they mean that the regression coefficient estimates for $\mathbf{X}$ obtained from the case-control study consistently estimate their population values. Additionally,

if D and $\mathbf{X}$ are conditionally independent given Y (i.e., D depends on $\mathbf{X}$ only through Y), then all regression coefficients except the intercept term will be valid.

A popular alternative when analyzing both cases and controls together is to include disease status as a predictor in the secondary outcome model. In the context of a binary secondary outcome Y, this means the effect of exposure X is adjusted for disease status D, as well as for possibly other covariates $\mathbf{Z}$ in the logistic regression model. Wang and Shete (2011a) show in simulation, however, that adjusting for disease status can still result in a biased estimate of the odds ratio (OR) of the exposure variable ($\text{OR}_{XY} \equiv \exp(\beta_X)$), and the magnitude of the bias depends on the prevalence of both the primary and secondary outcomes in the general population. As also described in Reilly *et al.* (2005), the results of such an adjusted analysis will not be valid if the original case status D modifies the effect of exposure X on Y. Even in situations in which there is no interaction, the adjusted estimates may not be scientifically relevant or meaningful (Lee *et al.*, 1997; Jiang *et al.*, 2006).

Another simple approach for analyzing secondary outcomes involves performing the analysis on the control subjects only. This strategy is appropriate only when the disease is rare, in which the control sample can be considered an approximation to a random sample from the general population (e.g., Jiang *et al.*, 2006; Monsees *et al.*, 2009; Lin and Zeng, 2009; Li *et al.*, 2010). This method is generally inefficient, however, as the information from the cases is completely ignored. Case-only estimators have similarly been investigated in simulation under the rare disease assumption (i.e., when disease prevalence in the population is assumed close to zero) in Monsees *et al.* (2009), Lin and Zeng (2009), Li *et al.* (2010), but notably Monsees *et al.* (2009), Lin and Zeng (2009) conducted their simulations under the assumption of no interaction between the exposure variable X and secondary outcome Y on disease risk. More specifically, they consider a primary disease model of the form

$$\text{logit}\{P(D = 1|Y, X)\} = \delta_0 + \delta_X X + \delta_Y Y. \tag{2.1}$$

A more general model allows for an interaction between X and Y on disease risk:

$$\text{logit}\{P(D = 1|Y, X)\} = \delta_0 + \delta_X X + \delta_Y Y + \delta_{XY} XY. \tag{2.2}$$

When the disease is rare, the relationship in (2.2) may be approximated as

$$P(D = 1|Y, X) \approx \exp(\delta_0 + \delta_X X + \delta_Y Y + \delta_{XY} XY). \tag{2.3}$$

Under the logistic regression

$$\text{logit}\{P(Y = 1|X)\} = \beta_0 + \beta_X X, \tag{2.4}$$

Li *et al.* (2010) focus on inference regarding $\log(\text{OR}_{XY}) = \beta_X$ in the general population, where $X \in \{0, 1\}$, $Y \in \{0, 1\}$. Viewing the observed case and control cell frequency vectors as realizations from two independent multinomial distributions, the estimator of β_X using cases only, denoted by $\hat{\beta}_{XCA}$, is *not* unbiased in general as the OR of X and Y among the cases is $\text{OR}_{XY}^{D=1} = \text{OR}_{XY} \exp(\delta_{XY})$. Li *et al.* (2010) further show that under (2.3), the maximum (retrospective) likelihood estimation of β_X in this simple setting is equivalent to analyzing data from the controls alone (see Section 2.4).

Indeed, if the association between the primary outcome D and secondary outcome Y does not depend on exposure variable X, then both the cases-only and controls-only estimators can provide (approximately) valid estimates of the exposure effect when the disease is rare. Consequently, in this situation these two estimates can be combined with inverse-variance weighting to yield a meta-analytic-type estimator. In the dichotomous Y and X scenario, this amounts to

$$\hat{\beta}_{XW} = w_{cc}\hat{\beta}_{XCO} + (1 - w_{cc})\hat{\beta}_{XCA}, \tag{2.5}$$

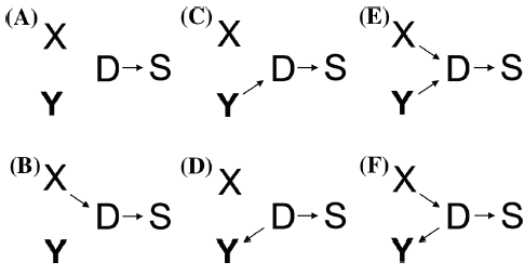

Figure 1: *The directed acyclic graphs (DAGs) showing six possible scenarios of association of X, Y, D, and sampling indicator S in the underlying population. (Figure taken from Ray and Basu (2017), with scenarios (A)–(F) originally considered and depicted in Monsees et al. (2009).)*

where $\hat{\beta}_{XCO}$ is the controls-only estimator with variance estimate $\hat{\sigma}^2_{CO}$, $\hat{\beta}_{XCA}$ was discussed above with variance estimate $\hat{\sigma}^2_{CA}$, and $w_{cc} = \hat{\sigma}^2_{CA}/(\hat{\sigma}^2_{CA} + \hat{\sigma}^2_{CO})$. The variance of $\hat{\beta}_{XW}$ can be approximated by $\hat{\sigma}^2_{CA}\hat{\sigma}^2_{CO}/(\hat{\sigma}^2_{CA} + \hat{\sigma}^2_{CO})$, leading to a more efficient estimate of β_X than $\hat{\beta}_{XCO}$, but it is only unbiased when $\delta_{XY} = 0$ (Li *et al.*, 2010).

Monsees *et al.* (2009) report the presence or absence of bias of naive methods that condition on case-control ascertainment and/or disease status in the six scenarios depicted in Figure 1 based on conditional probability rules. It is assumed that sampling mechanism S depends only on the case-control variable D, with the direction of an arrow indicating a causal relationship. Table 1 summarizes the conclusions from Monsees *et al.* (2009) regarding bias under the null (no direct effect of X on Y) or alternative hypothesis (direct effect of X on Y). They further examine the magnitude of the bias and its effect on inference in simulation. With continuous secondary outcome Y, the secondary outcome model considered in Monsees *et al.* (2009), and also Lin and Zeng (2009), is given by

$$Y = \beta_0 + \beta_X X + \epsilon, \tag{2.6}$$

where $\epsilon \sim N(0, \sigma^2)$. Assuming disease model (2.1) and a genomic exposure variable X, i.e., X = number of minor alleles of a diallelic single nucleotide polymorphism (SNP), Monsees *et al.* (2009) illustrated that the degree of bias in estimating β_X is dependent on the rarity of the primary disease, the strength of the association between the secondary and primary outcomes, and the association between the exposure variable and the primary outcome. From their simulations, they make the following conclusions about the three ((i)–(iii)) types of naive analysis discussed above for scenarios A–C and E in Figure 1:

- Bias: None of the naive analyses are biased when $\delta_Y = 0$. When $\delta_Y \neq 0$, the analyses can be biased, although for rare disease, naive methods (ii) and (iii) have no perceptible bias when $\delta_X = 0$. The bias for naive method (i) is small when δ_Y is modest. When both the secondary outcome Y and the exposure variable X are independently associated with disease risk ($\delta_Y \neq 0$ and $\delta_X \neq 0$), the unadjusted analysis of naive method (i) can be noticeably biased.

- Type I error: All naive methods had the correct Type I error rate for testing $H_0 : \beta_X = 0$ provided either $\delta_Y = 0$ or $\delta_X = 0$. For rare disease, they did not detect inflation in Type I error for naive methods (ii) and (iii), even when both $\delta_Y \neq 0$ and $\delta_X \neq 0$. For more common disease, however, they did observe inflation when $\delta_Y \neq 0$ and $\delta_X \neq 0$. Tests from naive method (i) had considerable Type I error inflation whenever $\delta_Y \neq 0$ and $\delta_X \neq 0$, and the severity of inflation decreased as the disease prevalence increased (and the sample became more representative of the full cohort).

Table 1: Presence or absence of bias in measures of $X - Y$ association under the null and alternative hypotheses of naive analyses that condition on case-control ascertainment and/or disease status (Table reproduced from Table 1 in Monsees *et al.* (2009))

		Conditioning Event (C)				Conditioning Event (C)	
		S^a	$\{S, D\}^b$			S^a	$\{S, D\}^b$
Scenario A	Null	No bias	No bias	Scenario D	Null	No bias	No bias
	Alternative[c]	No bias	No bias		Alternative	Bias	Bias
Scenario B	Null	No bias	No bias	Scenario E	Null	Bias	Bias
	Alternative	No bias	No bias		Alternative	Bias	Bias
Scenario C	Null	No bias	No bias	Scenario F	Null	Bias	No bias
	Alternative	Bias[d]	Bias		Alternative	Bias	Bias

[a]Conditions only on ascertainment, $S = 1$; e.g., analyses that ignore case-control sampling. [b]Conditions on ascertainment (by restricting to case-control sample) and case-control status D, e.g., analyses restricted to controls (or cases) or stratified by case-control status. [c]Relationships among X, Y, and D as in the corresponding null scenario, with the addition of a directed edge from X to Y. [d]Measures of the $X - Y$ relationship conditional on C may not reflect the $X - Y$ relationship in the general population.

Similarly assuming $\delta_{XY} = 0$, Lin and Zeng (2009) show theoretically that if the secondary outcome is not related to the case-control status (i.e., $\delta_Y = 0$), then all three naive methods are valid. If the exposure variable X is not associated with the case-control status (i.e., $\delta_X = 0$), then all three naive methods yield correct estimates of ORs for dichotomous traits, but the least-squares estimates for quantitative traits produced by the three methods are biased unless $\beta_X = 0$ or $\delta_Y = 0$. When the disease is rare ($< 2\%$), they conclude naive methods (ii) and (iii), but not (i) are approximately valid.

2.3. Weighting methods

2.3.1. Inverse probability weighting

Horvitz and Thompson (1952) introduced the concept of inverse probability weighting (IPW) in the context of sample surveys. Reilly *et al.* (2005) discuss the approach in the context of a two-stage epidemiological design (e.g., Breslow and Cain, 1988; Flanders and Greenland, 1991; Zhao and Lipsitz, 1992; Reilly, 1996), in which "easy" (cheap) measurements are obtained for a sample in the first stage, and more difficult (expensive) measurements are obtained for a subsample of these subjects in the second stage. The depth of the second-stage sampling within the strata defined by the first-stage variables can induce bias, but provided that the sampling probabilities depend only on variables from the first-stage, valid estimates at the second stage may be obtained by implementing a weighted analysis. Comparisons of this type of weighted procedure and various types maximum-likelihood procedures have been well-studied in the analysis of primary outcomes (e.g., Scott and Wild, 1986, 2002; Breslow *et al.*, 2013). Reilly *et al.* (2005) note that in the analysis of *secondary* outcomes, one can think of the available data from the case-control sample as the second stage of the study, sampled within the strata defined by disease status from the first stage. Thus, the idea is to weight each individual's contribution to the score or estimating equations from the secondary outcome model by the reciprocal of the individual's selection probability. Jiang *et al.* (2006) and Richardson *et al.*, (2007) discuss this approach in the analysis of binary secondary traits using weighted logistic regression, while Monsees *et al.* (2009) consider IPW for more general types of secondary outcomes and exposures. The IPW approach has also been successfully used in the context of secondary survival outcomes in nested case-control studies (e.g., Kim and Kaplan, 2014).

For concreteness, let S_i be an inclusion indicator in the case-control sample, and assume that the probability of selection into the sample depends only on disease status, such that weight w_i is given by

$w(D_i) = 1/P(S_i = 1|D_i, Y_i, \mathbf{X}_i)$. Let $\mu(\mathbf{X}_i;\boldsymbol{\beta}) = E(Y_i|\mathbf{X}_i)$, where the expectation is taken with respect to the population. Following Sofer *et al.* (2017a), the IPW estimation problem takes the general form

$$\sum_{i=1}^{n} U_{\text{IPW},i}(\boldsymbol{\beta}) = \sum_{i=1}^{n} S_i w(D_i) h(\mathbf{X}_i)[Y_i - \mu(\mathbf{X}_i;\boldsymbol{\beta})] = \mathbf{0}, \tag{2.7}$$

where $U_{\text{IPW},i}$ is an unbiased estimating function corresponding to the secondary outcome model for parameters collected in $\boldsymbol{\beta}$, and $h(\mathbf{X}_i)$ is a $q \times 1$ user-specified function such that $E(\partial U_{\text{IPW},i}/\partial\boldsymbol{\beta})$ is invertible. Jiang *et al.* (2006), Richardson *et al.*, (2007), and Monsees *et al.* (2009) take $U_i \equiv h(\mathbf{X}_i)[Y_i - \mu(\mathbf{X}_i;\boldsymbol{\beta})]$ to be the score function in logistic or linear regression, where $h(\mathbf{X}_i) = \mathbf{X}_i'$.

With correct specification of the sampling probabilities, using IPW has the advantages of being robust to sampling bias and being relatively easy to implement in standard software, so the notion of IPW can readily be extended to more complicated models involving estimating functions (see, e.g., Schifano *et al.*, 2013; Sofer *et al.*, 2017b). IPW can be inefficient, however, as compared to other methods that explicitly account for the sampling design (e.g., Jiang *et al.*, 2006; Richardson *et al.*, 2007; Monsees *et al.*, 2009; Wang and Shete, 2011a) and requires known sampling weights, such as in nested case-control studies within a prospective cohort, or when external information about disease prevalence is available (e.g., Wang and Shete, 2011a). For retrospective case-control studies, Wang and Shete (2011a) introduce the "extended IPW" approach, in which M individuals are assumed in the finite general population. With a disease prevalence π, the sampling fractions for cases and controls are $n_1/(\pi M)$ and $n_0/[(1-\pi)M]$, respectively. Setting the weight as 1 for cases, the weight for the controls is $n_1(1-\pi)/(n_0\pi)$ which is free of M. Richardson *et al.* (2007) noted that the IPW approach is best suited for large case-control studies in which the primary outcome is not extremely rare.

Sofer *et al.* (2017a) proposed an approach to improve upon the efficiency of the IPW estimator using a so-called control function. The approach still requires the use of known sampling fractions, but instead of using (2.7), they propose solving the estimating equation:

$$\sum_{i=1}^{n} S_i w(D_i) \left(h_1(\mathbf{X}_i)[Y_i - \mu(\mathbf{X}_i;\boldsymbol{\beta})] - h_2(\mathbf{X}_i, D_i)\right) = \mathbf{0}, \tag{2.8}$$

where $h_2(\mathbf{X}_i, D_i)$ is a $q \times 1$ vector control function satisfying $E\{S\,w(D)h_2(\mathbf{X}, D)|\mathbf{X}\} = 0$. Since (2.8) is inverse probability weighted and has mean zero for all h_2, the left-hand side of (2.8) is also unbiased. Clearly taking $h_2(\mathbf{X}_i, D_i) = 0$ reduces (2.8) to (2.7), but clever selection of $h_2(\mathbf{X}, D) \neq 0$, as well as h_1 can lead to a more efficient estimator. The authors show that $h_2(\mathbf{X}, D)$ may be expressed in terms of a selection bias function (Tchetgen Tchetgen, 2014, see also Section 2.4.2 below) with either the identity or log link functions. Specifically for the identity link, they consider $h_2(\mathbf{X}, D) = \gamma(\mathbf{X})[D - P(D = 1|\mathbf{X})]$, where $P(D = 1|\mathbf{X})$ is the conditional probability of disease given covariates in the target population and $\gamma(\mathbf{X}) = E(Y|D = 1, \mathbf{X}) - E(Y|D = 0, \mathbf{X})$ is the selection bias function. Since $P(D = 1|\mathbf{X})$ is unknown, they further study the semiparametric efficiency implications of imposing nonparametric, semiparametric, or parametric models for $P(D = 1|\mathbf{X})$. Sofer *et al.* (2017a) show that their proposed estimators are indeed more efficient than the standard IPW estimator when the disease model is either parametric or semiparametric, while also being unbiased provided the population mean model and conditional disease models are correctly specified; no other distributional assumptions are required for the outcomes. Their approach is implemented in the R package `RECSO` (Sofer, 2013).

Both Sofer *et al.* (2017a) and Tapsoba *et al.* (2014) note that the efficiency of the IPW-type estimators can also generally be improved in nested case-control studies, in which additional information can be obtained from the underlying cohort. In such settings, augmented inverse-probability weighted

estimators (AIPW) (Robins *et al.*, 1994) can be used. For N representing the total number of subjects in the *cohort* (versus $n(< N)$ samples in the case-control study), and W_i denoting the variables observed for all $i = 1, \ldots, N$, Tapsoba *et al.* (2014) propose an AIPW estimator solving

$$\sum_{i=1}^{N} \frac{S_i}{\pi_i} U_i + \sum_{i=1}^{N} \left(1 - \frac{S_i}{\pi_i}\right) E(U_i|W_i) = 0,$$

where $\pi_i = E(S_i|W_i)$ and $U_i = h(\mathbf{X}_i)[Y_i - \mu(\mathbf{X}_i;\boldsymbol{\beta})]$ as before. In simulation, Tapsoba *et al.* (2014) compare the standard IPW, AIPW, and most efficient augmented probability weighted (EIPW) estimators with ML-based methods (specifically, SPML2 methods; see Section 2.4.1 below) in terms of efficiency and robustness for secondary trait genetic association (i.e., X is a genetic variant, coded as 0, 1, or 2 depending on the number of variant alleles). Consistent with the results of Jiang *et al.* (2006) discussed in more detail in Section 2.4.1 below, they show that 1) when the primary disease model is incorrectly specified ML (SPML2) estimation can be severely biased and can consequently produce large inflation of type I error in some situations, 2) a nonparametric model for the nuisance disease risk model (SPML1) may increase robustness, but yields nearly the same efficiency as the weighted (IPW, AIPW, EIPW) estimators, 3) the weighted estimators were all robust against model misspecification, 4) there was little improvement in efficiency from the IPW estimator to the EIPW estimator for binary secondary traits and when only disease status information is available for all subjects within the cohort, and 5) when the (continuous) secondary trait is available for the entire cohort, their proposed AIPW estimator yielded a 15–20% efficiency gain over the standard IPW approach.

The general framework proposed by Xing *et al.* (2016) is also based on inverse-probability weighted estimating equations, but differs fundamentally from the sampling-probability approaches discussed above. Instead, the approach of Xing *et al.* (2016) inversely weights by a subject's probability of being a case (or a control), conditional on their exposure and secondary trait information. Let $U_\beta(Y_i, X_i) = h(X_i)[Y_i - \mu(X_i;\beta)]$ for each subject i, $i = 1, \ldots, n$, where $h(X_i) = \partial\mu(X_i;\beta)/\partial\beta V^{-1}(X_i)$ and $V(X_i)$ is a working model for $\mathrm{Var}(Y_i|X_i)$. They define the following weighted estimating functions

$$\tilde{U}_\beta^0 = \sum_{i=1}^{n_0} \frac{U_\beta(Y_i, X_i)}{1 - p_{X_i,Y_i}}, \qquad \tilde{U}_\beta^1 = \sum_{i=1}^{n_1} \frac{U_\beta(Y_i, X_i)}{p_{X_i,Y_i}},$$

where p_{XY} denotes the probability of being a case given X and Y in the population. This probability is a population quantity, but it can be estimated from a case-control sample either when the disease is rare or when the population disease prevalence is known; see Xing *et al.* (2016) Section 2.2 for details. Let $\hat{\beta}_X^{(d)}$ be the estimator of the exposure effect of interest obtained from solving $\tilde{U}_\beta^d = \mathbf{0}$, $d = 0, 1$. Noting that $\hat{\beta}_X^{(0)}$ and $\hat{\beta}_X^{(1)}$ are estimating the same parameter β_X, Xing *et al.* (2016) propose a combined estimator by taking a weighted combination, i.e., $\hat{\beta}_X^W = a_0\hat{\beta}_X^{(0)} + a_1\hat{\beta}_X^{(1)}$, where weights a_i are such that $a_0 + a_1 = 1$ and are chosen to minimize the variance of $\hat{\beta}_X^W$. Their simulations considered a continuous secondary outcome Y under scenarios in which the disease was rare with prevalence unknown (IPW_R), the disease was common with disease prevalence known (IPW_K), with normal or non-normal (χ^2) random errors, and using disease model (2.1) to estimate the weights. Their weighting approach performs better than the traditional sampling-weight IPW approach in simulations, having both higher power and smaller mean-squared error for the estimates of β_X. Although their method was slightly less efficient than the likelihood-based method of Lin and Zeng (2009) (see Section 2.4.1 below) when the likelihood is correctly specified, it was also more robust to deviations from normality for the error distribution. Fortran code for running their simulations is available in their Supplemental Materials.

2.3.2. Counterfactual weighted estimating equations

Wei *et al.* (2016) and Song *et al.* (2016a) consider solving different weighted estimating equations (WEE) that combine observed and pseudo (counterfactual) outcomes in order to provide unbiased estimation in the analysis of secondary outcomes. Counterfactual outcomes are often used in causal inference, where the counterfactual outcome is the potential (unobserved) outcome of a subject if he were assigned an *alternative* treatment or exposure group than what he actually experienced. In the context of secondary outcomes, the idea is to define the potential (unobserved) secondary outcome of each subject if his case-control status was reversed, and then use both the observed and counterfactual outcomes to estimate the exposure-secondary outcome association using the case-control sample.

Let $S(Y, \mathbf{X}, \boldsymbol{\beta})$ be an estimating function such that for randomly selected subjects in the general population $E_Y[S(Y, \mathbf{X}, \boldsymbol{\beta}^*)|\mathbf{X}] = \mathbf{0}$ at the true value $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. For example, $S(Y, \mathbf{X}, \boldsymbol{\beta})$ can be taken as $h(\mathbf{X}_i)[Y_i - \mu(\mathbf{X}_i; \boldsymbol{\beta})]$ as in (2.7), which includes score functions for generalized linear models (GLM) as a special case. The authors also consider quantile regression, for which $S(Y, \mathbf{X}, \boldsymbol{\beta}) = [\tau - I\{Y \leq \boldsymbol{\beta}'\mathbf{X}\}]\mathbf{X}$ for any quantile $\tau \in (0, 1)$. As indicated in Section 2.2, solving $S(Y, \mathbf{X}, \boldsymbol{\beta}) = \mathbf{0}$ directly in a case-control sample can lead to biased estimates of the population effects. Conditioning on D, however,

$$\begin{aligned} E_Y[S(Y, \mathbf{X}, \boldsymbol{\beta}^*)|\mathbf{X}] &= E_Y[S(Y, \mathbf{X}, \boldsymbol{\beta}^*)|\mathbf{X}, D = 0]P(D = 0|\mathbf{X}) + E_Y[S(Y, \mathbf{X}, \boldsymbol{\beta}^*)|\mathbf{X}, D = 1]P(D = 1|\mathbf{X}) \\ &= \mathbf{0}. \end{aligned}$$

If for each Y_i in the sample, $i = 1, \ldots, n$, we were also able to observe its counterfactual outcome $\tilde{Y}_i$ under the *alternate* disease status, then the corresponding sample estimating equations would be

$$\sum_{i=1}^{n} \left[S(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) P(D_i|\mathbf{X}_i) + S\left(\tilde{Y}_i, \mathbf{X}_i, \boldsymbol{\beta}\right) P(1 - D_i|\mathbf{X}_i) \right] = \mathbf{0}, \tag{2.9}$$

where the weights $P(D_i|\mathbf{X}_i)$ and $P(1 - D_i|\mathbf{X}_i)$ are the conditional probabilities of being the observed disease status and counterfactual (alternate) disease status, respectively, given $\mathbf{X}_i$. Both the counterfactual secondary outcomes and the weights in (2.9) are unknown, however, and need to be estimated in practice. Song *et al.* (2016a) propose two approaches for estimating $\tilde{Y}_i$ depending on the linearity of $S(Y_i, \mathbf{X}_i, \boldsymbol{\beta})$ in Y_i; see their manuscript for details. To estimate the weight, one could use models from the primary analysis, or assume a logistic model of the form $\text{logit}\{P(D = 1|\mathbf{X})\} = \gamma_0 + \gamma'\mathbf{X}$, with the estimate of γ_0 appropriately tuned to match the overall disease prevalence in the general population. For inference, the authors propose using a bootstrap procedure to obtain standard error estimates for the regression coefficients to take into account the uncertainty from the estimated weights and counterfactual outcomes. Their approach is implemented in the R package `WEE` (Song *et al.*, 2016b). The authors show in simulation with relatively common disease that their methods are robust to misspecification of disease prevalence and disease model (i.e., inclusion of $X \times Y$ interaction), and that the traditional IPW approach provides robust and similar efficient estimates to those of their proposed approach when Y given D is a random sample. If the sampling scheme is more complicated, however, their proposed approach performs better than IPW. They further warn, as have others (e.g., Tapsoba *et al.*, 2014), that likelihood-based approaches such as SPML (see Section (2.4.1)), can lead to biased estimates, and their efficiency may not be achieved, with a misspecified underlying disease model.

2.4. Joint modeling of primary and secondary outcomes

2.4.1. Maximum likelihood, semiparametric maximum likelihood (SPML), and related methods

In the context of a binary secondary outcome, Lee *et al.* (1997) proposed modelling a bivariate response, where the original case-control status D and the secondary outcome Y are the two responses

measured on the same individual. The goal is to estimate the coefficients $\boldsymbol{\beta}$ of (categorical) covariates $\mathbf{X}$ in the marginal logistic regression model for $P(Y = 1|\mathbf{X})$ while adjusting for the dependency between the primary and secondary outcomes D and Y. The authors consider two parameterizations for $P(D, Y|\mathbf{X})$. The first is the Palmgren model (Palmgren, 1989), in which both $P(Y = 1|\mathbf{X})$ and $P(D = 1|\mathbf{X})$ are modeled through logistic regression, and the log-OR between Y and D is modeled as a linear function of the covariates $\mathbf{X}$. That is,

$$\text{logit}\{P(Y = 1|\mathbf{X})\} = \boldsymbol{\beta}^\top \mathbf{X}^{(1)}, \quad \text{logit}\{P(D = 1|\mathbf{X})\} = \boldsymbol{\gamma}^\top \mathbf{X}^{(2)}, \quad \log \text{OR}(Y, D|\mathbf{X}) = \boldsymbol{\alpha}^\top \mathbf{X}^{(3)}, \tag{2.10}$$

where $\mathbf{X}^{(j)}$, $j = 1, 2, 3$ are vectors derived from $\mathbf{X}$ by dropping or transforming covariates. The second option again takes $P(Y = 1|\mathbf{X})$ as a logistic regression, but the primary disease model also considers the secondary outcome Y as a covariate, i.e., $P(D = 1|Y, \mathbf{X})$, and is modeled through logistic regression. In this case, the OR between Y and D is independent of $\mathbf{X}$, i.e., $\log \text{OR}(Y, D) = \alpha_0$. In each parameterization, the three specified 'submodels' specify the joint distribution of Y and D given the covariate vector $\mathbf{X}$ as a function of the regression coefficients from each submodel collected in vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\alpha}^\top)^\top$; the parameters in $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are considered as nuisance parameters. They use the technique of Scott and Wild (1995) to estimate the joint distribution, which is parameterized in terms of the marginal distribution of the secondary outcome, and yields the so-called pseudoconditional likelihood (PCL) estimate of $\boldsymbol{\beta}$.

Jiang *et al.* (2006) also consider the situation of a binary secondary outcome, and further examine joint modeling options for Y, D, and $\mathbf{X}$ via semiparametric maximum likelihood (SPML) methods. More broadly, likelihood-based methods for estimating $\boldsymbol{\beta}$ often involve the joint distribution of D, Y and $\mathbf{X}$ at some level, where the joint distribution can be represented as

$$P(Y, D, \mathbf{X}) = P(\mathbf{X})P(Y|\mathbf{X}; \boldsymbol{\beta})P(D|Y, \mathbf{X}). \tag{2.11}$$

The primary interest of the secondary outcome analysis is in parameter $\boldsymbol{\beta}$, and all other terms and parameters are considered nuisances. In the model termed SPML1, both $P(\mathbf{X})$ and $P(D|Y, \mathbf{X})$ are modeled nonparametrically. Jiang *et al.* (2006) state that this modeling is only possible via maximum likelihood if the variables are sufficiently discrete and the data set is large enough so that it is feasible via asymptotic theory to estimate $P(D|Y, \mathbf{X})$ for each combination of Y and $\mathbf{X}$. In the model termed SPML2, $P(D|Y, \mathbf{X})$ is modeled parametrically, while $P(\mathbf{X})$ is still modeled nonparametrically. With the bivariate modeling of D and Y in SPML3, they consider models for $P(D, Y|\mathbf{X})$ for which the marginal distribution of Y is $P(Y|\mathbf{X}, \boldsymbol{\beta})$. Specifically, Jiang *et al.* (2006) consider $P(D|Y, \mathbf{X})$ modeled using logistic regression for SPML2 (similar to the second parameterization in Lee *et al.* (1997)) and the Palmgren model (2.10) for SPML3 in their simulation studies, and compared the SPML methods with the IPW method. They noted, as have others (see Section 2.3), that while the weighted method was easier to implement and enjoys good robustness properties for estimating $\boldsymbol{\beta}$, it is generally less efficient than the SPML2 and SPML3 methods which impose additional modeling assumptions. SPML1, which imposes no additional modeling assumptions, had efficiencies nearly identical to IPW. There is a robustness-efficiency trade-off, however, as both SPML2 and SPML3 can suffer from considerable lack of robustness when the nuisance parts of the likelihood are misspecified. In particular, Jiang *et al.* (2006) note the importance of including at least the apparently significant Y by $\mathbf{X}$ interactions in the $D|Y, \mathbf{X}$-model when using the SPML2 method for analysis, and that the robustness of the SPML3 method is more sensitive to the misspecification of logOR-model than to the disease model.

The joint distribution (2.11) also appears in the retrospective likelihood considered in Lin and

Zeng (2009), among others, given by

$$\prod_{i=1}^{n} P(Y_i, \mathbf{X}_i | D_i) = \prod_{i=1}^{n} \left\{ \frac{P(\mathbf{X}_i)P(Y_i|\mathbf{X}_i;\beta)P(D_i = 1|Y_i, \mathbf{X}_i)}{P(D_i = 1)} \right\}^{D_i} \left\{ \frac{P(\mathbf{X}_i)P(Y_i|\mathbf{X}_i;\beta)P(D_i = 0|Y_i, \mathbf{X}_i)}{P(D_i = 0)} \right\}^{1-D_i}$$
$$= \prod_{i=1}^{n} \frac{P(\mathbf{X}_i)P(Y_i|\mathbf{X}_i;\beta)P(D_i|Y_i, \mathbf{X}_i)}{P(D_i)}. \tag{2.12}$$

In this work, $P(Y_i|\mathbf{X}_i;\beta)$ is modeled with logistic or linear regression for binary or quantitative secondary outcome Y, respectively, the disease model is given by

$$\text{logit}\{P(D = 1|Y, \mathbf{X})\} = \delta_0 + \delta_X^\top \mathbf{X} + \delta_Y Y, \tag{2.13}$$

and $P(D = 1) = 1 - P(D = 0) = \int P(\mathbf{X})P(Y|\mathbf{X};\beta)P(D = 1|Y, \mathbf{X})d\mu(\mathbf{X})d\mu(Y)$ where $\mu(\cdot)$ is a Lebesgue measure for a continuous random variable and a counting measure for a discrete random variable. The density of the covariates, $P(\mathbf{X})$, is regarded as nuisance using a profile likelihood approach, and as such, this represents an SPML2 model in the terminology of Jiang *et al.* (2006). The authors consider three scenarios: (i) known population disease-prevalence rate, (ii) rare disease, and (iii) general unknown disease prevalence rate. Their retrospective likelihood approach performs very well under scenarios (i) and (ii) when the population disease-prevalence rate is accurate, but can perform poorly in scenario (iii). In this situation, Lin and Zeng (2009) suggest performing sensitivity analysis to the population disease-prevalence rate. As noted by Tchetgen Tchetgen (2014), Lin and Zeng (2009) also established that their likelihood approach is well approximated by the "naive" conditional approach (i.e., naive method (ii) in Section 2.2) under the following conditions: (LZ.1) a rare disease assumption about the disease outcome defining case-control status, (LZ.2) no interaction between the secondary outcome and covariates in a regression model for the case-control outcome, and (LZ.3) the secondary outcome is normally distributed. Their approach is implemented in the C program `SPREG` (version 2.0) available at http://dlin.web.unc.edu/software/. The authors show in simulation that SPREG provides accurate control of the type I error and has high efficiency under correct model specification, but others have noted that both measures can suffer with model misspecification (e.g., Jiang *et al.*, 2006; Li *et al.*, 2010; Tapsoba *et al.*, 2014; Song *et al.*, 2016a).

Li *et al.* (2010) further consider the retrospective likelihood (2.12) in the context of binary X, Y and D under the assumption that the disease is rare (scenario (ii) from Lin and Zeng (2009)). With the saturated disease model (2.3) that includes the interaction term ($\delta_{XY} \neq 0$), Li *et al.* (2010) show that the maximum likelihood estimator (MLE) of β_X is the same as the control-only estimator $\hat{\beta}_{XCO}$ given in Section 2.2. Li *et al.* (2010) state that the MLE of β_X from this model corresponds to a SPML1 approach. Assuming $\delta_{XY} = 0$ as in Lin and Zeng (2009) corresponds to possibly misspecified parametric modeling within the SPML2 framework, such that if $\delta_{XY} \neq 0$, the MLE of Lin and Zeng (2009) is not robust and the resulting test does not control the type I error.

Recall from Section 2.2 under the rare disease setting that the meta-analytic estimator $\hat{\beta}_{XW}$ in (2.5) was a more efficient estimator of β_X than $\hat{\beta}_{XCO}$, but was only unbiased when $\delta_{XY} = 0$. To enjoy the efficiency of $\hat{\beta}_{XW}$ while avoiding the bias induced when $\delta_{XY} = 0$, Li *et al.* (2010) propose an adaptively weighted estimator for the rare disease setting that combines the control-only estimator with the meta-analytic-type estimator $\hat{\beta}_{XW}$. Specifically, they propose the estimator

$$\hat{\beta}^R_{XAW} = \frac{\hat{\delta}^2_{XY}}{\hat{\sigma}^2_{CO} + \hat{\delta}^2_{XY}} \hat{\beta}_{XCO} + \frac{\hat{\sigma}^2_{CO}}{\hat{\sigma}^2_{CO} + \hat{\delta}^2_{XY}} \hat{\beta}_{XW}, \tag{2.14}$$

where $\hat{\delta}_{XY} = (\hat{\beta}_{XCA} - \hat{\beta}_{XCO})$ and $\hat{\sigma}^2_{CO}$ was defined in Section 2.2. Clearly, the influence of $\hat{\beta}_{XCO}$ and $\hat{\beta}_{XW}$ depends adaptively on the estimate of the interaction $\hat{\delta}_{XY}$, with more weight placed on $\hat{\beta}_{XW}$ when the estimated interaction effect is small. Li *et al.* (2010) also provide a variance estimate for $\hat{\beta}^R_{XAW}$ from which one can construct a Wald statistic. In simulation, they show that the meta-analytic estimator in (2.5) does not control the type I error and cannot be trusted if it is plausible that $\delta_{XY} \neq 0$. The adaptively weighted estimator in (2.14), however, has lower MSE and greater power than the controls-only method, with only slight inflation in type I error rates.

Li and Gail (2012) proposed a different adaptively weighted estimator that is appropriate for the common disease situation (scenarios (i) and (iii) from Lin and Zeng (2009)). They generalize the approach of Lin and Zeng (2009) and suggest using a weighted sum of two retrospective likelihood-based estimators for β_X that differ in their assumed primary disease models. If disease prevalence is known (scenario (i)), then the MLE of β_X resulting from these models can be computed and is denoted by $\hat{\beta}_{XFM}$. The subscript FM stands for "full model", in reference to primary disease model (2.2) that includes the interaction term $\delta_{XY} \neq 0$. Li and Gail (2012) show that for dichotomous X and assuming (2.2) and (2.4) for the primary and secondary models, respectively, that $\hat{\beta}_{XFM}$ is the same as the IPW estimate from Richardson *et al.*, (2007) (see Section 2.3), which is known to be inefficient. If, however, one instead considers the reduced model (RM) in (2.1) in which $\delta_{XY} = 0$, while still assuming disease prevalence is known, the corresponding MLE, $\hat{\beta}_{XRM}$, is more efficient than $\hat{\beta}_{XFM}$ but can also be misleading if, in fact, $\delta_{XY} \neq 0$. Li and Gail (2012) thus propose the adaptive estimator

$$\hat{\beta}^C_{XAW} = \frac{\hat{\delta}^2_{XYFM}}{\hat{\delta}^2_{XYFM} + \hat{\sigma}^2_{FM}}\hat{\beta}_{XFM} + \frac{\hat{\sigma}^2_{FM}}{\hat{\delta}^2_{XYFM} + \hat{\sigma}^2_{FM}}\hat{\beta}_{XRM}, \tag{2.15}$$

where $\hat{\delta}_{XYFM}$ is the MLE of the interaction in the full model and $\hat{\sigma}^2_{FM}$ is the estimated variance of $\hat{\beta}_{XFM}$. Thus, $\hat{\beta}^C_{XAW}$ preferentially weights $\hat{\beta}_{XFM}$ over $\hat{\beta}_{XRM}$ when there is evidence that $\delta_{XY} \neq 0$. Li and Gail (2012) show in simulation that their proposed estimator $\hat{\beta}^C_{XAW}$ tends to outperform the full model and IPW estimators in terms of power, regardless of whether an interaction δ_{XY} truly exists. They also examine the performance of the reduced model estimator for β_X, both with and without external disease prevalence information. With knowledge of the disease prevalence (scenario (i)), the estimator $\hat{\beta}_{XRM}$ is most efficient when $\delta_{XY} = 0$, but does not control the type I error when $\delta_{XY} \neq 0$ and can be quite inaccurate. Without this external information (scenario (iii)), the reduced model estimator does not control the type I error regardless of the value of δ_{XY}, and is not recommended.

Ghosh *et al.* (2013) also take a retrospective likelihood approach, but do so by specifying models for the joint distribution of Y and D. Writing the retrospective likelihood in terms of this joint distribution, the contribution of the i^{th} observation is given by $P(Y_i, \mathbf{X}_i|D_i) = P(\mathbf{X}_i)P(D_i, Y_i|\mathbf{X}_i)/P(D_i)$, $i = 1, \ldots, n$. As an SPML3 approach, $P(D_i, Y_i|\mathbf{X}_i)$ is modeled parametrically while $P(\mathbf{X}_i)$ is modeled nonparametrically. Regardless of whether the secondary outcome Y is binary or continuous, the parameterization of the joint distribution of Y and D is chosen to respect the marginal logistic distribution for the disease trait D. For binary Y, this can be accomplished with the Palmgren model (2.10) as in Lee *et al.* (1997) and Jiang *et al.* (2006) which results in logistic marginals for both D and Y. For continuous Y, the natural choice for the marginal distribution of Y is normal. There is no standard bivariate distribution that yields logistic and normal marginals for D and Y, respectively, so Ghosh *et al.* (2013) propose a two-stage latent variable approach that starts with a bivariate normal distribution: let $\mu_1 = \gamma_0 + \gamma_1 X + \boldsymbol{\gamma}_2^\top \mathbf{Z}$ and $\mu_2 = \beta_0 + \beta_X X + \boldsymbol{\beta}_Z^\top \mathbf{Z}$, and define V such that

$$\begin{pmatrix} V \\ Y \end{pmatrix} \Bigg| X, \mathbf{Z} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma_2^2 \\ \rho\sigma_2^2 & \sigma_2^2 \end{bmatrix}\right),$$

where $\log[(1+\rho)/(1-\rho)] = \alpha_0 + \alpha_1 X$. To obtain the desired logistic marginal, $\text{logit}\{P(D=1|X,\mathbf{Z})\} = \gamma_0 + \gamma_1 X + \boldsymbol{\gamma}_2^\top \mathbf{Z}$, further define $U = \mu_1 + \log[\Phi(V-\mu_1)/(1-\Phi(V-\mu_1))]$, where Φ is the standard normal cumulative distribution function, and set $D = 1$ if $U \geq 0$ and $D = 0$ if $U < 0$. As usual, the interest is in inference for β_X and Ghosh *et al.* (2013) propose two ways of doing this: an approximate profile likelihood approach and pseudo-likelihood approach. The performance of the two approaches in simulation under both binary and continuous secondary outcome experiments were nearly identical in terms of bias, MSE, coverage probabilities, type I error, and power. The simulations also consider several naive analyses, the adaptively weighted method of Li *et al.* (2010), and IPW. Only the proposed likelihood-based methods and IPW approaches maintained the correct type I error rate throughout all simulation settings considered, but the likelihood-based methods can be much more efficient and powerful than IPW. Consistent with the results in Jiang *et al.* (2006), the SPML3 method performs well even when the data are simulated according to the model of Lin and Zeng (2009). Ghosh *et al.* (2013) note that the likelihood-based methods do not, in principle, require specification of the disease prevalence, and can be used to analyze secondary outcomes for data from case-control studies of both rare and common primary diseases. When disease prevalence is assumed unknown, however, the estimation of intercept parameter in the marginal disease model (γ_0) can lead to numerically unstable solutions (see also Lin and Zeng, 2009; Li and Gail, 2012, with scenario (iii)).

For a single normally distributed secondary outcome Y, the model of Ghosh *et al.* (2013) is actually special case of a Gaussian copula model. He *et al.* (2012) propose a Gaussian copula model framework where again, the joint distribution is modeled in terms of the marginals for the primary and secondary outcomes, and the multivariate normal distribution is used to model the correlation between them. Their method can also be applied to multiple correlated secondary outcomes, and can handle a variety of secondary outcome types provided their distributions come from an exponential family. Unlike in Ghosh *et al.* (2013), however, there is no model specified to capture the relationship between X and the correlation of Y and D. Assuming a single exposure variable X corresponding to a SNP under Hardy-Weinberg equilibrium and an additive genetic model (i.e., $X \in \{0,1,2\}$), no additional covariates, and known disease prevalence, He *et al.* (2012) propose maximizing the resulting retrospective likelihood numerically using a Gauss-Newton type algorithm to estimate the model parameters. They compare their approach in simulation to the Lin and Zeng (2009) approach and the IPW method for the scenario of a single continuous secondary outcome. Robustness was evaluated by simulating data from the model of Lin and Zeng (2009) (i.e., models (2.1) and (2.6)). Both the copula-based and IPW approaches controlled the type I error rates and were robust to this model misspecification, while the Lin and Zeng approach had inflated type I error rates when the data was simulated under the copula model, particularly when the primary and secondary outcomes were highly correlated. For power, when the exposure variable X is not associated with disease, the copula-based approach and Lin and Zeng's approach have similar power, and both outperform the IPW method. When the X is associated with D, the power of Lin and Zeng's method depends on the correlation between the primary and secondary outcomes, while both the copula-based and IPW approaches have roughly constant power across different correlation levels. In particular, the copula-based approach is more powerful than Lin and Zeng's approach when the primary and secondary outcomes are positively correlated.

Motivated by the recent emergence of next-generation sequencing and the associated challenges with studies of rare genetic variants, Kang *et al.* (2017) proposed an approach to jointly model the primary binary outcome and the (continuous or binary) secondary outcome by using the so-called set-value (SV) model. As they describe, "the SV model is to model the relationship between independent variables and a set-valued dependent variable that can be generated by a quantization process of the corresponding continuous latent or unknown variable." In the context of secondary outcome analysis,

the SV model is used to model the dichotomizing process of the continuous variable for the primary disease outcome, and a similar dichotomization process is used if the secondary outcome is also binary. Let D_{lv} be the continuous underlying latent variable corresponding to D, and similarly let Y_{lv} be the continuous underlying latent variable corresponding to the binary secondary outcome Y. With both binary D and Y, the proposed model is given by

$$Y_{lv} = \beta_0 + \beta_X X + \epsilon, \quad Y = I_{[Y_{lv}>0]}, \tag{2.16}$$

$$D_{lv} = \delta_0 + \delta_X X + \delta_Y Y + \epsilon_{lv}, \quad D = I_{[D_{lv}>0]}, \tag{2.17}$$

where ϵ (ϵ_{lv}) are independently normally distributed with mean 0 and variance σ^2 (σ^2_{lv}), and I is an indicator function. Thus the conditional probabilities of the secondary and primary outcomes are $P(Y = 1|X) = F_\sigma(\beta_0 + \beta_X X)$ and $P(D = 1|X, Y) = F_{\sigma_{lv}}(\delta_0 + \delta_X X + \delta_Y Y)$, respectively, where $F_a(\cdot)$ is the cumulative distribution function of a normal distribution with mean 0 and variance a^2. If instead Y is continuous, (2.16) is replaced by (2.6), and where again ϵ (ϵ_{lv}) are independently normally distributed with mean 0 and variance σ^2 (σ^2_{lv}). To estimate these SV model parameters, they propose maximizing the retrospective likelihood given in (2.12). Here, the authors assume a known disease prevalence, so that maximizing (2.12) is equivalent to maximizing its numerator with $P(Y_i|X_i;\beta)$ and $P(D_i|Y_i, X_i)$ specified from the SV model. For $P(X_i)$ with $X_i \in \{0, 1, 2\}$ corresponding to the number of minor alleles of a SNP, they assume Hardy-Weinberg equilibrium, and parameterize this distribution with its minor allele frequency (MAF) p. With known disease prevalence, p can also be expressed in terms of the SV model parameters. Thus, the retrospective likelihood can be expressed as a function of only SV model parameters, and can be numerically optimized to find the MLE. They also derived a closed-form Fisher information matrix based on estimated parameters, so that they can construct a Wald statistic for inference. Their approach is implemented within the R package `SV2bc` (version v0.1) available at https://www.stjuderesearch.org/site/depts/biostats/sv2bc. In simulation, Kang *et al.* (2017) compared their approach to those of Lin and Zeng (2009) and Ghosh *et al.* (2013) in a variety of settings to assess the effect of different MAFs (both rare and common variants), different disease prevalences, different link functions, different correlations between primary and secondary outcomes, and different types of associations between the genetic variants and the primary and secondary outcomes. Interestingly, there was no obvious effect of the disease prevalence on the estimated type I error or power of the SV method, but it did have an impact on the performance of the other methods particularly for rare variants. Moreover, its use of (latent) continuous outcomes contributed to the method's efficiency and robustness when estimating the model parameters.

Also in the context of studying rare genetic variants, Liu and Leal (2012) propose to jointly model the primary and secondary outcomes conditional on the study subjects being ascertained, where the sampling mechanism is incorporated by means of a prospective likelihood approach. The secondary outcome is assumed to be continuous, but they consider not only case-control sampling, but also extreme trait and multiple trait studies. For case-control studies, the primary and secondary outcomes are assumed to follow a multivariate GLM with $F_D(\boldsymbol{\mu}_D) = \gamma_0 + \gamma_X X + \gamma'_Z \mathbf{Z}$ and $F_Y(\boldsymbol{\mu}_Y) = \beta_0 + \beta_X X + \beta_D D + \beta'_Z \mathbf{Z}$ modeled jointly, where $F_D(\boldsymbol{\mu}_D)$ and $F_Y(\boldsymbol{\mu}_Y)$ are link functions, and $\boldsymbol{\mu}_D$ and $\boldsymbol{\mu}_Y$ are the model parameters related to the primary and secondary outcomes. Assuming no covariates for simplicity, the conditional likelihood is given by $\prod_{i=1}^n P(D_i, Y_i|S_i = 1, X_i)$, where S_i is again the sampling indicator. With probit and normal link functions specified for $F_D(\cdot)$ and $F_Y(\cdot)$, respectively, the model can be simplified. Score statistics are proposed for detecting associations with X.

2.4.2. Relaxing parametric assumptions

All of the methods discussed in Section 2.4.1 have assumed a parametric model for secondary outcome

Y given exposure X. As has been noted above, the semiparametric efficient estimates can suffer from lack of robustness when the underlying model is misspecified. To address this issue, Wei *et al.* (2013) relax the parametric assumption under the continuous secondary outcome framework by instead requiring that the secondary outcome regression is "homoscedastic", i.e., the secondary regression model is of the form $Y = \beta_0 + \mu(\mathbf{X}, \boldsymbol{\beta}) + \epsilon$, where β_0 is an intercept, $\mu(\cdot)$ is a known function, and where ϵ has mean 0 and is independent of covariates $\mathbf{X}$, but its density is otherwise unspecified. The disease model retains a parametric form with a logistic regression model, such that $\text{logit}\{P(D = 1|Y, X)\} = \delta_0 + m(Y, X, \boldsymbol{\delta})$, where $m(\cdot)$ is a known function with unknown parameter vector $\boldsymbol{\delta}$; in their simulations, they use model (2.1). The authors use an estimating equation approach for parameter estimation, and a bootstrap procedure to estimate the variance of the estimated parameters; please consult the original manuscript for further details. They show theoretically that the regression parameters can be estimated consistently even if the assumed model for Y given $\mathbf{X}$ is incorrect under the assumption that the disease prevalence is known or well estimated; this assumption can be dropped when the disease is rare. Their simulations were conducted under the rare disease assumption and using either Gaussian or Gamma random errors. Their proposed estimator had small bias and nearly nominal coverage probability in all scenarios considered. As expected, the approach that imposed a parametric model for the secondary outcome regression can suffer from bias and lower coverage probability when the model is misspecified (under Gamma random errors). In Gazioglu *et al.* (2013), the authors modify the work of Wei *et al.* (2013) so that it may apply to penalized spline regression, where $Y = f(X) + \epsilon$ with $f(\cdot)$ an unknown smooth function and ϵ having mean zero and being independent of X, but its distribution is otherwise unspecified. Their simulation results indicate that their proposed method improves efficiency significantly upon spline regression using the controls only.

Tchetgen Tchetgen (2014) describes the homoscedastic assumption of Wei *et al.* (2013) as assuming that any association between the vector of covariates and the secondary outcome is completely captured by a location shift model. He further notes that their inferential framework relies critically on this assumption, and may exhibit bias if the assumption is not satisfied. Tchetgen Tchetgen (2014), in turn, proposes a generalization of the "naive" conditional approach (i.e., naive approach (ii) in Section 2.2) to allow for possible violation of any or all of assumptions (LZ.1)–(LZ.3) (see Section 2.4.1), without assuming the location shift model of Wei *et al.* (2013). His approach relies on a nonparametric reparameterization of the conditional model for the secondary outcome given the primary outcome and covariates, in terms of (a) the population regression of interest for the secondary outcome given covariates and (b) the population regression of the primary outcome on covariates. Models (a) and (b) can take any functional form, but the approach is developed for the identity, log, and logit link functions in model (a). Consider here the identity link and let $\boldsymbol{\mu}(\mathbf{X}) = E(Y|\mathbf{X})$ denote the population mean model of interest. If $\tilde{\boldsymbol{\mu}}(\mathbf{X}, D) = E(Y|\mathbf{X}, D)$ denotes the conditional mean, then Tchetgen Tchetgen (2014) shows that the unconditional and conditional means are related as $\tilde{\boldsymbol{\mu}}(\mathbf{X}, D) = \boldsymbol{\mu}(\mathbf{X}) + \gamma(\mathbf{X})[D - P(D = 1|\mathbf{X})]$, where $\gamma(\mathbf{X}) = \tilde{\boldsymbol{\mu}}(\mathbf{X}, 1) - \tilde{\boldsymbol{\mu}}(\mathbf{X}, 0)$ is the so-called *selection bias function*. Thus, $\tilde{\boldsymbol{\mu}}(\mathbf{X}, D)$ can be parameterized in terms of the population regression function of interest $\mu(\mathbf{X})$, and also $\gamma(\mathbf{X})$ and $P(D = 1|\mathbf{X})$. This re-parameterization is described as "nonparametric" because it is not restricted to a particular choice of models for $(P(D = 1|\mathbf{X}), \gamma(\mathbf{X}), \mu(\mathbf{X}))$ so that parametric, semiparametric, and nonparametric models could be used for each of these in principle. Also note that under the assumptions (ETT.1) $\gamma(\mathbf{X}) = \gamma$ does not depend on $\mathbf{X}$ and (ETT.2) the disease is rare in the population, and the specification of $\boldsymbol{\mu}(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$, then $\tilde{\boldsymbol{\mu}}(\mathbf{X}, D) \approx \beta_0 + \boldsymbol{\beta}'\mathbf{X} + D\gamma$. Thus, under assumptions (ETT.1) and (ETT.2), the naive conditional model that includes a main effect for D in the regression to adjust for case-control sampling (i.e., naive method (ii) in Section 2.2), is approx-

imately correct. Lin and Zeng (2009) noted a similar result under assumptions (LZ.1)–(LZ.3), where (LZ.1)–(LZ.3) imply (ETT.1) and (ETT.2), but the converse is generally not true. For other links, Tchetgen Tchetgen (2014) provides similar re-parameterizations of the conditional means $E(Y|\mathbf{X}, D)$ in terms of the unconditional means $E(Y|\mathbf{X})$. A set of estimating equations is described for inference, and a potentially more efficient approach is also provided. In this second approach, the estimator achieves the semiparametric efficiency bound in the absence of model misspecification, and remains consistent even if the error distribution for the outcome is incorrectly specified (provided the form of the mean is correctly specified). When the disease is not rare, the approach requires that sampling fractions are known for cases and controls or that the population disease prevalence is known. Simulations were conducted under the common disease setting using models with the identity link and normally distributed random error. The results confirm that IPW, the proposed estimating equation approach, and proposed locally efficient approach all have small bias and produce 95% confidence intervals with appropriate coverage. The locally efficient estimation can also outperform IPW and the estimating approach in terms of efficiency.

Ma and Carroll (2016) also relax the assumptions required for the analysis of a continuous secondary outcome. The secondary regression model here takes the form

$$Y = m(\mathbf{X}, \boldsymbol{\beta}) + \epsilon, \tag{2.18}$$

where function $m(\cdot)$ known up to a parameter $\boldsymbol{\beta}$, but the only assumption is that $E(\epsilon|\mathbf{X}) = 0$. While Wei *et al.* (2013) used a similar model and also did not assume a distributional form for the error, they still assumed a homoscedastic distribution for ϵ independent of $\mathbf{X}$ and made a rare disease approximation. To avoid making such assumptions, the authors use the notion of a "superpopulation." As they describe, "the main idea behind a superpopulation is to enable us to view the case-control sample as a sample of independent and identically distributed observations from the superpopulation. Conceptually, a superpopulation is simply a proportional expansion of the case-control sample to ∞...The ability to view the case-control sample as a random sample permits us to use classical semiparametric approaches (Bickel *et al.*, 1993; Tsiatis, 2006), regardless of whether the disease rate in the real population is rare or not, or is known or not." With the superpopulation having the same case-to-control ratio as the case-control sample, it retains the same joint distribution of $\mathbf{X}$ and Y given D as in the true population. The density of $(D, Y, \mathbf{X})$ in the superpopulation is given as

$$f_{\mathbf{X},Y,D}(\mathbf{x}, y, d) = f_D(d) f_{\mathbf{X},Y|D}(\mathbf{x}, y, d) = \frac{n_d}{n} f^{\text{true}}_{\mathbf{X},Y|D}(\mathbf{x}, y, d), \tag{2.19}$$

where $f^{\text{true}}_{\mathbf{X},Y|D}(\mathbf{x}, y, d)$ is defined similarly as in Lin and Zeng (2009) with the primary model given by (2.13), but the secondary model is no longer fully parametric and instead given through (2.18) such that $E(\epsilon|\mathbf{X}) = 0$. The goal is to estimate $(\boldsymbol{\delta}, \boldsymbol{\beta})$, while the marginal distributions of $\mathbf{X}$ and the conditional distribution of Y given $\mathbf{X}$ are considered nuisances and are not estimated. Ma and Carroll (2016) also establish identifiability conditions for their model, which are more general than those of Lin and Zeng (2009), in that only a mean function is assumed for the secondary model.

To estimate $(\boldsymbol{\delta}, \boldsymbol{\beta})$, Ma and Carroll (2016) first derive a semiparametric efficient estimating equation in the superpopulation when the densities $f^{\text{true}}_{Y|X}$ and f^{true}_{X} are known. They then remove these assumptions, and modify the estimating equation accordingly so that it has mean 0 asymptotically. Thus, even if the model posited for the secondary regression errors is incorrect, the model will still provide robust and consistent estimation results. The authors further show under regularity that their proposed estimator is asymptotically normal. We refer the interested reader to their manuscript for the details of their estimation procedure. Their method, referred to as "Semi" in the manuscript, was

compared under a variety of settings to the methods of Lin and Zeng (2009), Wei *et al.* (2013), and a control-only analysis based on ordinary least squares with a sandwich estimator for its variance. As expected, under settings in which the simulated data are generated to agree with the assumptions of a particular method (e.g., rare disease, normal errors and homoscedasticity for Lin and Zeng (2009); rare disease and homoscedasticity for Wei *et al.* (2013)), the proposed method compares favorably with the competing methods in terms of bias, standard error, coverage probabilities, and MSE efficiency as compared to control-only analysis. When the assumptions are violated, however, the performance of the Lin and Zeng (2009) and Wei *et al.* (2013) methods may decline. In contrast, "Semi", which makes no assumptions about known/rare disease, normal errors or homoscedasticity, performs well with coverage probabilities that achieve the nominal rates. Fortran code and an executable file for implementing this method, as well as Matlab and R code to generate initial values required by the method, are available at https://www.stat.tamu.edu/~carroll/matlab_programs/software.php.

Very recently, Liang *et al.* (2018) also used the notion of superpopulation for semiparametrically efficient estimation in secondary quantile regression. In contrast to the secondary quantile regression method of Wei *et al.* (2016), Liang *et al.* (2018) allow the disease rate to be unknown in the true population and assume a quantile regression model only at one specific quantile level (not at all quantile levels). Despite the weaker assumptions in Liang *et al.* (2018), their proposed semiparametric estimator still performed better than the estimator of Wei *et al.* (2016) in all simulation settings considered. As in Ma and Carroll (2016), the idea is to view the case-control sample as a random sample taken from the superpopulation. Please consult the original paper for more details.

2.4.3. Proportional odds models

Lutz *et al.* (2014) note that the method of Lin and Zeng (2009) can be computationally intensive for continuous secondary outcome Y under the null hypothesis $H_0 : \beta_X = 0$ when the likelihood surface that needs to be optimized is relatively flat. This can be particularly problematic in the GWAS context, where most of the SNPs are expected to have no association with the secondary outcome. If the main objective of the secondary outcome analysis is hypothesis testing rather than parameter estimation, they propose an alternative likelihood decomposition, which does not require maximizing a relatively flat likelihood surface, but does depend crucially on the exposure variable X representing a genomic value for a particular SNP. More specifically, they use $\prod_{i=1}^{n} P(Y_i, X_i|D_i) = \prod_{i=1}^{n} P(X_i|Y_i, D_i)P(Y_i|D_i)$. The likelihood ratio test statistic for assessing whether X is independent of Y is then $-2\log[\prod_i P(X_i|D_i)/\prod_i P(X_i|Y_i, D_i)]$, which is distributed χ^2_1 under the null hypothesis. For an additive genetic model for X (i.e., $X \in \{0, 1, 2\}$), the authors suggest using a cumulative logistic regression model with proportional odds for $P(X|D)$ and $P(X|Y, D)$. Thus, the authors recommend to: 1) test all SNPs with their proposed proportional odds logistic regression approach, and then for the significant SNPs, 2) apply the Lin and Zeng (2009) method to obtain parameter estimates and confidence intervals. While the power of their proposed method is comparable to that of Lin and Zeng (2009) in simulation, the savings in computation time with their proposed approach can be substantial.

Recently, Ray and Basu (2017) also consider regressing X on Y using a proportional odds model (POM), but rather than including D as a covariate, they include an estimated propensity score (i.e., the conditional probability of case status) as a covariate. As discussed in Ray and Basu (2017) and references therein, propensity scores are traditionally used to estimate an average treatment effect by comparing treated and untreated groups in a non-randomized study, with four main approaches: 1) propensity score matching, which involves creating matched sets of treated and untreated subjects that share a similar value of the propensity score; 2) propensity score stratification, which involves

using estimated scores to rank subjects, who are then grouped into subsets based on thresholds defined *a priori*; 3) IPW, where each subject's weight is equal to the inverse of estimated propensity score; and 4) adjusting for propensity score as a covariate in the regression model. In a retrospective case-control study, Ray and Basu (2017) note that one can potentially use any of these approaches (see, e.g., Schifano *et al.*, 2015, for matching), that the IPW method using propensity score based weights showed inflated type I error for this approach in their preliminary studies under some scenarios, and that the other methods are not as amenable to the analysis of multiple secondary outcomes. In contrast, both the Lutz *et al.* (2014) and Ray and Basu (2017) POM methods can easily accommodate multiple secondary outcomes since Y is included as a predictor, but only the proposed method of Ray and Basu (2017) can control the type I error rate in the simulation settings considered. R code is available for both POM methods on github: https://github.com/SharonLutz/SecondaryPhenotype and https://github.com/RayDebashree/POM-PS.

2.4.4. Bias correction methods

In the context of a binary secondary outcome Y, Wang and Shete (2011a) consider the primary disease model (2.1) with no exposure by secondary-outcome interaction and the secondary outcome model given in (2.4). The interest was specifically in estimating $\text{OR}_{XY} = \exp(\beta_X)$. Conditional on n_1 and n_0, the authors indicate that the OR can be written as a function of the expected number of individuals E_{ki}, where $Y = k \in \{0, 1\}$ and $X = i \in \{0, 1\}$. That is,

$$\text{OR}_{XY} = \frac{E_{11}E_{00}}{E_{01}E_{10}}, \quad \text{where } E_{ki} = \sum_j \frac{n_j}{n} P(Y = k, X = i | D = j),\ i, j, k = 0, 1. \tag{2.20}$$

As in Lin and Zeng (2009), the retrospective probability in E_{ki} can be written as $P(X = i)P(Y = k|X = i)P(D = j|Y = k, X = i)/P(D = j)$, for $i, j, k = 0, 1$. The two conditional probabilities are specified by models (2.1) and (2.4), and $P(D = 1) = 1 - P(D = 0)$ is the disease prevalence in the general population. Here, X represents a binary SNP count under either a dominant or recessive genetic model, so the probabilities $P(X = i)$ are related to genotypic frequencies of the SNP. Consequently, OR_{XY} can be expressed as a function of the unknown parameters in models (2.1) and (2.4). To avoid bias, however, the authors propose incorporating known information about the true prevalences of *both* the primary and secondary outcomes when estimating the intercepts δ_0 and β_0. The logistic regression model (2.1) fit to the case-control sample will provide estimates of δ_X and δ_Y, and the genotypic frequencies needed for estimating $P(X = i)$ can be estimated from the data. Noting that

$$P(D = 1) = \sum_{i,k} P(X = i)P(Y = k|X = i)P(D = 1|Y = k, X = i), \tag{2.21}$$

$$P(Y = 1) = \sum_i P(X = i)P(Y = 1|X = i), \tag{2.22}$$

and that the estimated prevalences $\widehat{P(D = 1)}$ and $\widehat{P(Y = 1)}$ in the general population can be obtained from the literature, the authors propose a method of moments approach for solving (2.20), (2.21), and (2.22) as a system of nonlinear equations with three unknown parameters $\delta_0, \beta_0, \beta_X$ to produce bias-corrected OR estimates. They also extend the approach to handle categorical covariate $X = i \in \{0, 1, 2\}$ representing an additive genetic model for a SNP, and propose a bias correction approach for the frequency-matched case-control study with respect to the secondary trait of interest. Confidence intervals for the OR are obtained via bootstrap. Wang and Shete (2011a) show in simulation that the

corrected OR obtained from their proposed approach was a more accurate estimator of the OR as compared to naive-type analyses, as well as IPW. While their approach does not require specification regarding whether the disease is rare or common, it does require knowledge of the prevalences of *both* the primary and secondary outcomes. Sensitivity analyses show, however, that the misspecification of the primary and secondary outcome prevalences do not have a large impact on the estimate of the corrected OR. In Wang and Shete (2011b), the authors separately investigate the type I error and power of their proposed bias correction approach, and compare their results to those obtained from "naive" logistic regression analyses, focusing on a case-control study design frequency matched on the secondary phenotype in their simulations. Indeed, the bias correction approach is more powerful for detecting the association and has better-controlled type I error rates. Wang and Shete (2012) further demonstrate that their method is robust even when $\delta_{XY} \neq 0$ in the disease model (2.2). Their approach, called OR_tilde, is implemented in R, and version 1.3 is currently available at https://sites.google.com/site/jianwangswebsite/biascorrection.

Chen *et al.* (2013) also propose bias correction formulas under both marginal and conditional analysis of the secondary outcome Y, but do so under a slightly more general modeling framework than Lin and Zeng (2009) and Wang and Shete (2011a) to allow for gene×environment ($X \times E$) interactions and more general sampling schemes. For binary secondary outcome Y, they operate under the following primary and secondary outcome models:

$$\begin{aligned} \text{logit}\{P(D = 1|Y, X, E)\} &= \delta_0 + \delta_Y Y + \delta_X X + \gamma_E E + \gamma_{XE} XE, \\ \text{logit}\{P(Y = 1|X, E)\} &= \beta_0 + \beta_X X + \beta_E E + \beta_{XE} XE. \end{aligned}$$

Let S again be the sampling indicator, where $S = 1$ for a subject included in the case-control sample. Chen *et al.* (2013) model the sampling probability as $P(S = 1|D, Y, X, E) = \pi_1(D)\pi_2(Y)\pi_3(X, E)$, which can accommodate pre-selection of subjects based on (X, E), and reduces to case-control ascertainment when $\pi_2(Y) = \pi_3(X, E) = 1$. Similar to Liu and Leal (2012), Chen *et al.* (2013) model the primary and secondary outcomes conditional on the study subjects being ascertained. Assuming π_k, $k = 1, 2, 3$, are unknown, the distribution of (D, Y) for a sample ascertained in this manner is given by

$$P(D, Y|X, E, S = 1) = \frac{\pi_1(D)P(D|Y, X, E)\pi_2(Y)P(Y|X, E)}{\sum_D \sum_Y \pi_1(D)P(D|Y, X, E)\pi_2(Y)P(Y|X, E)}. \tag{2.23}$$

Under this sampling design, maximizing the joint prospective likelihood in (2.23) yields an association parameter estimator β_X identical to that from maximizing the retrospective likelihood on the basis of $P(Y, X, E|D)$. Based on this joint likelihood, they derive expressions for $P(Y|D, X, E, S = 1)$ and $P(Y|X, E, S = 1)$ for conditional and marginal analysis, respectively, and also their corresponding bias-correction formulas for the parameter estimates of $(\beta_X, \beta_E, \beta_{XE})$. Correction for both types of analysis can be accomplished by using computationally simple formulas involving estimates from a logistic regression analysis of D given Y, X, and E and the disease prevalence. If the baseline disease prevalence is unknown (scenario (iii) in Lin and Zeng (2009)), the authors suggest computing the corrected estimate for $(\beta_X, \beta_E, \beta_{XE})$ using a range of values, and obtaining sensitivity plots over this range. Variance estimates of the corrected estimators are provided so that testing of $H_0 : \beta_X = \beta_{XE} = 0$ is possible. Chen *et al.* (2013) also provide correction formulas when either or both of the primary and secondary outcomes are quantitative, and further provide correction formulas that allow for interactions between the secondary outcome Y and the exposure X or the environmental factors E. Simulation results revealed good performance in terms of (lack of) bias and appropriate empirical type I error rates. Please see their manuscript for further details.

3. Secondary analysis in a lung cancer case-control study

The Environment And Genetics in Lung cancer Etiology (EAGLE) study is a population-based case-control study from the Lombardy region of Italy including over 2000 lung cancer cases and over 2000 disease-free controls (Landi *et al.*, 2008). A GWAS was conducted to investigate the genetic determinants of lung cancer risk, with genotypic and select phenotypic data available at dbGaP (Study Accession: phs000093.v2.p2). In this section, we illustrate the various types of secondary outcome analyses with nicotine dependence (smoking behavior) serving as the secondary outcome Y, lung cancer status serving as the primary outcome D, and SNP rs8034191 on chromosome 15q25.1 serving as the exposure variable X. Genetic variants in this 15q25.1 region have previously been reported to increase the risks of lung cancer, nicotine dependence, and associated smoking behavior (e.g., VanderWeele *et al.*, 2012, and references therein). Consequently, we expect to see the effects of sampling bias in the naive analyses with this example.

For reasons described in the paragraphs to follow, this dataset will not be used for examining the *true population* association between SNP rs8034191 and nicotine dependence. This dataset, however, will be used to illustrate the similarities and differences between the various methods when both X and Y are associated with D. First, although the control participants were matched on location, gender, and age, (and thus not selected based solely on disease status), the participants were not matched on any smoking-related variables. Frequency-matched case-control designs with respect to the secondary outcome that do not take this matching into account may additionally contribute to bias in the estimates for the effect of the exposure variable on the secondary outcome (e.g., Wang and Shete, 2011a, 2011b, 2012). Matching on other variables has received limited attention, with some discussions found in Reilly *et al.* (2005), Wei *et al.* (2013), Chen *et al.* (2013), and Tchetgen Tchetgen (2014). We proceed for illustration by adjusting all secondary analyses (when possible) for age (ordinal, but modeled as continuous) and gender, as in Hancock *et al.* (2015) and tseng *et al.* (2014). Note also that all participants self-identified as White, and that no location information was included in the dbGaP data. In addition to SNP, when possible the same covariates were also used in any methods requiring a primary disease model, selection model (i.e., in RESCO from Sofer *et al.*, 2017a), and Palmgren model (i.e., in SPML3 from Jiang *et al.*, 2006). Not all implementations allow for covariate adjustment, however, so to facilitate comparisons with these methods, we also present results from analyses that do not adjust for age and gender in any (secondary, primary, selection, Palmgren) model.

Second, nicotine dependence was assessed using the Fagerstrom Test for Nicotine Dependence (FTND) (Heatherton *et al.*, 1991) among participants who reported smoking > 100 cigarettes in their lifetime. For former smokers who quit more than 6 months previously, dependence during the time in which they smoked was reported (Tseng *et al.*, 2014). The test has a score range of 0–10, with higher scores indicating greater dependence. As nicotine dependence was assessed only among current and former smokers using the FTND scale, for illustration purposes we set FTND to 0 for nonsmokers in our secondary analyses to minimize further potential complications of selection bias. The FTND score defined in this manner, referred to as FT henceforth, serves more as a surrogate for 'smoking behavior', rather than as a true measure of nicotine dependence (e.g., Etter *et al.*, 1999). We thus re-emphasize that the results presented below, even those based on methods discussed in Sections 2.3 and 2.4, likely do not reflect the true population association between SNP rs8034191 and nicotine dependence. Indeed, the purpose of this section is not to confirm or invalidate results from previous studies, but rather to show the similarities and differences in results across the various methods using publicly available data. We examine this FT score as both a continuous and binary secondary outcome,

Table 2: Comparison of estimates of β_X, their standard errors (SE), and corresponding p-values, using continuous secondary outcome FT without (left) and with adjustment (right) for age and gender

	Method	No covariate adjustment			Covariate adjustment		
		Estimate	SE	P-value	Estimate	SE	P-value
Naive Methods	No disease status adjustment	0.3196	0.06486	8.722e−07	0.2983	0.06272	2.042e−06
	Adjusted for disease status	0.1349	0.05879	2.178e−02	0.1123	0.05640	4.657e−02
	Control-only	0.0505	0.07971	5.265e−01	0.0372	0.07743	6.306e−01
	Case-only	0.2241	0.08678	9.903e−03	0.1790	0.08205	2.929e−02
Weighting Methods	IPW*†	0.0519	0.08155	5.245e−01	0.0386	0.07943	6.268e−01
	RECSO* (Sofer *et al.*, 2017a)	0.0519	0.08155	5.244e−01	0.0386	0.07943	6.268e−01
	IPW$_R^{\ddagger}$ (Xing *et al.*, 2016)	0.1251	0.06033	3.808e−02			
	WEE*$^{\#}$ (Song *et al.*, 2016a)	0.0519	0.08111	5.224e−01	0.0386	0.07991	6.289e−01
Joint Modeling Methods	SPREG‡/SPML2 (Lin and Zeng, 2009)	0.1349	0.05879	2.173e−02	0.1123	0.05640	4.650e−02
	SPML3*$^{\spadesuit}$ (Gghosh *et al.*, 2013)	0.0622	0.08197	4.481e−01	0.0447	0.07856	5.694e−01
	SV* (Kang *et al.*, 2017)	0.0599	0.02289	8.836e−03			
	Semi$^{\clubsuit\ddagger}$ (Ma and Carroll, 2016)	0.1257	0.09462	1.839e−01			

* Using disease prevalence $\pi = 0.0015$. †Weights defined as in Wang and Shete (2011) for Extended IPW. ‡Disease model includes SNP, FT (no interaction). $^{\#}$Calculated based on 10000 bootstrap samples. $^{\spadesuit}$Model for correlation includes only SNP (no adjustment for age and gender). $^{\clubsuit}$Based on modified initial values.

where for the latter, FT is set to 1 for scores greater than 3 and set to 0 otherwise (Hancock *et al.*, 2015).

The analyses below include a total of $n = 3791$ subjects, with $n_0 = 1973$ controls and $n_1 = 1818$ cases, for which there is complete data for FT, age, gender, case-control status, and SNP. We use an additive genetic model for the SNP, with $X \in \{0, 1, 2\}$ representing the number of (minor) C alleles. The observed MAF for this SNP is 0.4296. Prior to running the secondary outcome analyses, we examined the relationship between the primary, secondary, and exposure variables. Based on the logistic regression model adjusted for age and gender, the interaction between FT and SNP was not statistically significant at the 0.05 level ($p = 0.1542$ and $p = 0.3432$, respectively for continuous and binary FT). The main effects of FT and SNP were both highly significantly associated with disease status ($p <$ 1e−06), further suggesting that the naive analyses will be affected by the sampling bias.

Table 2 provides estimates for β_X, their estimated standard errors (SE), and p-values for the association of SNP with continuous FT, with (right) or without (left) covariate adjustment, for any continuous secondary outcome method that allows for non-binary X, is applicable in the rare-disease setting, and in which there was either publicly available software, or in which the authors kindly shared their code. Table 3 similarly summarizes the results for the association of SNP with binary FT. Several of the methods required estimates of the disease prevalence and are indicated with an asterisk (*); we use the value of 0.0015 as determined from Solipaca and Ricciardi (2016) using data from 2010.

We first note that the results on the left and right portions of Tables 2 and 3 are not substantially different, indicating that age and gender do not have large confounding effects with SNP, but the differences are large enough to warrant inclusion of the left portions of the tables for more fair comparisons with the implementations that do not allow for covariate adjustment. The overall patterns appear the same across methods regardless of covariate inclusion. Second, we note that even when ignoring results from the naive analyses, there is some disagreement in the estimates of β_X within a column of a given table. Examining the results for continuous FT in Table 2, we make the following observations:

- Since FT and SNP are highly associated with disease status, we expect the "No disease status adjustment" method to be highly affected by the biased sampling scheme, and indeed, the estimates from this method are very large and substantially different from the other estimates.

- "Adjusted for disease status" is expected to perform similarly to the method of Lin and Zeng (2009) ("SPREG/SPML2") when the disease is rare, there is no interaction between the secondary outcome and covariates in the disease model, and the secondary outcome is normally distributed. The results in Table 2 numerically confirm this result, as the estimates for both of these methods are actually the same (p-values differ based on t vs standard normal null distribution). While lung cancer is clearly rare, the assumption of 'no interaction' between FT and SNP is questionable based on its moderate p-value (see Jiang *et al.* (2006) discussion in Section 2.4.1), as well as the noticeable difference in "Control-only" and "Case-only" estimates for β_X. This assumption of 'no interaction' likely contributes to the slightly larger estimates (left: –0.13; right: –0.11) shared by "Adjusted for disease status" and "SPREG/SPML2", and as well as "IPW$_R$" and "Semi" as compared to other approaches which do not make this assumption.

- The "Control-only" analysis should be approximately valid (at least in unmatched studies; see Reilly *et al.* (2005)) when the disease is rare, but is indeed quite inefficient. The "IPW" approach provided very similar results as "Control-only", as expected under rare disease, and should approximate the inference drawn from the population if the sampling probabilities are correctly specified. The "RECSO" and "WEE" approaches provided nearly identical results to "IPW", with no noticeable gains in efficiency over IPW. As indicated in Sofer *et al.* (2017a), the results in RECSO were not sensitive to the choice of covariates in the selection model (results not shown).

- The incorporation of the interrelationship between SNP, FT, and disease status via the inclusion of SNP in the specification for the correlation model between Y and the latent continuous version of D in SPML3 (Gghosh *et al.*, 2013) leads to similar estimates for β_X as those from the weighting methods "IPW", "RECSO", and "WEE". Somewhat surprisingly, we did not observe large differences in standard error estimates between these weighting methods and the SPML3 method.

- The "SV" method of Kang *et al.* (2017) yielded an estimate similar to, yet slightly higher than, the estimates from the "IPW", "RECSO", and "WEE" weighting methods. It is conjectured that the lack of a SNP by FT interaction in the (latent) disease model in the current implementation may be contributing to the higher estimate. Additionally, the "SV" method noticeably yielded the smallest SE estimate across all methods. This observation is in agreement with their simulations results, where Kang *et al.* (2017) found that the SE for SV tended to be lower than that of "SPREG" (Lin and Zeng, 2009) or "SPML3" (Gghosh *et al.*, 2013), and was not affected by low disease prevalence.

- For the "Semi" method of Ma and Carroll (2016), we first remark that estimates based on the computed initial values from their provided code were very different than those reported in Table 2, with a large estimate of intercept δ_0 in the disease model ($\hat{\delta}_0 = -0.1438$) that is not expected with our knowledge that lung cancer is a rare disease. The initial values were calculated using standard logistic regression, and their method does not require specification of a disease prevalence or the assumption of rare disease. The estimate for β_X and its SE produced using the initial values computed using the provided code were 0.2717, 0.05084, respectively. Providing a different (lower) initial value for δ_0 of (-5) yielded the results reported in the Table 2, and provided an estimate of δ_0 that was more in agreement with other methods that assumed rare disease or a known disease prevalence ($\hat{\delta}_0 = -6.5785$), but was also quite unstable (SE = 19.3286). The estimate of β_X in the Table (produced from the modified initial value) is on the larger side, consistent with other methods that do not include the SNP by FT interaction in the disease model.

Table 3: Comparison of estimates of β_X, their standard errors (SE), and corresponding p-values, using binary secondary outcome FT without (left) and with adjustment (right) for age and gender

	Method	No covariate adjustment			Covariate adjustment		
		Estimate	SE	P-value	Estimate	SE	P-value
Naive methods	No disease status adjustment	0.2253	0.04699	1.630e−06	0.2212	0.04801	4.069e−06
	Adjusted for disease status	0.1278	0.05084	1.192e−02	0.1162	0.05221	2.608e−02
	Control-only	0.0761	0.07551	3.137e−01	0.0702	0.07697	3.615e−01
	Case-only	0.1710	0.06898	1.316e−02	0.1496	0.07133	3.596e−02
Weighting methods	IPW*†	0.0771	0.07563	3.079e−01	0.0713	0.07695	3.541e−01
	WEE*$^{\#}$ (Song *et al.*, 2016a)	0.0650	0.07722	4.000e−01	0.0595	0.07891	4.508e−01
Joint modeling methods	SPML2*‡ (Jiang *et al.*, 2006)	0.1285	0.05075	1.132e−02	0.1174	0.05208	2.415e−02
	SPML2*$^{\diamond}$ (Jiang *et al.*, 2006)	0.0778	0.07554	3.028e−01	0.0653	0.07645	3.931e−01
	SPREG‡/SPML2 (Lin and Zeng, 2009)	0.1277	0.05082	1.195e−02	0.1165	0.05215	2.546e−02
	SPML3* (Jiang *et al.*, 2006)	0.0772	0.07524	3.046e−01	0.0714	0.07670	3.519e−01
	SPML3*$^{\spadesuit}$ (Gghosh *et al.*, 2013)	0.0772	0.07525	3.046e−01	0.0732	0.07678	3.401e−01
	SV* (Kang *et al.*, 2017)	0.0874	0.02912	2.665e−03			
	OR_tilde*$^{\bullet}$ (Wang and Shete, 2012)	0.0650	0.11678	5.751e−01			

*Using disease prevalence $\pi = 0.0015$. †Weights defined as in Wang and Shete (2011) for Extended IPW. ‡Disease model includes SNP, FT (no interaction). $^{\diamond}$Disease model includes SNP, FT, and SNP×FT interaction. $^{\bullet}$Allows for FT and SNP interaction in primary outcome model; using for secondary outcome prevalence 0.0775; p-value and standard error based on 10,000 bootstrap samples. $^{\#}$Calculated based on 10,000 bootstrap samples, and $T = 10$. $^{\spadesuit}$Model for the log-OR includes only SNP as a predictor, with no adjustment for age and gender.

- Both "IPW$_R$" (Xing *et al.*, 2016) and "Semi" (Ma and Carroll, 2016) do not assume normality of the errors of secondary outcome model, but the "IPW$_R$" method assumes a rare disease. This may contribute to the lower observed SE for "IPW$_R$" as compared to "Semi".

We additionally ran the POM methods of Lutz *et al.* (2014) and Ray and Basu (2017). Since the results based on these methods are not directly comparable with the estimates in Table 2, we just report their p-values for testing the association between FT and SNP: in covariate-adjusted analyses, the method of Lutz *et al.* (2014) yielded a p-value of 5.978e−02 while the POM-PS (v1.15) method of Ray and Basu (2017) yielded a p-value of 9.334e−02. In their simulations, neither Lutz *et al.* (2014) nor Ray and Basu (2017) examine the scenario of X by Y interaction in the population model for disease, so it is not clear how these results should be interpreted given what we have observed above.

Many of the same broad observations for continuous FT exist for the methods using the dichotomized FT in Table 3. In particular, the "No disease status adjustment" method produced the largest estimate of β_X across all methods considered and was substantially different from the other estimates; the "Control-only" and "IPW" results are very similar; the "SV" method produced similar but slightly higher estimates of β_X than "IPW" and had the smallest SE among all methods considered; and "Adjusted for disease status" approach behaved similarly to "SPREG/SPML2" and also "SPML2‡" from Jiang *et al.* (2006) when there was no interaction included in the disease model. Some additional observations follow:

- Again, the moderate p-value for the interaction between FT and SNP, as well as the noticeable difference in the "Control-only" and "Case-only" estimates of β_X calls into question the appropriateness of analyses that assume "no interaction" in the disease model. When the SPML2 approach of Jiang *et al.* (2006) includes the FT by SNP interaction ("SPML2$^{\diamond}$"), the results are much more similar to the "Control-only" and "IPW" results. The "SPML3" results from Jiang *et al.* (2006), which include SNP in the specification for the log-OR between Y and D, were even more similar to the "Control-only" and "IPW" results. This similarity of results between SPML2 with interaction and SPML3 was also observed in Jiang *et al.* (2006) in their data application. Interestingly,

when SNP was not included in the log-OR specification in SPML3, the results behaved similarly to those from the SPML2 models that did not include the FT by SNP interaction in the disease model (estimate = 0.1139, SE = 0.05218, p-value = 2.908e−02 in covariate adjusted model).

- With the binary FT outcome, the "WEE" approach yielded the smallest estimate for β_X and was similar to the "IPW" result, but not as close as in the continuous outcome setting. This is not entirely surprising, since the estimation strategies used in Song *et al.* (2016a) are different for continuous and binary secondary outcomes, with the latter requiring generation of T sets of pseudo counterfactual observations. The results reported used $T = 10$, as suggested by the authors.

- The "OR_tilde" bias correction method, which allows for FT by SNP interaction in the disease model, produces similar estimates as "WEE", but the estimated bootstrap SE based on 10000 bootstrap samples is higher. Also notable is that the methods of Wang and Shete require an estimate of the prevalence for the binary secondary outcome. For this, we used 0.0775 (Gallus *et al.*, 2005). The estimates of β_X were not highly sensitive to this choice, but the SEs tended to decrease with larger prevalence values for the binary secondary outcome (prevalence = 0.0385: estimate = 0.0676, SE = 0.16367; prevalence = 0.155: estimate = 0.0623, SE = 0.08887; both results based on 10,000 bootstrap samples).

- Under the binary FT framework, the model to be fit based on the implementation from Ghosh *et al.* (2013) is a special case of the model to be fit based the implementation from Jiang *et al.* (2006). With the implementation from Jiang *et al.* (2006), any subset of covariates can be specified in the three submodels in (2.10), while for the implementation from Ghosh *et al.* (2013) the same set of covariates are used in the two logistic models, and only the SNP is allowed in the log-OR model with no adjustment for additional covariates. Indeed, on the left side of Table 3, the results for the two SPML3 methods are nearly identical. When adjusting for covariates, if we rerun the SPML3 method from Jiang *et al.* (2006) with only SNP as a predictor in the log-OR model, we get identical results as the SPML3 method from Ghosh *et al.* (2013) on the right side of Table 3.

- The incorporation of the interrelationship between SNP, FT, and disease status (via the inclusion of the FT by SNP interaction in the disease model in SPML2 or "OR_tilde", or inclusion of SNP in the specification for the log-odds ratio between Y and D in SPML3) leads to similar estimates for β_X as those from the weighting methods ("IPW" and "WEE"). As in the continuous FT scenario, we again did not observe large differences in SE estimates between the weighting methods and these SPML methods which account for the three-way interrelationship.

4. Discussion

In this article, we began by briefly discussing the issues of various naive secondary analyses that do not account explicitly for case-control ascertainment, and in which situations these naive analyses are expected to provide valid inference. Unfortunately, these situations often cannot not be identified in practice, and can otherwise bias inference. As such, we reviewed a large collection of methods designed to explicitly account for the sampling bias in Sections 2.3 and 2.4. A pervasive theme throughout these two sections was the trade-off between robustness and efficiency. Generally speaking, the likelihood-based methods in Section 2.4.1 can often be substantially more efficient than most of the weighting approaches discussed in Section 2.3. At the same time, likelihood-based methods can suffer from severe lack of robustness when certain parts of the model are misspecified.

Finally, we compared the results of many of the reviewed methods in an example from a lung cancer case-control study. Echoing the "robustness" comments above, what was particularly striking in this example was the impact of the specification of the disease model in the methods that required it. Despite the "non-significant" interaction between Y and X on D, whether or not the three-way interrelationship of the variables was included in the model specification produced quite different results for β_X. We note that the example focused specifically on a situation in which both Y and X were associated with disease, in which the effects of sampling bias in naive analyses are expected to be the worst. In other scenarios, and with less evidence for an interaction, the differences in the results may not be as drastic. Nonetheless, the example illustrates that such differences can and do exist, and that researchers must exercise caution when conducting inference.

We conclude with a quote adapted from Jiang *et al.* (2006): "Unfortunately the above does not lead to any easy answers for practitioners. Because the potential efficiency losses of the weighted [IPW-type] method can be so severe, both [IPW-type and other more efficient joint modeling methods] should clearly be performed. Conclusions borne out by both types of analyses are particularly credible. Effects detected only by the [more efficient methods that rely on distributional assumptions] should be regarded with some suspicion and subjected to further investigation."

Acknowledgement

The author is most appreciative to Yanning Jiang, Chris Wild, and Fei Zou for sharing their code.

References

Breslow NE, Amorim G, Pettinger MB, and Rossouw J (2013). Using the whole cohort in the analysis of case-control data: application to the women's health initiative, *Statistics in Biosciences*, **5**.

Breslow NE and Cain KC (1988). Logistic regression for two-stage case–control data, *Biometrika*, **75**, 11–20.

Breslow NE and Day NE (1980). *Statistical Methods in Cancer Research Volume 1, The Analysis of Case Control Studies*, International Agency for Research on Cancer, Lyon.

Chen HY, Kittles R, and Zhang W (2013). Bias correction to secondary trait analysis with case-control design, *Statistics in Medicine*, **32**, 1494–1508.

Etter JF, Duc TV, and Perneger TV (1999). Validity of the Fagerström test for nicotine dependence and of the Heaviness of Smoking Index among relatively light smokers, *Addiction*, **94**, 269–281.

Flanders WD and Greenland S (1991). Analytic methods for two-stage case-control studies and other stratified designs, *Statistics in Medicine*, **10**, 739–747.

Gallus S, Pacifici R, Colombo P, La Vecchia C, Garattini S, Apolone G, and Zuccaro P (2005). Tobacco dependence in the general population in Italy, *Annals of Oncology*, **16**, 703–706.

Gazioglu S, Wei J, Jennings EM, and Carroll RJ (2013). A note on penalized regression spline estimation in the secondary analysis of case-control data, *Statistics in Biosciences*, **5**, 250–260.

Ghosh A, Wright F, and Zou F (2013). Unified analysis of secondary traits in case-control association studies, *Journal of the American Statistical Association*, **108**, 566–576.

Hancock DB, Reginsson GW, Gaddis NC, *et al.* (2015). Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence, *Translational Psychiatry*, **5**, e651.

He J, Li H, Edmondson AC, Rader DJ, and Li M (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies, *Biostatistics*, **13**, 497–508.

Heatherton TF, Kozlowski LT, Frecker RC, and Fagerstrom KO (1991). The Fagerström test for

nicotine dependence: a revision of the Fagerström tolerance questionnaire, *British Journal of Addiction*, **86**, 1119–1127.

Horvitz DG and Thompson DJ (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

Jiang Y, Scott AJ, and Wild CJ (2006). Secondary analysis of case-control data, *Statistics in Medicine*, **25**, 1323–1339.

Kang G, Bi W, Zhang H, *et al.* (2017). A robust and powerful set-valued approach to rare variant association analyses of secondary traits in case-control sequencing studies, *Genetics*, **205**, 1049–1062.

Kim RS and Kaplan RC (2014). Analysis of secondary outcomes in nested case-control study designs, *Statistics in Medicine*, **33**, 4215–4226.

Landi MT, Consonni D, Rotunno M, *et al.* (2008). Environment and Genetics in Lung Cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer, *BMC Public Health*, **8**, 203.

Lee AJ, McMurchy L, and Scott AJ (1997). Re-using data from case-control studies, *Statistics in Medicine*, **16**, 1377–1389.

Li H and Gail MH (2012). Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies, *Human Heredity*, **73**, 159–173.

Li H, Gail MH, Berndt S, and Chatterjee N (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies, *Genetic Epidemiology*, **433**, 427–433.

Liang L, Ma Y, Wei Y, and Carroll RJ (2018). Semiparametrically efficient estimation in quantile regression of secondary analysis, *Journal of the Royal Statistical Society, Series B*, **80**, 625–648.

Lin DY and Zeng D (2009). Proper analysis of secondary phenotype data in case-control association studies, *Genetic Epidemiology*, **33**, 256–265.

Liu DJ and Leal SM (2012). A flexible likelihood framework for detecting associations with secondary phenotypes in genetic studies using selected samples: application to sequence data, *European Journal of Human Genetics*, **20**, 449–456.

Lutz SM, Hokanson JE, and Lange C (2014). An alternative hypothesis testing strategy for secondary phenotype data in case-control genetic association studies, *Frontiers in Genetics*, **5**, 188.

Ma Y and Carroll RJ (2016). Semiparametric estimation in the secondary analysis of case-control studies, *Journal of the Royal Statistical Society, Series B*, **78**, 127–151.

Monsees GM, Tamimi RM, and Kraft P (2009). Genome-wide association scans for secondary traits using case-control samples, *Genetic Epidemiology*, **33**, 717–728.

Nagelkerke NJ, Moses S, Plummer FA, Brunham RC, and Fish D (1995). Logistic regression in case-control studies: the effect of using independent as dependent variables, *Statistics in Medicine*, **14**, 769–775.

Palmgren J (1989). *Regression models for bivariate binary responses*, School of Public Health and Community Medicine, University of Washington, 101.

Prentice RL and Pyke R (1979). Logistic disease incidence models and case-control studies, *Biometrika*, **66**, 403–411.

Ray D and Basu S (2017). A novel association test for multiple secondary phenotypes from a case-control GWAS, *Genetic Epidemiology*, **41**, 413–426.

Reilly M (1996). Optimal sampling strategies for two-stage studies, *American Journal of Epidemiology*, **143**, 92–100.

Reilly M, Torrang A, and Klint A (2005). Re-use of case–control data for analysis of new outcome

variables, *Statistics in Medicine*, **24**, 4009–4019.

Richardson DB, Rzehak P, Klenk J, and Weiland SK (2007). Analyses of case-control data for additional outcomes, *Epidemiology*, **18**, 441–445.

Robins JM, Rotnitzky A, and Zhao LP (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **89**, 198–203.

Rothman KJ (1986). *Modern Epidemiology*, Little Brown & Company, Boston.

Schifano ED, Bar H, and Harel O (2015). Methods for analyzing secondary outcomes in public health case-control studies, Chen DG (Din) and Wilson JR (Eds), (Chapter 1, pp. 3–15) *Innovative Statistical Methods for Public Health Data*, Springer, Switzerland.

Schifano ED, Li L, Christiani DC, and Lin X (2013). Genome-wide association analysis for multiple continuous phenotypes, *American Journal of Human Genetics*, **92**, 744–759.

Schlesselman JJ (1981). *Case-Control Studies: Design, Conduct, Analysis*, Oxford University Press, Oxford.

Scott AJ and Wild CJ (1986). Fitting logistic models under case-control or choice based sampling, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **48**, 170–192.

Scott AJ and Wild CJ (1995). *Maximum likelihood estimation for case-control data*, Department of Statistics, University of Auckland, 4.

Scott AJ and Wild CJ (2002). On the robustness of weighted methods for fitting models to case-control data, *Journal of the Royal Statistical Society: Series B*, **64**, 207–219.

Sofer T (2013). RECSO: Robust and Efficient Analysis using Control function approach, of a Secondary Outcome, R package version 1.0. Available from: https://CRAN.R-project.org/package=RECSO

Sofer T, Cornelis MC, Kraft P, and Tchetgen Tchetgen, EJ (2017a). Control function assisted IPW estimation with a secondary outcome in case-control studies, *Statistica Sinica*, **27**, 785–804.

Sofer T, Schifano ED, Christiani DC, and Lin X (2017b). Weighted pseudolikelihood for SNP set analysis of multiple secondary phenotypes in case–control genetic association studies, *Biometrics*, **73**, 1210–1220.

Solipaca A and Ricciardi W (2016). Rapporto Osservasalute: Stato di salute e qualità dell'assistenza nelle regioni italiane, *Osservatorio Nazionale Sulla Salute Nelle Regioni Itaniane*, 212–213.

Song X, Ionita-Laza I, Liu M, Reibman J, and Wei Y (2016a). A general and robust framework for secondary traits analysis, *Genetics*, **202**, 1329–1343.

Song X, Ionita-Laza I, Liu M, Reitman J, and Wei Y (2016). WEE: weighted estimated equation (WEE) approaches in genetic case-control studies, R package version 1.0. Available from: https://CRAN.R-project.org/package=WEE

Tapsoba JdeD, Kooperberg C, Reiner A, Wang CY, and Dai JY (2014). Robust estimation for secondary trait association in case-control genetic studies, *American Journal of Epidemiology*, **179**, 1264–1272.

Tchetgen Tchetgen E (2014). A general regression framework for a secondary, *Biostatistics*, **15**, 117–128.

Tseng TS, Park JY, Zabaleta J, Moody-Thomas S, Sothern MS, Chen T, Evans DE, and Lin HY (2014). Role of nicotine dependence on the relationship between variants in the nicotinic receptor genes and risk of lung adenocarcinoma, *PLoS ONE*, **9**, e107268.

VanderWeele TJ, Asomaning K, Tchetgen Tchetgen EJ, *et al.* (2012). Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction, *American Journal of Epidemiology*, **175**, 1013–1020.

Wang J and Shete S (2011a). Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases, *Genetic Epidemiology*, **35**, 190–200.

Wang J and Shete S (2011b). Power and type I error results for a bias-correction approach recently shown to provide accurate odds ratios of genetic variants for the secondary phenotypes associated with primary diseases, *Genetic Epidemiology*, **35**, 739–743.

Wang J and Shete S (2012). Analysis of secondary phenotype involving the interactive effect of the secondary phenotype and genetic variants on the primary disease, *Annals of Human Genetics*, **76**, 484–499.

Wei J, Carrroll RJ, Muller U, Van Keilegon I, and Chatterjee N (2013). Locally efficient estimation for homoscedastic regression in the secondary analysis of case-control data, *Journal of the Royal Statistical Society, Series B*, **75**, 186–206.

Wei Y, Song X, Liu M, Ionita-Laza I, and Reibman J (2016). Quantile regression in the secondary analysis of Case-control data, *Journal of the American Statistical Association*, **111**, 344–354.

Xing C, M McCarthy J, Dupuis J, Adrienne Cupples L, B Meigs J, Lin X, and S Allen A (2016). Robust analysis of secondary phenotypes in case-control genetic association studies, *Statistics in Medicine*, **35**, 4226–4237.

Zhao LP and Lipsitz S (1992). Designs and analysis of two-stage studies, *Statistics in Medicine*, **11**, 769–782.

Received August 13, 2018; Revised December 22, 2018; Accepted January 23, 2019