*Biostatistics* (2014), **15**, 1, *pp.* 117–128
doi:10.1093/biostatistics/kxt041
Advance Access publication on October 22, 2013

# A general regression framework for a secondary outcome in case–control studies

ERIC J. TCHETGEN TCHETGEN

*Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA and*
*Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA*

etchetgen@gmail.com

## Summary

Modern case–control studies typically involve the collection of data on a large number of outcomes, often at considerable logistical and monetary expense. These data are of potentially great value to subsequent researchers, who, although not necessarily concerned with the disease that defined the case series in the original study, may want to use the available information for a regression analysis involving a secondary outcome. Because cases and controls are selected with unequal probability, regression analysis involving a secondary outcome generally must acknowledge the sampling design. In this paper, the author presents a new framework for the analysis of secondary outcomes in case–control studies. The approach is based on a careful re-parameterization of the conditional model for the secondary outcome given the case–control outcome and regression covariates, in terms of (a) the population regression of interest of the secondary outcome given covariates and (b) the population regression of the case–control outcome on covariates. The error distribution for the secondary outcome given covariates and case–control status is otherwise unrestricted. For a continuous outcome, the approach sometimes reduces to extending model (a) by including a residual of (b) as a covariate. However, the framework is general in the sense that models (a) and (b) can take any functional form, and the methodology allows for an identity, log or logit link function for model (a).

*Keywords*: Case–control studies; Generalized linear models; Statistical genetics; Secondary outcomes.

## 1. Introduction

Case–control studies typically collect information on a large number of outcomes, often at considerable cost. These data are of potentially great value for studying associations, involving a secondary outcome other than the disease outcome defining case–control status. For instance, secondary outcomes analyses are now routine in genetic epidemiology, with several recent papers on genetic variants influencing human quantitative traits such as height, body mass index, and lipid levels, using data mostly from case–control studies of complex diseases (diabetes, cancer, and hypertension) (Lettre *and others*, 2008; Loos *and others*, 2008; Sanna *and others*, 2008; Weedon *and others*, 2007). Other examples have emerged in environmental epidemiology, such as the recent study of Weuve *and others* (2009), which uses data

taken, in part, from a case–control study nested within the Nurses' Health Study (NHS). In the NHS Lead Study, Boston-area NHS participants had extensive lead exposure assessment (bone and blood measures). Associations of lead measures with hypertension, bone mineral density/metabolism, and cognition were then assessed. However, the Lead Study selected women on the basis of their blood pressure status. Therefore, analyses that aim to evaluate risk factors of osteoporosis (a binary outcome) and cognitive function decline (a continuous outcome), may be affected by the case–control sampling design. In fact, Monsees *and others* (2009) and Lin and Zeng (2009) established that the non-random ascertainment from the study base, when ignored, can sometimes lead to inflated Type I error rate for tests of associations of a secondary outcome in re-purposed case–control samples. They further showed that commonly used analytic techniques, such as least-squares regression for quantitative traits, can sometimes give biased estimates, and that such bias can be present when covariates in the regression model in view, are associated with the case–control outcome, which itself is independently associated with the secondary outcome.

A number of analytic strategies have been proposed to eliminate selection bias associated with oversampling of cases in analyses of secondary outcomes; see, for instance, Nagelkerke *and others* (1995), Lee *and others* (1997), Jiang *and others* (2006), Reilly *and others* (2005), Richardson *and others* (2007), Lin and Zeng (2009), Monsees *and others* (2009), Li *and others* (2010), Wang and Shete (2011), and Wei *and others* (2013). Suggested strategies include: (i) weighting the standard analysis by the inverse of sampling probabilities (IPW); (ii) performing the analysis only in controls; (iii) analyzing cases and controls separately, i.e. stratifying the analysis by case–control status; (iv) including case–control status as a covariate in the regression model of the secondary outcome.

The first strategy (i) gives a viable simple solution as it recovers correct inferences about association measures, without the burden of additional modeling that would be required had data been sampled independently of case–control status. However, simply weighting by sampling rates will often be inefficient (Robins *and others*, 1994; Tchetgen Tchetgen, 2012). The second method is appropriate only when the disease status is rare in the population but does not use data on cases which might render it relatively inefficient. Methods that adjust for the primary disease status by either (iii) or (iv) may yield flawed conclusions because the associations between a secondary outcome and an exposure of interest in the case and control groups can be quite different from the association in the underlying target population. More formal likelihood methods have also appeared in the literature. For instance, (v) Jiang *and others* (2006) considered various likelihood methods for categorical secondary outcomes that can be more efficient than (i). (vi) Recently, Lin and Zeng (2009) further generalized the likelihood framework for a continuous secondary outcome by assuming the latter follows a specific parametric distribution, with special emphasis given to a normal model.

They also establish that the likelihood approach is well approximated by strategy (iv) under the following specific 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.

Thus, Lin and Zeng (2009) justify formally via a maximum likelihood argument, the conditional approach (iv) under conditions (LZ.1)–(LZ.3). More recently, (vii) Wei *and others* (2013) develop an estimating equations approach for a continuous secondary outcome which relaxes the distributional assumption made in (v) somewhat, and instead requires that the secondary outcome regression is "strongly homoscedastic" in the following sense. They assume that residuals from the secondary outcome regression are independent of covariates, but their density is otherwise unrestricted. In other words, they suppose that any association between the vector of covariates and the secondary outcome is completely captured by a location shift model. Their inferential framework relies crucially on this assumption, and may be biased if the assumption does not hold exactly.

An additional approach is proposed by Chen *and others* (2013) who uses a bias correction formula for an odds ratio parameter, while Ghosh *and others* (2013) adopt a retrospective likelihood framework, further extending the likelihood framework of Lin and Zeng (2009).

In this paper, the author generalizes the conditional approach (iv) to allow for possible violation of any or all of assumptions (LZ.1)–(LZ.3), without assuming the location shift model of Wei *and others* (2013). The new approach is based on a careful non-parametric re-parameterization of the conditional model for the secondary outcome given the case–control outcome and regression covariates, in terms of: (a) the population regression of interest for the secondary outcome given covariates and (b) the population regression of the case–control outcome on covariates.

Because non-parametric inference may not be practical for regression analysis with numerous covariates, parametric, and semiparametric models will invariably be used in practice for (a) for (b). Crucially, the re-parameterization ensures models for (a) and (b) are variation independent, in the sense that, whether parametric or semiparametric, a choice of model for (a) places no restriction on a corresponding choice of model for (b) and vice-versa. An important feature of the proposed strategy is that the error distribution for the secondary outcome conditional on covariates and case–control status is unrestricted. For a continuous outcome, the approach sometimes simplifies to extending model (a) by including the residual of (b) as an additional regression covariate producing a regression model for the secondary outcome conditional on case–control status, that is directly parameterized in terms of model (a). We show such a re-parameterization can appropriately account for selection bias without compromising inference about the population regression parameter. The framework is general in the sense that models (a) and (b) can take any functional form, and the methodology is developed to allow the use of the identity, log (described in supplementary material available at *Biostatistics* online) or logit link function in model (a). For inference, a simple estimating equation framework is first developed, and a strategy for obtaining a semiparametric locally efficient estimator is subsequently described. Simulations and an empirical example are used to illustrate the approach.

## 2. Regression with an identity link function

### 2.1 *Re-parameterization of conditional regression function*

Consider an unmatched case–control sample of i.i.d data consisting of the case–control status $D$, a continuous secondary outcome $Y$, and covariates $\mathbf{X}$. Unless otherwise stated, we will assume that the sampling fraction is known for cases and controls, respectively. This is often a reasonable assumption, that is, fairly standard in the literature on secondary outcomes (e.g. Jiang *and others*, 2006; Lin and Zeng, 2009; Wei *and others*, 2013), and the assumption is usually satisfied by design in nested case–control studies. An equivalent assumption is that the disease prevalence is known to be $\bar{p} = \Pr(D = 1)$ in the target population, and $\bar{\pi} = \Pr(D = 1|S = 1)$ in the case–control sample, where $S$ indicates selection into the case–control sample. Formally, $\bar{\pi}$ may be taken as the limiting proportion of cases in the case–control study with increasing sample size. As we will see below, the assumption that $\bar{p}$ is known will usually not be needed when the disease is rare for all levels of $\mathbf{X}$ in the population. The main target of inference is the population mean model for $Y$ given $\mathbf{X}$ which we denote for the identity link, $\mu(\mathbf{X}) = E(Y|X)$. A familiar example of such a model is

$$\mu(\mathbf{X}) = (1, \mathbf{X}^{\mathrm{T}})\beta_0. \tag{2.1}$$

We will also consider the conditional mean of $Y$ given $(\mathbf{X}, D)$, which we denote $\tilde{\mu}(\mathbf{X}, D) = E(Y|\mathbf{X}, D) = E(Y|\mathbf{X}, D, S = 1)$, where the second equality holds by design, because selection into the unmatched case–control sample is independent of $(Y, \mathbf{X})$ given $D$. Then, the following relation between $\mu(\mathbf{X})$ and $\tilde{\mu}(\mathbf{X}, D)$

also holds,

$$\begin{aligned}\mu(\mathbf{X}) &= \tilde{\mu}(\mathbf{X},1)\Pr(D=1|\mathbf{X}) + \tilde{\mu}(\mathbf{X},0)\Pr(D=0|\mathbf{X})\\ &\Leftrightarrow \begin{cases}\tilde{\mu}(\mathbf{X},1) = \mu(\mathbf{X}) + (1-\Pr(D=1|\mathbf{X}))\{\tilde{\mu}(\mathbf{X},1) - \tilde{\mu}(\mathbf{X},0)\},\\ \tilde{\mu}(\mathbf{X},0) = \mu(\mathbf{X}) + (0-\Pr(D=1|\mathbf{X}))\{\tilde{\mu}(\mathbf{X},1) - \tilde{\mu}(\mathbf{X},0)\},\end{cases}\\ &\Leftrightarrow \tilde{\mu}(\mathbf{X},D) = \mu(\mathbf{X}) + t\{D-\Pr(D=1|\mathbf{X})\}\{\tilde{\mu}(\mathbf{X},1) - \tilde{\mu}(\mathbf{X},0)\},\\ &= \mu(\mathbf{X}) + \{D-p(\mathbf{X})\}\gamma(\mathbf{X}),\end{aligned} \tag{2.2}$$

where $\gamma(\mathbf{X}) \equiv \{\tilde{\mu}(\mathbf{X},1) - \tilde{\mu}(\mathbf{X},0)\}$ describes the association between $Y$ and $D$ on the mean difference scale, within levels of $\mathbf{X}$, and $p(\mathbf{X}) \equiv \Pr(D=1|\mathbf{X})$ is the population risk of $D$ within levels of $\mathbf{X}$. From the alternative representation of $\tilde{\mu}(\mathbf{X}, D)$ given in the display above, one learns that the conditional mean function $\tilde{\mu}(\mathbf{X}, D)$ can be directly parameterized in terms of the population regression function of interest $\mu(\mathbf{X})$, and the additional functions $\{p(\mathbf{X}), \gamma(\mathbf{X})\}$. Note that the proposed re-parameterization is non-parametric, in the sense that it is not restricted to a particular choice of models for $\{p(\mathbf{X}), \gamma(\mathbf{X}), \mu(\mathbf{X})\}$ and therefore in principle, parametric, semiparametric, and non-parametric models can be used for each of these functions. Crucially, these functions are also variation independent, so that the choice of parameterization does not *a priori* rule out any possible data-generating mechanism. The function $\gamma(\mathbf{X})$ may be viewed as a selection bias function induced by an association between $D$ and $Y$ within levels of $\mathbf{X}$. Thus, the re-parameterization confirms what we might naturally expect, that the marginal and conditional regressions of $Y$ on $\mathbf{X}$ coincide exactly when selection bias is absent on the additive scale, i.e. $\tilde{\mu}(\mathbf{X}, D) = \mu(\mathbf{X})$ if $\gamma(\mathbf{X}) \equiv 0$. Furthermore, the re-parameterization ensures that even when $\gamma(\mathbf{x})$ is not zero for at least one level of $\mathbf{x}$, as one would hope, upon averaging over $D$ in the underlying population, $\tilde{\mu}(\mathbf{X}, D)$ reduces to $\mu(\mathbf{X})$ exactly. Additionally, one learns from the re-parameterization that when, (ETT.1)$\gamma(\mathbf{X}) = \gamma$ does not vary with $\mathbf{X}$, and, (ETT.2) the disease is rare in the population, so that $\tilde{\mu}(\mathbf{X}, D=0) \approx \mu(\mathbf{X})$ and $\tilde{\mu}(\mathbf{X}, D=1) = \mu(\mathbf{X}) + \{1-p(\mathbf{X})\}\gamma_0 \approx \mu(\mathbf{X}) + \gamma_0$, then $\tilde{\mu}(\mathbf{X}, D) \approx \mu(\mathbf{X}) + D\gamma_0$, which in the special case of model (2.1) takes the standard linear form $\tilde{\mu}(\mathbf{X}, D) \approx (1, \mathbf{X}^{\mathrm{T}})\beta_0 + D\gamma_0$. This implies that, under conditions (ETT.1) and (ETT.2), the simple strategy of extending the population model of interest $\mu(\mathbf{X})$ by adding a main effect for $D$ to the regression to adjust for case–control sampling, is approximately correct. Although the ensuing approximation to the regression $\tilde{\mu}(\mathbf{X}, D)$ is equal to that of Lin and Zeng (2009), we note that their required assumptions (LZ.1)–(LZ.3) imply assumptions (ETT.1) and (ETT.2), while the converse is not generally true. Specifically, it is straightforward to verify that assumptions (LZ.2) and (LZ.3) imply the no-heterogeneity assumption (ETT.1). However, without the normality assumption, (LZ.2) and (ETT.2) are not necessarily equivalent. The appeal of (ETT.2) is that it does not require distributional assumptions for the secondary outcome. Finally, one should note that (LZ.2) and (ETT.1) are empirically testable, and can be relaxed to account for possible effect heterogeneity. Specifically, as we will see in the next section, (ETT.1) may be relaxed by modeling $\gamma(\mathbf{X})$, which leads to the modified approximation $\tilde{\mu}(\mathbf{X}, D) \approx \mu(\mathbf{X}) + D\gamma(\mathbf{X})$. For instance, taking $\gamma(\mathbf{X}) = (1, \mathbf{X}^{\mathrm{T}})\boldsymbol{\gamma}_0$ gives under model (2.1), the standard linear form $\tilde{\mu}(\mathbf{X}, D) \approx (1, \mathbf{X}^{\mathrm{T}})\beta_0 + D(1, \mathbf{X}^{\mathrm{T}})\boldsymbol{\gamma}_0$. Crucially, one may note that while the modified approximation now incorporates possible interactions between $D$ and $\mathbf{X}$, i.e. $D(1, \mathbf{X}^{\mathrm{T}})\boldsymbol{\gamma}_0$, both the main effect of $D$ and its interactions with $\mathbf{X}$ are not interpretable as part of the marginal association between $\mathbf{X}$ and $Y$ for the population, only the first term of the approximate expression for $\tilde{\mu}(\mathbf{X}, D)$, i.e. $(1, \mathbf{X}^{\mathrm{T}})\beta_0$ encodes the marginal association of interest.

We should also note that, while we have assumed, and we will continue to do so unless otherwise noted, that the data arise from an unmatched case–control study, the above parameterization would continue to hold in the presence of matching, provided that the factors defining the matched set were also included in $\mathbf{X}$. This is illustrated in the matched case–control ovarian cancer study reported in Section 6.

### 2.2 *Inference via simple estimating equations*

Next, let $\pi(\mathbf{X}) \equiv \Pr(D = 1|\mathbf{X}, S = 1)$ denote the risk function of $D$ within levels of $\mathbf{X}$ in an unmatched case–control sample. $\pi(\mathbf{X})$ and $p(\mathbf{X})$ are well known to satisfy the following relation:

$$\operatorname{logit} p(\mathbf{X}) = \operatorname{logit} \pi(\mathbf{X}) + \log \frac{\bar{p}(1-\bar{\pi})}{\bar{\pi}(1-\bar{p})},$$

so that population and the case–control risks of $D$ agree on the logit scale, up to a constant shift in the intercept. Suppose that $\pi(\mathbf{X})$ follows a logistic model

$$\operatorname{logit} \pi(\mathbf{X}; \psi_0, \eta_0) = \eta_0 + m(\mathbf{X}; \psi_0), \tag{2.3}$$

where $m(\cdot; \psi)$ is a known function indexed by a parameter $\psi$ satisfying $m(0; \psi) = 0$, with unknown intercept $\eta_0$ and slope $\psi_0$. In the following, we will assume for simplicity that $m(\mathbf{X}; \psi_0) = \mathbf{X}^{\mathrm{T}}\psi_0$, although more elaborate models can also be used. Thus, $\operatorname{logit} p(\mathbf{X}; \eta_0, \psi_0) = \mathbf{X}^{\mathrm{T}}\psi_0 + \eta_0 + \log(\bar{p}(1-\bar{\pi})/\bar{\pi}(1-\bar{p}))$. For simplicity, we will also assume without loss of generality that $\gamma(\mathbf{X}) = \gamma(\mathbf{X}; \alpha_0)$, where $\gamma(\mathbf{X}; \alpha) = (1, \mathbf{X}^{\mathrm{T}})\alpha$. Together, these various parametric assumptions produce a corresponding model for $\tilde{\mu}(\mathbf{X}, D)$:

$$\begin{aligned} &\tilde{\mu}(\mathbf{X}, D; \theta_0) = (1, \mathbf{X}^{\mathrm{T}})\beta_0 + \{D - p(\mathbf{X}; \psi_0, \eta_0)\}(1, \mathbf{X}^{\mathrm{T}})\alpha_0, \\ &\text{where} \quad \theta_0 = (\beta_0^{\mathrm{T}}, \eta_0, \psi_0^{\mathrm{T}}, \alpha_0^{\mathrm{T}})^{\mathrm{T}}. \end{aligned} \tag{2.4}$$

We propose to estimate $(\eta_0, \psi_0')$ by standard logistic maximum likelihood of (2.3) using data on $(\mathbf{X}, D)$, i.e. by maximizing the partial log-likelihood $\sum_i L_i(\psi_0, \eta_0)$ w.r.t. $(\eta_0, \psi_0^{\mathrm{T}})$ where $L_i(\psi_0, \eta_0) = D_i \operatorname{logit} \pi(\mathbf{X}_i; \psi_0, \eta_0) + \log(1 - \pi(\mathbf{X}_i; \psi_0, \eta_0))$. For any $\theta = (\beta^{\mathrm{T}}, \eta, \psi^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$, let $\varepsilon(\theta) = Y - \tilde{\mu}(\mathbf{X}, D; \theta)$ and define the estimating function

$$\mathbf{U}(\theta) = \frac{\partial \tilde{\mu}(\mathbf{X}, D; \theta)}{\partial(\beta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}} \varepsilon(\theta). \tag{2.5}$$

This corresponds for the simple parametric models considered above to, $\mathbf{U}(\theta) = (1, \mathbf{X}^{\mathrm{T}}, (1, \mathbf{X}^{\mathrm{T}})\{D - p(\mathbf{X}; \psi, \eta)\})^{\mathrm{T}}\varepsilon(\theta)$, where $\varepsilon(\theta) = Y - (1, \mathbf{X}^{\mathrm{T}})\beta - \{D - p(\mathbf{X}; \psi, \eta)\}(1, \mathbf{X}')\alpha$. We propose to estimate $(\beta_0^{\mathrm{T}}, \alpha_0^{\mathrm{T}})$, with $(\hat{\beta}^{\mathrm{T}}, \hat{\alpha}^{\mathrm{T}})$ which solves $\mathbf{W}(\hat{\theta}) = \sum_i \mathbf{U}_i(\hat{\theta}) = 0$. In principle, one may specify any vector $\mathbf{h}(\mathbf{X}, D, \theta)$ of dimension $\dim((\beta_0^{\mathrm{T}}, \alpha_0^{\mathrm{T}})^{\mathrm{T}})$ in place of $\partial\tilde{\mu}(\mathbf{X}, D; \theta_0)/\partial(\beta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$ in (2.5), to obtain $\mathbf{U}(\boldsymbol{\theta}, \mathbf{h}) = \mathbf{h}(\mathbf{X}, D, \theta)\varepsilon(\theta)$ provided the derivative of the resulting estimating equation, more precisely its expectation, is not singular, and the variance–covariance matrix of $\mathbf{U}(\theta, \mathbf{h})$ is finite. Interestingly, IPW estimation is recovered upon setting $h(\mathbf{X}, D, \theta) = h'(\mathbf{X}, \theta)/\Pr(S = 1|D)$ and $\alpha = 0$, so that $D$ only appears in the inverse-probability weight for selection into the sample, and no longer in the outcome regression. From this observation, we also have that one can expect IPW to be suboptimal compared with the proposed approach, in the absence of model mis-specification, since by the proposition given in Section 4, setting $\alpha = 0$ and $h(\mathbf{X}, D, \theta) = h'(\mathbf{X}, \theta)/\Pr(S = 1|D)$ is inefficient. One may also verify using the proposition given in Section 4, that assuming $p(\mathbf{X})$ is known, the optimal choice of $\mathbf{h}$ is $\mathbf{h}_{\mathrm{opt}}(\mathbf{X}, D, \theta_0) = \{\partial\tilde{\mu}(\mathbf{X}, D; \theta_0)/\partial(\beta_0^{\mathrm{T}}, \alpha_0^{\mathrm{T}})^{\mathrm{T}}\}\mathrm{var}(\varepsilon(\theta_0)|\mathbf{X}, D)^{-1}$, and therefore $\mathbf{U}(\theta, \mathbf{h}_{\mathrm{opt}})$ is optimal, in the sense of producing an estimator with minimal asymptotic variance among regular and asymptotically linear estimators (RAL), when $\varepsilon(\theta)$ is homoscedastic and $p(\mathbf{X})$ is known. A standard argument shows that under standard regularity conditions, the resulting estimator $\hat{\theta}$ is in large sample approximately:

$$\hat{\theta} \dot{\sim} N(\theta_0, n^{-1}\Sigma(\theta_0)), \tag{2.6}$$

where $\Sigma(\theta)$ is the variance–covariance matrix of $\mathbb{E}[\partial(\mathbf{U}'(\theta), \mathbf{S}'(\psi, \eta))/\partial\theta]^{-1} \times (\mathbf{U}'(\theta), \mathbf{S}'(\psi, \eta))'$ with $\mathbf{S}(\psi, \eta) = \partial L(, \eta)/\partial((\psi', \eta')')$.

## 3. Regression with a logit link function

Next, suppose that $Y$ is binary. We introduce a similar re-parameterization of $\mathbb{E}(Y|\mathbf{X}, D) = \Pr(Y = 1|\mathbf{X}, D)$ on the logit scale, in terms of $\mathbb{E}(Y|\mathbf{X}) = \Pr(Y = 1|\mathbf{X})$. To proceed, let $\text{ODDS}(\mathbf{X}, D) = \Pr(Y = 1|\mathbf{X}, D)/\Pr(Y = 0|\mathbf{X}, D)$ denote the odds of $\{Y = 1\}$ within levels of $(\mathbf{X}, D)$. Likewise, let $\text{ODDS}(\mathbf{X}) = \Pr(Y = 1|\mathbf{X})/\Pr(Y = 0|\mathbf{X})$ denote the odds of $\{Y = 1\}$ within levels of $\mathbf{X}$. Then, note that

$$\begin{aligned}\frac{\tilde{\mu}(\mathbf{X}, D)}{1 - \tilde{\mu}(\mathbf{X}, D)} \equiv \text{ODDS}(\mathbf{X}, D) &= \frac{\text{ODDS}(\mathbf{X}, D)}{\text{ODDS}(\mathbf{X})} \times \text{ODDS}(\mathbf{X}) \\ &= \frac{\text{ODDS}(\mathbf{X}, D)}{\text{ODDS}(\mathbf{X}, D = 0)} \times \left\{ \sum_{d^*=0}^{1} \frac{\text{ODDS}(\mathbf{X}, d^*)}{\text{ODDS}(\mathbf{X}, D = 0)} \Pr(D = d^*|\mathbf{X}, Y = 0) \right\}^{-1} \times \text{ODDS}(\mathbf{X}) \\ &= \exp\left\{ \log \frac{\mu(\mathbf{X})}{1 - \mu(\mathbf{X})} + \nu(\mathbf{X}, D) - \bar{\nu}(\mathbf{X}) \right\}, \end{aligned} \tag{3.1}$$

where $\mu(\mathbf{X}) = \Pr(Y = 1|\mathbf{X})$ is the outcome risk function in the population, the function $\nu(\mathbf{X}, D) = \log \text{ODDS}(\mathbf{X}, D)/\text{ODDS}(\mathbf{X}, D = 0)$ measures the log-odds ratio association between $D$ and $Y$ within levels of $\mathbf{X}$, and accounts for selection bias due to the sampling design. As shown in supplementary material available at *Biostatistics* online, the parameter $\bar{\nu}(\mathbf{X}) = \log\{\exp\{\nu(\mathbf{X}, D = 1)\} \Pr(D = 1|\mathbf{X}, Y = 0) + \Pr(D = 0|\mathbf{X}, Y = 0)\}$ is not a free parameter, and is introduced to ensure that upon marginalization over $D$ in the target population, as one would hope to be the case, the conditional risk function $\tilde{\mu}(\mathbf{X}, D) = \Pr(Y = 1|\mathbf{X}, D)$ marginalizes to $\mu(\mathbf{X}) = \Pr(Y = 1|\mathbf{X})$ exactly. Interestingly, note that the population density of $D$ used in the above re-parameterization conditions on $\{Y = 0\}$ and hence differs from the density function of $D$ involved in previous re-parameterizations for the identity or log-link functions. This choice of parameterization is tied to a property of probability odds functions which is key to our developments. We have that, while $\mathbb{E}\{\text{ODDS}(\mathbf{X}, D)|\mathbf{X}\} \neq \text{ODDS}(\mathbf{X})$, it is, however, the case that $\mathbb{E}\{\text{ODDS}(\mathbf{X}, D)|\mathbf{X}, Y = 0\} = \text{ODDS}(\mathbf{X})$, in other words, marginalization of the conditional odds with respect to disease status in the underlying population free of the secondary outcome recovers the marginal odds function of primary interest. Equation (3.1) is equivalently written as a conditional logistic regression, $\tilde{\mu}(\mathbf{X}, D) = \Pr(Y = 1|D, \mathbf{X}) = [1 + \exp\{-\mu^\dagger(\mathbf{X}; \beta_0) - \nu(\mathbf{X}, D) + \bar{\nu}(\mathbf{X})\}]^{-1}$, where $\mu^\dagger(\mathbf{X}; \beta_0) = \log \mu(\mathbf{X})/\{1 - \mu(\mathbf{X})\}$. Suppose that the log-odds function $\mu^\dagger(\mathbf{X}; \beta_0) = (1, \mathbf{X})^{\mathrm{T}}\beta_0$, and the log-odds ratio function $\nu(\mathbf{X}, D) = \nu(\mathbf{X}, D; \alpha_0) = D(1, \mathbf{X}^{\mathrm{T}})\alpha_0$. We redefine

$$\text{logit}\, \pi(\mathbf{X}; \psi_0, \eta_0) = \text{logit}\, \Pr(D = d^*|\mathbf{X}, Y = 0, S = 1; \psi_0, \eta_0) = \eta_0 + m(\mathbf{X}; \psi_0), \tag{3.2}$$

using as before the convenient choice $m(\mathbf{X}; \psi_0) = \mathbf{X}^{\mathrm{T}}\psi_0$. Again, more elaborate models could be used to incorporate interactions and non-linearities. Let $\text{logit}\, \Pr(D = d^*|\mathbf{X}, Y = 0; \psi_0, \eta_0) = \text{logit}\, \pi(\mathbf{X}; \psi_0, \eta_0) + \log(\bar{p}(1 - \bar{\pi})/\bar{\pi}(1 - \bar{p}))$ denote the corresponding model in the population, accounting for retrospective ascertainment. The resulting parametric model for $\Pr(Y = 1|D, \mathbf{X})$ is given by

$$\text{logit}\, \Pr(Y = 1|D, \mathbf{X}; \theta_0) = (1, \mathbf{X})^{\mathrm{T}}\beta_0 + D(1, \mathbf{X}^{\mathrm{T}})\alpha_0 - \bar{\nu}(\mathbf{X}; \psi_0, \eta_0, \alpha_0), \tag{3.3}$$

where $\theta_0 = (\beta_0^{\mathrm{T}}, \eta_0, \psi_0^{\mathrm{T}}, \alpha_0^{\mathrm{T}})^{\mathrm{T}}$. Estimation and inference about $\theta_0$ can then proceed as in the identity or log link settings, by solving the estimating equation $\mathbf{W}(\hat{\theta}) = \mathbb{P}_n\mathbf{U}(\hat{\theta}) = 0$ given by (2.5), upon substituting in (3.3) for the conditional mean model $\tilde{\mu}(\mathbf{X}, D; \theta)$, but with $(\hat{\psi}, \hat{\eta})$ the maximum likelihood estimator

obtained using the log-likelihood function $\sum_i L_i(\boldsymbol{\psi}, \eta)$ where $L(\psi, \eta) = (1 - Y)\{D_i \operatorname{logit} \pi(\mathbf{X}; \psi, \eta) + \log(1 - \pi(\mathbf{X}; \psi, \eta))\}$. The asymptotic distribution of $\hat{\theta}$ is then given by (2.6) once the above substitution is made. Finally, we briefly note that when $D$ is rare, the logit link is well approximated by the log link and $\Pr(D = 1|\mathbf{X}, Y = 0) \approx \Pr(D = 1|\mathbf{X})$ and therefore the approximate approach developed for the log link also applies here, see supplementary material available at *Biostatistics* online.

## 4. Semiparametric locally efficient estimation

In this section, we present an alternative, potentially more efficient strategy for estimating $\theta_0$, based on semiparametric efficiency theory. To proceed, first note that as argued by Breslow *and others* (2000), the law of the observed data is formally given by the conditional density $f(Y, \mathbf{X}|D) = f(Y|\mathbf{X}, D) f(\mathbf{X}|D)$ which is up to a proportionality constant equivalent to the density of an experiment in which $D$ is itself randomly sampled from a Bernoulli density with known event probability equal to $\bar{\pi}$. Thus, we derive the efficient score for i.i.d data $(Y, \mathbf{X}, D)$ sampled from the joint density

$$f(Y|\mathbf{X}, D) f(\mathbf{X}|D) \bar{\pi}^D (1 - \bar{\pi})^{1-D} = f(Y|\mathbf{X}, D) \frac{f(D|\mathbf{X}) f(\mathbf{X})}{f(D)} \bar{\pi}^D (1 - \bar{\pi})^{1-D}$$
$$\propto f(Y|\mathbf{X}, D) f^*(D|\mathbf{X}) f^*(\mathbf{X}), \tag{4.1}$$

where $f(Y|\mathbf{X}, D)$ is the population density of $Y$ given $(\mathbf{X}, D)$, $f(D)$ is the known marginal density of $D$ in the target population; $f(D = 1|\mathbf{X}) = p(\mathbf{X})$ is the population probability that $D = 1$ given $\mathbf{X}$; $\operatorname{logit} f^*(D = 1|\mathbf{X}) = \operatorname{logit} \pi(\mathbf{X}) = \operatorname{logit} p(\mathbf{X}) - \log(\bar{p}(1 - \bar{\pi})/\bar{\pi}(1 - \bar{p}))$ is the probability that $D = 1$ given $\mathbf{X}$ in the case–control sample; $f^*(\mathbf{X}) \propto f(\mathbf{X})(f(D = 0|\mathbf{X})/f^*(D = 0|\mathbf{X}))$ is the case–control density of $\mathbf{X}$. Define the semiparametric model $\mathcal{M}_1$, with sole restrictions given by the restricted mean model $\tilde{\mu}(\mathbf{X}, D; \theta)$ for $Y$ given $(\mathbf{X}, D)$, with identity link (2.4) or log link (see supplementary material available at *Biostatistics* online); and the parametric model (2.3) for $D$ given $\mathbf{X}$. The model is otherwise non-parametric in the density of $\varepsilon(\theta) = Y - \tilde{\mu}(\mathbf{X}, D; \theta)$ given $(\mathbf{X}, D)$, as well as in the population density $f(\mathbf{X})$ and thus in $f^*(\mathbf{X})$.

The following theorem gives the efficient score for $\theta_0$ in model $\mathcal{M}_1$, a similar result for the logit link is relegated to supplementary material available at *Biostatistics* online.

Proposition 1 The efficient score of $\theta_0$ in model $\mathcal{M}_1$ is given by

$$\mathbf{R}(\theta_0) = \begin{pmatrix} \mathbf{R}_{(\beta,\alpha)}(\theta_0) \\ \mathbf{R}_{(\eta,\psi)}(\theta_0) \end{pmatrix},$$

with $\mathbf{R}_{(\beta,\alpha)} = \{\partial \tilde{\mu}(\mathbf{X}, D; \theta)/\partial(\beta', \alpha')'\}\mathbf{Z}(\theta)$ and $\mathbf{R}_{(\eta,\psi)}(\theta) = \mathbf{S}(\psi, \eta) + [\partial \tilde{\mu}(\mathbf{X}, D; \theta)/\partial(\psi', \eta)']\mathbf{Z}(\theta)$, where $\mathbf{Z}(\theta) = \operatorname{var}(\varepsilon(\theta)|\mathbf{X}, D)^{-1} \varepsilon(\theta)$.

Next, suppose that $\hat{\sigma}^2(\mathbf{X}, D, \hat{\theta}) = \widehat{\operatorname{var}}(\varepsilon(\hat{\theta})|\mathbf{X}, D)$ is a consistent estimate of the conditional variance $\sigma^2(\mathbf{X}, D, \theta_0) = \operatorname{var}(\varepsilon(\theta_0)|\mathbf{X}, D)$, then, upon defining $\hat{\mathbf{R}}(\theta)$ as $\mathbf{R}(\theta)$ by replacing $\sigma^2(\mathbf{X}, D)$ with $\hat{\sigma}^2(\mathbf{X}, D, \hat{\theta})$, the estimator $\hat{\theta}_{\text{eff}}$ that solves $\sum_i \hat{\mathbf{R}}_i(\hat{\theta}_{\text{eff}}) = 0$ is regular and asymptotically linear, with large sample variance the semiparametric efficiency bound in $\mathcal{M}_1$ which is given by $\mathbb{E}\{\mathbf{R}(\theta_0)\mathbf{R}^{\mathrm{T}}(\theta_0)\}^{-1}$. In practice, $\hat{\sigma}^2(\mathbf{X}, D, \hat{\theta})$ may be based on a parametric/semiparametric model, and therefore, may be inconsistent if mis-specified. Then, $\hat{\theta}_{\text{eff}}$ would still be RAL, although not necessarily asymptotically efficient. For this reason, $\hat{\theta}_{\text{eff}}$ is known as a semiparametric locally efficient estimator that is consistent and asymptotically normal regardless of whether $\hat{\sigma}^2(\mathbf{X}, D, \hat{\theta})$ is consistent or not, and that is asymptotically efficient at the submodel where $\hat{\sigma}^2(\mathbf{X}, D, \hat{\theta})$ is consistent. Interestingly, upon close inspection of the efficient

Table 1. *Simulation results*

| | Absolute bias | Variance | Coverage |
|---|---|---|---|
| $\beta_1 = 0$ | | | |
| Standard OLS | 0.734 | $2.2 \times 10^{-3}$ | 0.000 |
| Conditional OLS | 0.227 | $2.8 \times 10^{-3}$ | $4 \times 10^{-3}$ |
| IPW | $1.95 \times 10^{-4}$ | $3.3 \times 10^{-3}$ | 0.970 |
| Locally efficient | $1.11 \times 10^{-3}$ | $1.8 \times 10^{-3}$ | 0.960 |
| Simple estimating equation | $8.05 \times 10^{-5}$ | $3.5 \times 10^{-3}$ | 0.978 |
| $\beta_1 = 4$ | | | |
| Standard OLS | 0.730 | $2.3 \times 10^{-3}$ | 0.000 |
| Conditional OLS | 0.231 | $2.7 \times 10^{-3}$ | $2 \times 10^{-3}$ |
| IPW | $4.0 \times 10^{-3}$ | $3.4 \times 10^{-3}$ | 0.957 |
| Locally efficient | $4.2 \times 10^{-4}$ | $2.0 \times 10^{-3}$ | 0.956 |
| Simple estimating equation | $3.51 \times 10^{-3}$ | $3.6 \times 10^{-3}$ | 0.985 |

score $\mathbf{R}_{(\eta,\psi)}(\theta)$ one notes that information about $(\psi, \eta)$ the parameter indexing the density of $D$ given $\mathbf{X}$, naturally comes from the score of the corresponding factor of the likelihood function, i.e. $\mathbf{S}(\psi, \eta)$; however, additional information is obtained from the factor corresponding to the conditional density of $Y$ given $(D, \mathbf{X})$. Although unusual, this is not entirely surprising given that this density was carefully re-parameterized to depend on $(\psi, \eta)$. This further reveals that the simple estimating equations approach that gave $\hat{\theta}$ in previous sections, do not generally exploit this additional information since $(\hat{\psi}, \hat{\eta})$ solve the score equation $\mathbb{P}_n\{\mathbf{S}(\psi, \eta)\} = 0$ instead of the efficient score equation $\mathbb{P}_n\{\mathbf{R}_{(\eta,\psi)}(\theta)\} = 0$, and is therefore generally inefficient, except perhaps when the disease is rare.

## 5. A SIMULATION STUDY

We performed a simulation study to compare in the context of simple linear regression, the performance of the locally efficient estimator to that of two common strategies used in practice. The first approach involves inverse-probability weighting by the selection probability given case–control status, while the second approach involves including case–control status as a covariate in the regression for the secondary outcome. We also compared these methods to ordinary linear regression based on the entire data set, which one expects to be significantly biased. We generated $X$ from a mixture of normals with density $N(0, 4)$ with probability 0.88 and density $N(2, 4)$ otherwise. The logistic model is logit $\Pr(D = 1|X) = -2.5 + \psi_0 X$, where $\psi_0 = 0.5$. The model for $Y$ given $X$ is the linear regression model, $Y = 50 + \beta_1 X + \epsilon$, where $\epsilon|X$ is a mean zero residual error, that is, generated such that model (2.4) holds with $\gamma(X; \alpha_0) = 3 + 2\mathbf{X}$, and $\varepsilon(\theta_0)|D, X \sim N(0, 4)$. The simulation study explores both null ($\beta_1 = 0$) and non-null ($\beta_1 = 4$) conditions. The rate of disease is approximately 0.12 in the target population and therefore, the rare disease approximation does not hold. The case–control study has 500 cases and 500 controls, we generated 1000 simulated data sets.

For the simulation study, the locally efficient approach is implemented by maximizing the log-likelihood $\log\{f(\varepsilon(\theta)|X, D) f^*(D|X; \eta, \psi)\}$ which corresponds exactly to solving the efficient score of Proposition 1, under homoscedastic normal error, i.e. assuming $\varepsilon(\theta)|X, D \sim N(0, \sigma^2)$. This specific choice of likelihood model facilitates the implementation of the locally efficient approach using standard off-the shelf software, we used Proc NLMIXED in SAS to implement the approach.

The simulation results given in Table 1 confirm that IPW and the locally efficient approach both have small bias and produce 95% confidence intervals with appropriate coverage under either the null or the

alternative hypothesis. In contrast, as expected, ordinary linear regression using the entire sample and ignoring the sampling design is noticeably biased with disastrous coverage (=0%) in all scenarios. Simply adding a main effect for disease status corrects some of the bias but still produces 95% confidence intervals with poor coverage. In terms of efficiency, as expected, locally efficient estimation clearly outperforms IPW in both scenarios with relative efficiency sometimes >200%. Although remarkable, this efficiency gain is not entirely surprising from a semiparametric perspective, since by altogether avoiding to model $\gamma(\mathbf{X})$ and $p(\mathbf{X})$, IPW essentially allows these two models to remain unrestricted in estimating $\beta_1$, i.e. nonparametric, whereas the proposed approach relies crucially on parametric models for these functions to estimate $\beta_1$. These additional restrictions for the most part explain the efficiency gain.

We also implemented the inefficient estimating equations of Section 2.2, together with standard logistic maximum likelihood estimation of $\psi_0$. Although both approaches show little bias (Table 1), as projected by Proposition 1, the locally efficient estimator outperforms this alternative strategy in terms of efficiency and demonstrates remarkable efficiency gain not only for the parameter of primary interest $\beta_0$ ($\mathrm{ARE}(\beta_0) = 115\%$, $\mathrm{ARE}(\beta_1) = 180\%$), where $\mathrm{ARE}(\beta) = \mathrm{var}(\hat{\beta})/\mathrm{var}(\hat{\beta}_{\mathrm{eff}})$ but also for the logistic regression parameter $\psi_0$ ($\mathrm{ARE}(\psi_0) = 300\%$). This result confirms that as projected by Proposition 1, the locally efficient approach can, when the disease is not rare, recover information about $\psi_0$ that standard logistic regression cannot exploit.

## 6. An empirical application

This section illustrates the locally efficient approach in an analysis of data from a population-based case–control study of ovarian cancer (Modan *and others*, 2001). Two controls per case were selected from a central population registry in Israel, matching on age within 2 years, area of birth and place, and length of residence. Blood samples were collected on both cases and controls and were tested for the presence of mutation in two major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. Additional data were collected on reproductive and gynecologic history, such as parity, number of years of oral contraceptive use, and gynecologic surgery. The main objective of the study was to examine the interplay of the BRCA1/2 genes and known reproductive/gynecologic risk factors for ovarian cancer. In reanalyses of these data, a number of authors have exploited a gene-environment independence assumption to obtain more efficient estimates of interactions between BRCA1/2, and parity and oral contraceptive use, respectively (Chatterjee and Carroll, 2005; Tchetgen Tchetgen and Robins, 2010; Tchetgen Tchetgen, 2011). Specifically, they assumed that in the target population BRCA1/2 is jointly independent of parity and oral contraceptive within levels of covariates. As a secondary analysis, we evaluate this hypothesis empirically and estimate the mean association in the target population, between BRCA1/2 status and years of oral contraceptive use ($Y_1$) and parity ($Y_2$), respectively, adjusting for covariates. Thus, let $\mathbf{X} =$ (BRCA1/2, age (categorical defined by decades), ethnic background (Ashkenazi or non-Ashkenazi), the presence of personal history of breast cancer, a history of gynecologic surgery, and family history of breast or ovarian cancer (no cancer vs. one breast cancer in the family vs. one ovarian cancer or two or more breast cancer cases in the family)). The analysis uses data on 832 cases and 747 controls who did not have bilateral oophorectomy and who were interviewed for risk factor information and successfully tested for BRCA1/2 mutations. To illustrate the method with both identity and log link functions, $Y_1$ is coded as number of years of oral contraceptive use and a linear regression of $Y_1$ on $\mathbf{X}$ is evaluated, while $Y_2$ is a count of live births, and a log-linear model is assumed for the regression of $Y_2$ on $\mathbf{X}$. As suggested by Chatterjee and Carroll (2005), we set the population rate of ovarian cancer to $\bar{p} = 8.7 \times 10^{-4}$ which implies the rare disease approximation is appropriate, and thus an estimate of the risk of ovarian cancer as a function of $\mathbf{X}$ is not strictly needed. Nonetheless, we performed both analyses, with and without the rare disease approximation, and obtained identical results.

Table 2. *Parameter estimates (standard errors) of mean effect of BRCA1/2 on oral contraceptive use and Parity*

| | $Y_1$ | $Y_2$ |
|---|---|---|
| | BRCA1/2 (se) | BRCA1/2 (se) |
| Standard OLS | 0.212 (0.144) | −0.053 (0.047) |
| IPW | 0.327 (0.570) | $-5 \times 10^{-4}$ (0.142) |
| Locally efficient without interaction | 0.332 (0.152) | −0.020 (0.033) |
| Locally efficient with interaction | 0.287 (0.109) | 0.094 (0.175) |

Analyses further adjust for age, ethnic background, personal history of breast cancer, history of gynecologic surgery, and family history of breast or ovarian cancer.

For each outcome, we compare inferences based on standard OLS ignoring case–control status, IPW and the locally efficient approach with and without possible effect heterogeneity by BRCA1/2 in the case–control adjustment, i.e. $\gamma(\mathbf{X}; \alpha_0) = \alpha_0$ vs. $\gamma(\mathbf{X}; \alpha_0) = \alpha_0 + \alpha_1 \times \text{BRCA1/2}$.

Table 2 summarizes the results for BRCA1/2 associations with $Y_1$ and $Y_2$. In both sets of analyses, standard OLS gives the largest point estimates for the effect of BRCA1/2 on the average years of oral contraceptive use and parity, respectively. For both outcomes, IPW and the locally efficient approach incorporating a $D \times \text{BRCA1/2}$ interaction correct the OLS estimate, nonetheless the three methods agree in their conclusion and none rejects the null hypothesis of no gene-environment association at the $\alpha = 0.05$ level. Interestingly, not including the interaction in the locally efficient approach has different effects in the two analyses. For $Y_1$, not including the interaction leads to a wider Wald 95% confidence interval that rejects the null hypothesis of no BRCA1/2 association, which suggests the need to account for the interaction. In contrast, removing the interaction in the $Y_2$ regression leads to a shorter confidence interval without altering the overall conclusion, suggesting that perhaps the interaction is not necessary.

## 7. CONCLUSION

In this paper, we have described a general, yet simple framework for performing regression analysis for a secondary outcome in the context of case–control sampling. The current results focused on the three most common link functions used in practice, the identity link typically used for a continuous outcome, the log link typically used with counts, and the logit link typically used for binary data. A simple set of estimating equations is described for inference, and a potentially more efficient approach is also given. A particular appeal of the approach is that it is readily implemented with off-the-shelf statistical software. The framework also gives a formal justification for including the case–control status as a covariate in the regression model in view to account for study design when the case–control disease is rare, without requiring the distributional assumptions that have previously appeared in the literature. It is also straightforward to extend our basic argument to justify this type of conditional approach for other link functions, such as the complementary log-log link, or the probit link, under rare disease. When the disease is not rare, the approach requires that sampling fractions are known for cases and non-case–controls, which may be a challenge in certain settings, but is usually feasible when the case–control sample is nested within a well-defined cohort study. As we also describe, it is straightforward to use the proposed methods for matched case–control studies, simply by including matching factors in $\mathbf{X}$. Also note that as was done in the simulation study, the locally efficient estimator can sometimes be implemented using simple maximum likelihood, in which case, standard likelihood-based methods, such as Akaike's information criterion or the Bayesian information criterion can also be used to assess goodness-of-fit. Due to space restrictions, specific methods for goodness-of-fit methods will be explored in detail elsewhere.

Finally, an interesting and important direction for future work is to further develop the framework to handle settings where the secondary outcome is a vector of correlated variables, arising either from a longitudinal process, or due to spatial or other potential sources of clustering.

## Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

## Acknowledgments

*Conflict of Interest:* None declared.

## References

Breslow, N. E., Robins, J. M. and Wellner, J. A. (2000). On the semiparametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6**(3), 447–455.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92**, 399–418.

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

Ghosh, A., Wright, F. and Zou, F. (2013). Unified analysis of secondary traits in case-control association studies. *Journal of the American Statistical Association*, doi:10.1080/01621459.2013.793121.

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

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

Lettre, G., Jackson, A., Gieger, C., Schumacher, F. R., Berndt, S. and Hirschhorn, J. (2008). Identification of ten loci associated with height and previously unknown biological pathways in human growth. *Nature Genetics* **40**(5), 584–591.

Li, H., Gail, M. H., 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* **34**, 427–433.

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

Loos, R., Lindgren, C. M., Li, S., Wheeler, E. and Zhao, J. (2008). Association studies involving over 90,000 samples demonstrate that common variants near MC4R influence fat mass, weight and risk of obesity. *Nature Genetics* **40**(6), 768–775.

Modan, M. D., Hartge, P., Hirsh-Yechezkel G., Chetrit A., Lubin F., Beller U., Ben-Baruch G., Fishman A., Menczer J., Struewing J. P. *and others* (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New England Journal of Medicine* **345**, 235–40.

Monsees, G., Tamimi, R. and Kraft, P. (2009). Genomewide association scans for secondary traits using case-control samples. *Genetic Epidemiology* **33**, 717–728.

Nagelkerke, N. J. D., Moses, S., Plummer, F. A., Brunham, R. C. and Fish, D. (1995). Logistic regression in case–control studies: the effect of using independent as dependent variables. *Statistics in Medicine* **14**, 769–755.

REILLY, M., TORRANG, A. AND KLINT, A. (2005). Reuse of case–control data for analysis of new outcome variables. *Statistics in Medicine* **24**, 4009–4019.

RICHARDSON, D. B., RZEHAK, P., KLENK, J. AND WEILAND, S. K. (2007). Analysis of case–control data for additional outcomes. *Epidemiology* **18**, 441–445.

ROBINS, J. M., ROTNITZKY, A. AND ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866. Reproduced courtesy of the American Statistical Association.

SANNA, S., JACKSON, A. U., NAGARAJA, R., WILLER, C. J., CHEN, W. M., BONNYCASTLE, L. L., SHEN, H., TIMPSON, N., LETTRE, G., USALA, G. *and others* (2008). Common variants in the GDF5-UQCC region are associated with variation in human height. *Nature Genetics* **40**(2), 198–203.

TCHETGEN TCHETGEN, E. (2011). Robust discovery of genetic associations incorporating gene-environment interaction and independence. *Epidemiology* **22**(2), 262–272.

TCHETGEN TCHETGEN, E. J. (2012). Leveraging auxiliary information to enhance power in the analysis of nested case-control GWAS. *Technical Report*. Harvard University.

TCHETGEN TCHETGEN, E. J. AND ROBINS, J. (2010). The semi-parametric case-only estimator. *Biometrics* **66**(4), 1138–1144.

TCHETGEN TCHETGEN, E. J. AND ROTNITZKY, A. (2011). Double-robust estimation of an exposure-outcome odds ratio adjusting for confounding in cohort and case-control studies. *Statistics in Medicine* **30**(4), 335–347.

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

WEEDON, M. N., LETTRE, G., FREATHY, R. M., LINDGREN, C. M., VOIGHT, B. F., PERRY, J. R., ELLIOTT, K. S., HACKETT, R., GUIDUCCI, C., SHIELDS, B. *and others* (2007). A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nature Genetics* **39**(10), 1245–1250.

WEI, J., CARROLL, R. J., MULLER, U., VAN KEILEGOM, 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.

WEUVE, J., KORRICK, S., WEISSKOPF, M., RYAN, L., SCHWARTZ, J., NIE, H., GRODSTEIN, F. AND HU, H. (2009). Cumulative exposure to lead in relation to cognitive function in older women. *Environmental Health Perspectives* **117**, 574–580.

[*Received July 4, 2013; revised August 27, 2013; accepted for publication September 2, 2013*]