Published in final edited form as:
Biometrics. 2010 December ; 66(4): 1138–1144. doi:10.1111/j.1541-0420.2010.01401.x.

The semi-parametric case-only estimator

Eric J. Tchetgen Tchetgen and **James Robins**
Departments of Epidemiology and Biostatistics, Harvard University

Abstract

We propose a semi-parametric case-only estimator of multiplicative gene-environment or gene-gene interactions, under the assumption of conditional independence of the two factors given a vector of potential confounding variables. Our estimator yields valid inferences on the interaction function if either but not necessarily both of two unknown baseline functions of the confounders is correctly modeled. Furthermore, when both models are correct, our estimator has the smallest possible asymptotic variance for estimating the interaction parameter in a semi-parametric model that assumes that at least one but not necessarily both baseline models are correct.

Keywords

Gene-Environment interaction; Gene-Environment independence; Generalized odds ratio; Double robustness; Local efficiency

1 Introduction

The case-only design has recently become a popular approach for making inferences on the statistical interaction on the risk ratio scale between the effects of a genetic factor G and an environmental factor E (or a second genetic factor G') on the risk of a dichotomous disease status Y. Henceforth, we always refer to the second factor as an environmental factor in order to simplify the exposition. For dichotomous E and G, the crude case-only estimator of the interaction parameter is the empirical marginal odds ratio (i.e. crude odds ratio) between gene and environment among cases, and thus data on unaffected individuals is not required (Piegorsch et al, 1994). The validity of this estimator relies crucially on the assumption that gene and environment are independent in the population from which cases arose (Albert et al, 2001). Efficiency considerations have also contributed to the appeal of the case-only estimator; when the disease is rare, the case-only estimator is well known to be nearly efficient even when data on unaffected individuals are available as, for example, in a case-control study with controls sampled from subjects who remain unaffected at the end of the study. In contrast, under gene-environment independence, the standard prospective logistic regression estimator of gene-environment interaction can be less efficient. This is expected since it does not make use of the assumed independence (Prentice and Pyke, 1979; Breslow et al 2000).

While the case-only design allows for consistent estimation of G-E interaction on the risk ratio scale under the assumption of G-E independence, it does not allow for the estimation of the main effects of either G or E on the risk of disease, without which a meaningful interpretation of interactions may be difficult. In spite of this limitation, qualitative prior knowledge concerning main effects of both G and E can provide an appropriate background

[1]Corresponding Author:Eric Joel Tchetgen Tchetgen, Assistant professor of Epidemiology, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. etchetge@hsph.harvard.edu.

context for meaningful interpretation of the interaction parameters. For instance, Vanderweele et al (2009) recently discuss situations where a case-only estimate of a statistical interaction together with prior qualitative knowledge of main effects, may be interpreted causally as indicating the presence of a mechanistic interaction between the genetic factor and the environmental exposure.

To guarantee that the gene-environment independence assumption holds, it is often necessary to condition on additional covariates **Z** (Piegorsch et al, 1994; Gatto et al, 2004). Furthermore, it is often necessary to condition on covariates **W** that confound the association between disease and the gene and/or the environmental factor. As covariates included in **Z** may very well overlap with those included in **W**, we simplify notation by letting **L** dente $(\mathbf{Z}^T, \mathbf{W}^T)^T$ and further assume that conditional on **L**, *G* and *E* are independent and unmeasured confounding is absent. Furthermore, as in Chatterjee and Carroll (2005), we assume that the genetic factor has finite support $\{g_0, \ldots, g_{T-1}\}$ and thus we make use of the vector $\{I(G = g_1), \ldots, I(G = g_{T-1})\}$ which we again denote by *G* to simplify notation. The main objective of this paper is to provide a semi-parametric framework for making inferences under a case-only design on the generalized conditional log-odds ratio function

$$\theta(G, E, g_0, e_0, \mathbf{L}) = \log\left\{\frac{h(G, E|\mathbf{L}, Y=1)\, h(g_0, e_0|\mathbf{L}, Y=1)}{h(G, e_0|\mathbf{L}, Y=1)\, h(g_0, E|\mathbf{L}, Y=1)}\right\} \quad (1)$$

between *G* and *E* given **L** among cases ($Y = 1$), where $h(G, E|\mathbf{L}, Y = 1)$ is the retrospective density of *G* and *E* given **L** and $Y = 1$, and (g_0, e_0) is a user specified point in the sampling space which generally corresponds to an interpretable reference level for *G* and *E* such as (0, 0) in the simple case where both variables are binary. In light of the previous discussion, we formalize our model by assuming that:

(a.1) cases follow a semi-parametric log-binomial model for the density of *Y* given *G*, *E*, **L** with

$$log(P(Y=1|G, E, \mathbf{L}; \boldsymbol{\psi}_0)) = m_G(G, \mathbf{L}) + \gamma(G, E, \mathbf{L}; \boldsymbol{\psi}_0) + m_E(E, \mathbf{L}) + m_{\mathbf{L}}(\mathbf{L}),$$

where $\gamma(G, E, \mathbf{L}, \boldsymbol{\psi})$ is a known function and $\boldsymbol{\psi}_0$ is an unknown *p*–dimensional parameter. The term $\gamma(G, E, \mathbf{L}; \boldsymbol{\psi})$ is the generalized interaction function for the effects of *E* and *G* on *Y* within levels of **L**, on the exponential scale, and takes the value 0 if either $G = g_0$, $E = e_0$ or $\boldsymbol{\psi} = 0$, so that $\boldsymbol{\psi}_0 = 0$ encodes the null hypothesis of no interaction on the risk ratio scale. The terms $m_E(E, \mathbf{L})$ and $m_G(G, \mathbf{L})$ represent main effects for environment and genetic status, respectively within levels of **L**. We assume that $m_E(e_0, \mathbf{L}) = m_G(g_0, \mathbf{L}) = 0$ and that there exists a function $B(\mathbf{L})$ such that $\max_{G,E}[\max\{|m_G(G, \mathbf{L})|, |m_E(E, \mathbf{L})|, |\gamma(G, E, \mathbf{L}; \boldsymbol{\psi}_0)|\}] \leq B(\mathbf{L})$ for almost all **L**. The functions $m_G(G, \mathbf{L})$ and $m_E(E, \mathbf{L})$ are otherwise unrestricted, while the function $m_{\mathbf{L}}(\mathbf{L})$ is restricted only by the condition $\max_{G,E} P(Y = 1|G, E, \mathbf{L}; \boldsymbol{\psi}_0) < 1$ for almost all **L**. Note that, with this model, $\exp\{\gamma(G, E, \mathbf{L}; \boldsymbol{\psi}_0)\}$ is the *L*-specific generalized relative risk interaction function, i.e. multiplicative G-E interaction function (see Vansteelandt et al, 2008 for further detail on this model);

(a.2) *E* and *G* are conditionally independent given **L** so that the joint conditional density of gene and environment given **L** in the overall population, factorizes as $f_{G,E|\mathbf{L}}(G, E|\mathbf{L}) = f_{G|\mathbf{L}}(G|\mathbf{L}) \times f_{E|\mathbf{L}}(E|\mathbf{L})$, with $f_{G|\mathbf{L}}(G|\mathbf{L})$ and $f_{E|\mathbf{L}}(E|\mathbf{L})$ unrestricted conditional density functions;

(a.3) the population marginal density $f_{\mathbf{L}}(\mathbf{L})$ of **L** is unrestricted;

(a.4) the observed data (E, G, $\mathbf{L}$) are randomly sampled from the density $h_{G,\,E,\,\mathbf{L}|Y\,=\,1}$ (E, G, $\mathbf{L}|Y = 1$)

Under these assumptions, we show that

$$\gamma(G, E, \mathbf{L}; \boldsymbol{\psi}_0) = \theta(G, E, g_0, e_0, \mathbf{L}) \tag{2}$$

in other words, the generalized interaction function between E and G within levels of $\mathbf{L}$ coincides with the generalized odds ratio function relating genetic and environmental factors within levels of $\mathbf{L}$ among cases only.

In section 2, we show that assumptions (a.1)–(a.4) induce a semi-parametric model for the observed data characterized by the sole restriction (2). As our interest is in $\boldsymbol{\psi}_0$, we would prefer that the nuisance functions $m_E(E, \mathbf{L})$, $m_G(G, \mathbf{L})$, $f_{G|\mathbf{L}}(G|\mathbf{L})$, $f_{E|\mathbf{L}}(E|\mathbf{L})$ and $f_{\mathbf{L}}(\mathbf{L})$ remain unrestricted. However, due to the curse of dimensionality, this is not possible, when as we assume throughout, $\mathbf{L}$ is highly multivariate and/or includes two or more continuous covariates. Thus, we adopt the practical yet flexible approach of Chen (2007) and Tchetgen Tchetgen, Robins and Rotnitzky (2009). Specifically, we construct a so-called double robust (DR) estimator of $\boldsymbol{\psi}_0$ that is regular and asymptotically linear (RAL), and thus consistent and asymptotically normal (CAN), in the semi-parametric union model that assumes that one but not necessarily both of the following statements hold: i) a parametric model $\rho_G(G, \mathbf{L}, \boldsymbol{\alpha})$ indexed by $\boldsymbol{\alpha}$, for the conditional density function $\rho_G(G, \mathbf{L}) = f_{G|E,\,Y=1,\,\mathbf{L}}(G|E = e_0, Y = 1, \mathbf{L})$ is correct; or ii) a parametric model $\rho_E(E, \mathbf{L}; \boldsymbol{\eta})$ indexed by $\boldsymbol{\eta}$, for the conditional density function $\rho_E(E, \mathbf{L}) = f_{E|G,\,Y=1,\,\mathbf{L}}(E|G = g_0; Y = 1, \mathbf{L})$ is correct. It is because of this remarkable property that our estimator is called doubly robust, to reflect the fact that it provides the analyst with two separate chances for getting the correct answer. At the intersection submodel where both i) and ii) hold, our estimator is locally semi-parametric efficient in the union model, in the sense that it is a semi-parametric estimator in the union model whose asymptotic variance attains the semi-parametric variance bound for the union model at the intersection submodel. We should emphasize that estimation of the functions $m_{\mathbf{L}}(\mathbf{L})$ and $f_{\mathbf{L}}(\mathbf{L})$ is not required by our semi-parametric approach. Also, under a rare disease assumption, case-only estimation (in particular our proposed doubly robust case-only approach) may also be used to evaluate the parameters of a gene-environment generalized interaction function operating on the logit scale; because when the outcome is rare within all strata defined by (E, G, $\mathbf{L}$), assumption a. 1) may often serve as a practical approximation to the alternate assumption a*.1) that cases arise prospectively according to a semi-parametric logistic model $\text{logit}\{P(Y = 1|G, E, \mathbf{L}; \boldsymbol{\psi}_0)\} = m_G(G, \mathbf{L}) + \gamma(G, E, \mathbf{L}; \boldsymbol{\psi}_0) + m_E(E, \mathbf{L}) + m_{\mathbf{L}}(\mathbf{L})$; in which case our results would also hold approximately in this latter model. However, under the logistic model of a*.1), it is well known that departures from the rare disease assumption can result in a severely biased estimate of interaction parameters (Schmidt and Schaid, 1999). Furthermore, as we argue below in the discussion section, even when by common standards, the disease is considered rare, case-only test statistics and $(1 - \alpha)$ confidence intervals for interaction parameters on the odds ratio scale may substantially deviate from their nominal levels. This is in contrast to the case of gene-environment interactions on the risk ratio scale, where the case-only estimator remains valid whether or not the disease is rare (Schmidt and Schaid, 1999; Yang et al, 2004)

This article is organized as follows. In Section 2, we formally derive equation (2) which we use to construct RAL estimators of $\boldsymbol{\psi}_0$ in the union model. In Section 3, results of simulations illustrate the finite sample efficiency and robustness of our new estimators. In Section 4, we describe a simple specification test for detecting which of the models i) and ii) is correctly specified under the union model. In Section 5, the method is illustrated with the data from an Israeli study of the interactions between reproductive risk factors and

BRCA1/2 in their effects on the risk of ovarian cancer. Some closing remarks are provided in Section 6.

In the following, to simplify notation, we suppose $g_0 = 0$ and $e_0 = 0$ throughout, so that $\gamma(G, 0, \mathbf{L}; \boldsymbol{\psi}) = \gamma(0, E, \mathbf{L}; \boldsymbol{\psi}) = \gamma(G, E, \mathbf{L}; 0) = 0$.

2 Our estimators and their properties

The following theorem motivates our proposed approach

Theorem 1

Assumptions (a.1)–(a.4) imply that the observed data E, G, **L** *follows a semi-parametric model with sole restriction given by* equation (2).

Proof: By assumption (a.3) the conditional density $h_{\mathbf{L}|Y=1}(\mathbf{L}|Y=1)$ is unrestricted, thus it is sufficient to show that under assumptions (a.1), (a.2) and (a.4), the conditional density $h_{E,G|\mathbf{L},Y=1}(E, G|Y=1, \mathbf{L}; \boldsymbol{\psi}_0)$ of G and E given $\mathbf{L}$ and $Y = 1$ can be written

$$= \frac{h_{G,E|\mathbf{L},Y=1}(G,E|\mathbf{L},Y=1;\boldsymbol{\psi}_0)}{\int \exp\{m_G(G,\mathbf{L})+\gamma(G,E,\mathbf{L};\boldsymbol{\psi}_0)+m_E(E,\mathbf{L})\} \times f_{G|\mathbf{L}}(G|\mathbf{L}) \times f_{E|\mathbf{L}}(E|\mathbf{L})}$$

$$= \frac{\exp\{\gamma(G,E,\mathbf{L};\boldsymbol{\psi}_0)\} \times f_{G|E,Y=1,\mathbf{L}}(G|E=0,Y=1,\mathbf{L}) \times f_{E|G,Y=1\mathbf{L}}(E|G=0,Y=1,\mathbf{L})}{\int \exp\{\gamma(g,e,\mathbf{L};\boldsymbol{\psi}_0)\} \times f_{G|E,Y=1,\mathbf{L}}(g|E=0,Y=1,\mathbf{L}) \times f_{E|G,Y=1,\mathbf{L}}(e|G=0,Y=1,\mathbf{L}) d\mu(e,g)}$$

where the first equality follows from the definition of the conditional joint density

$$h_{E,G|\mathbf{L},Y=1}(E,G,\mathbf{L}|Y=1;\boldsymbol{\psi}_0) = \frac{\exp\{m_G(G,\mathbf{L})+\gamma(G,E,\mathbf{L};\boldsymbol{\psi}_0)+m_E(E,\mathbf{L})+m_{\mathbf{L}}(\mathbf{L})\} \times f_{G|\mathbf{L}}(G|\mathbf{L}) \times f_{E|\mathbf{L}}(E|\mathbf{L}) \times f_{\mathbf{L}}(\mathbf{L})}{\int \exp\{m_G(g,\mathbf{L})+\gamma(g,e,\mathbf{L};\boldsymbol{\psi}_0)+m_E(e,\mathbf{L})+m_{\mathbf{L}}(\mathbf{L})\} \times f_{G|\mathbf{L}}(g|\mathbf{L}) \times f_{E|\mathbf{L}}(e|\mathbf{L}) \times f_{\mathbf{L}}(\mathbf{L}) d\mu(g,e,\mathbf{L})}$$

and the second equality follows from (a.2.) and the cancellation of all functions of only **L** appearing in both the denominator and the numerator. Therefore,

$h_{G,E|\mathbf{L},Y=1}(G,E|\mathbf{L},Y=1;\boldsymbol{\psi}_0) = \frac{\exp\{\gamma(G,E,\mathbf{L};\boldsymbol{\psi}_0)\}\rho_G(G,\mathbf{L})\rho_E(E,\mathbf{L})}{\int \exp\{\gamma(g,e,\mathbf{L};\boldsymbol{\psi}_0)\}\rho_G(g,\mathbf{L})\rho_E(e,\mathbf{L})d\mu(e,g)}$ is a conditional density of $G, E|\mathbf{L}, Y = 1$ indexed by the conditional odds ratio function $\theta(G, E, g_0, e_0, \mathbf{L}) = \gamma(G, E, \mathbf{L}; \boldsymbol{\psi}_0)$. Because $\{m_G(G, \mathbf{L}), f_{G|\mathbf{L}}(G|\mathbf{L})\}$ are unrestricted, and $\{m_E(E, \mathbf{L}), f_{E|\mathbf{L}}(E|\mathbf{L})\}$ are unrestricted by (a.1) and (a.2), $\rho_G(G, \mathbf{L}) = f_{G|e_0,Y=1,\mathbf{L}}(G|E = e_0, Y = 1, \mathbf{L})$ and $\rho_E(E, \mathbf{L}) = f_{E|g_0,Y=1,,\mathbf{L}}(E|G = g_0, Y = 1, \mathbf{L})$ are unrestricted baseline conditional densities satisfying $\int \exp\{\gamma(g, e, \mathbf{L}; \boldsymbol{\psi})\} \rho_G(g, \mathbf{L}) \rho_E(e, \mathbf{L}) d\mu(g, e) < \infty$ for almost all **L**.

Consider working models $\rho_G(G, \mathbf{L}; \boldsymbol{\alpha})$ and $\rho_E(E, \mathbf{L}; \boldsymbol{\eta})$ for density functions $\rho_G(G, \mathbf{L})$ and $\rho_E(E, \mathbf{L})$ respectively, then an individual's conditional likelihood (given **L** and $Y = 1$) contribution for this model is $h_{G,E|\mathbf{L},Y=1}(G, E|\mathbf{L}, Y = 1; \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) =$

$$\frac{\exp\{\gamma(G,E,\mathbf{L};\boldsymbol{\psi})\}\rho_G(G,\mathbf{L};\alpha)\rho_E(E,\mathbf{L};\boldsymbol{\eta})}{\int \exp\{\gamma(g,e,\mathbf{L};\boldsymbol{\psi})\}\rho_G(g,\mathbf{L};\alpha)\rho_E(e,\mathbf{L};\boldsymbol{\eta}) d\mu(g,e)} \quad (3)$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are variation independent parameters. Now, we could in principle obtain likelihood-based inferences on $\boldsymbol{\psi}_0$ by taking one of the following three approaches: 1) we could obtain the maximum likelihood estimator (mle) $(\hat{\boldsymbol{\psi}}_{mle}, \hat{\boldsymbol{\eta}}_{mle}, \hat{\boldsymbol{\alpha}}_{mle})$ in model (3), which

is consistent provided the model is correct. However, because we can never know that the intersection submodel (where both $\rho_G(\cdot,\cdot;\boldsymbol{\alpha})$ and $\rho_E(\cdot,\cdot;\boldsymbol{\eta})$ hold) applies, this approach may be too restrictive; and thus we may prefer the alternate approach 2) which obtains the conditional mle (cmle) $((\hat{\boldsymbol{\psi}}_{cmle,1}, (\hat{\boldsymbol{\alpha}}_{cmle,1})$ based on the conditional likelihood of the density model for G given E, $\mathbf{L}$ and $Y = 1$; given by

$$h_{G|E,\mathbf{L},Y=1}(G|E,\mathbf{L},Y=1;\boldsymbol{\psi},\alpha)=\frac{\exp\{\gamma(G,E,\mathbf{L};\boldsymbol{\psi})\}\rho_G(G,\mathbf{L};\alpha)}{\int\exp\{\gamma(g,e,\mathbf{L};\boldsymbol{\psi})\}\rho_G(g,\mathbf{L};\alpha)d\mu(g,e)},$$

which does not depend on $\rho_E(E, \mathbf{L})$ unrestricted, but requires that $\rho_G(G, \mathbf{L}; \boldsymbol{\alpha})$ is correct; or alternately, we could obtain 3) the conditional mle $(\hat{\boldsymbol{\psi}}_{cmle,2}, \hat{\boldsymbol{\eta}}_{cmle,2})$ by maximizing the conditional likelihood for the density model of E given G, $\mathbf{L}$ and $Y = 1$; given by

$$h_{E|G,\mathbf{L},Y=1}(E|G,\mathbf{L},Y=1;\eta,\boldsymbol{\psi})=\frac{\exp\{\gamma(G,E,\mathbf{L};\boldsymbol{\psi})\}\rho_E(E,\mathbf{L};\eta)}{\int\exp\{\gamma(g,e,\mathbf{L};\boldsymbol{\psi})\}\rho_E(e,\mathbf{L};\eta)d\mu(g,e)},$$

which leaves $\rho_G(G, \mathbf{L})$ unrestricted but requires that $\rho_E(E, \mathbf{L}; \boldsymbol{\eta})$ is correct. Although more robust than the first approach, the analyst will generally not know which, if any, of the two models 2) or 3) is correct. For this reason, we construct a CAN estimator which is guaranteed to be consistent for the interaction function if at least one, but not necessarily both, of the baseline models $\rho_E(E, \mathbf{L}; \boldsymbol{\eta})$ or $\rho_G(G, \mathbf{L}; \boldsymbol{\alpha})$ is correct.

To proceed, let $\varepsilon(\boldsymbol{\psi}, \boldsymbol{\alpha}) = G - E(G|E, \mathbf{L}, Y = 1; \boldsymbol{\psi}, \boldsymbol{\alpha})$ where $E(\cdot|\cdot, \mathbf{L}, Y = 1; \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ denotes conditional expectations with respect to $h_{G,E|\mathbf{L},Y=1}(G, E|\mathbf{L}, Y = 1; \boldsymbol{\psi}, \boldsymbol{\eta}, \boldsymbol{\alpha})$. Define $\boldsymbol{\Psi}(E, \mathbf{L}; \boldsymbol{\psi}, \boldsymbol{\alpha}) = E\{\varepsilon(\boldsymbol{\psi}, \boldsymbol{\alpha})^{\otimes 2}|E, \mathbf{L}; \boldsymbol{\psi}, \boldsymbol{\alpha}\}$ and for a user-supplied $p \times (T-1)$–dimensional function $\mathbf{K} = \mathbf{k}(E, \mathbf{L})$, let $\tilde{U}(\mathbf{k}; \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \{\mathbf{K} - \tilde{E}(\mathbf{K}|\mathbf{L}; \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\eta})\} \times \varepsilon(\boldsymbol{\psi}, \boldsymbol{\alpha})$, be a function of p–dimensions; where $\tilde{E}\{\mathbf{K}|\mathbf{L}; \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\eta}\} = E\{\mathbf{k}(E, \mathbf{L}) \times \boldsymbol{\Psi}(E, \mathbf{L}; \boldsymbol{\psi}, \boldsymbol{\alpha})|\mathbf{L}; \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\eta}\} \times E\{\boldsymbol{\Psi}(E, \mathbf{L}; \boldsymbol{\psi}, \boldsymbol{\alpha})|\mathbf{L}; \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\eta}\}^{-1}$. By theorem 2 of Tchetgen Tchetgen and Robins (2009), the estimator $\hat{\boldsymbol{\psi}}(\mathbf{k})$ is RAL in the union model, where $\hat{\boldsymbol{\psi}}(\mathbf{k})$ is the solution to

$$\sum_{i=1}^{n}\tilde{U}_i\{\mathbf{k};\boldsymbol{\psi},\widehat{\alpha}(\boldsymbol{\psi}),\widehat{\eta}(\boldsymbol{\psi})\}=0,$$

and

$$\widehat{\alpha}(\boldsymbol{\psi})=\arg\max_{\alpha}\sum_{i=1}^{n}\log\{h_{G|E,\mathbf{L},Y=1}(G_i|E_i,\mathbf{L}_i,Y_i=1;\boldsymbol{\psi},\alpha)\},$$

is the profile MLE of $\boldsymbol{\alpha}$ at a fixed $\boldsymbol{\psi}$,

$$\widehat{\eta}(\boldsymbol{\psi})=\arg\max_{\eta}\sum_{i=1}^{n}\log\{h_{E|G,\mathbf{L},Y=1}(E_i|G_i,\mathbf{L}_i,Y_i=1;\boldsymbol{\psi},\eta)\}$$

is the profile MLE of $\boldsymbol{\eta}$ at a fixed $\boldsymbol{\psi}$. Furthermore, Tchetgen Tchetgen and Robins (2000) also show that the estimator $\hat{\boldsymbol{\psi}}_{eff} = \hat{\boldsymbol{\psi}}(\hat{\mathbf{k}}_{eff})$ is locally semi-parametric efficient in the union model at the intersection submodel, where $\widehat{\mathbf{k}}_{eff}(E,\mathbf{L})=\frac{\partial\{\gamma^T(E,\mathbf{L};\boldsymbol{\psi})\}}{\partial\boldsymbol{\psi}}|_{\psi=\tilde{\psi}}$, with $\gamma(E, \underset{\sim}{\mathbf{L}}; \boldsymbol{\psi})$ the $(T-1)\times 1$ vector with *jth* component equal to $\gamma(g_j, E, \mathbf{L}, \boldsymbol{\psi})$, $1 \le j \le T-1$; and $\tilde{\boldsymbol{\psi}}$ any preliminary estimator of $\boldsymbol{\psi}_0$ that is consistent at the intersection submodel (this could be the mle or either cmles). Furthermore, these authors prove that under the union model and

regardless of whether or not the intersection submodel holds, we have $\sqrt{n}\left\{\widehat{\boldsymbol{\psi}}\left(\widehat{\mathbf{k}}_{eff}\right)-\boldsymbol{\psi}_0\right\}$ is asymptotically normal, with variance consistently estimated by

$$\widehat{\sum}=n^{-1}\sum_{i=1}^{n}\left(\left[\begin{array}{c} n^{-1}\sum_{j=1}^{n}\frac{\partial}{\partial\boldsymbol{\psi}}M_j\left\{\left(\boldsymbol{\psi},\widehat{\eta}\left(\widehat{\boldsymbol{\psi}}_{eff}\right),\widehat{\alpha}\left(\widehat{\boldsymbol{\psi}}_{eff}\right);\widehat{\mathbf{k}}_{eff}\right)\mid_{\boldsymbol{\psi}=\widehat{\boldsymbol{\psi}}_{eff}}\right\}^{-1} \\ \times M_i\left\{\widehat{\boldsymbol{\psi}}_{eff},\widehat{\eta}\left(\widehat{\boldsymbol{\psi}}_{eff}\right),\widehat{\alpha}\left(\widehat{\boldsymbol{\psi}}_{eff}\right);\widehat{\mathbf{k}}_{eff}\right\}\end{array}\right]^{\otimes 2}\right)$$

where

$$\begin{aligned} M(\boldsymbol{\psi},\eta,\alpha;\mathbf{k})= \quad & \tilde{U}(\mathbf{k},\boldsymbol{\psi},\eta,\alpha) \\ & -E\left\{\frac{\partial}{\partial\eta}\tilde{U}(\mathbf{k};\boldsymbol{\psi},\eta,\alpha)\right\}E^{-1}\left\{\frac{\partial}{\partial\boldsymbol{\eta}}C(\boldsymbol{\psi},\boldsymbol{\eta})\right\}C(\boldsymbol{\psi},\eta) \\ & -E\left\{\frac{\partial}{\partial\alpha}\tilde{U}(\mathbf{k};\boldsymbol{\psi},\eta,\alpha)\right\}E^{-1}\left\{\frac{\partial}{\partial\alpha}B(\boldsymbol{\psi},\alpha)\right\}B(\boldsymbol{\psi},\alpha), \end{aligned} \tag{4}$$

$C(\boldsymbol{\psi},\eta)=\frac{\partial}{\partial\theta}\log\{h_{E|G,\mathbf{L},Y=1}(E_i|G_i,\mathbf{L}_i,Y_i=1;\boldsymbol{\psi},\eta)\}$ and $B(\boldsymbol{\psi},\alpha)=\frac{\partial}{\partial\alpha}\log\{h_{G|E,\mathbf{L},Y=1}(G_i|E_i,\mathbf{L}_i,Y_i=1;\boldsymbol{\psi},\alpha)\}$ are the scores for $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ respectively; which can be used to obtain Wald type confidence intervals for components of $\boldsymbol{\psi}_0$. Alternatively, inferences on $\boldsymbol{\psi}_0$ can be based on the non-parametric bootstrap.

An alternate approach that is approximately DR locally efficient was proposed by Chen (2007), however, his method requires the use of a computationally intensive alternating conditional expectations (ACE) algorithm to approximate the efficient score. Because in our setting, the efficient score can be written in closed form, we do not need Chen's approximate approach.

3 A Simulation Study

We perform a simulation study where we study the finite sample performance of $\hat{\psi}_{cmle,1}$, $\hat{\psi}_{cmle,2}$, $\hat{\psi}_{mle}$ and $\hat{\psi}_{eff}$ in 1000 data samples (of sample size n=200, 1000) consisting of variables $\mathbf{L}=\{(L_1,L_2),G,E\}$ generated by repeatedly sampling L_1 from a Bernoulli(1/2) density, L_2 from a normal(0,.75^2); and dichotomous pairs (G, E) from the conditional probability mass function (3) with

$\text{logit}\{\rho_G(1,\mathbf{L};\alpha)\}=\mathbf{L}_G'\alpha=(1,\mathbf{L}_1,L_2^2,L_1L_2^3)\alpha$, $\text{logit}\{\rho_E(1,\mathbf{L};\boldsymbol{\eta})\}=L_E'\boldsymbol{\eta}=(1,L_1,L_2^2,L_1L_2^3)\boldsymbol{\eta}$, where α' = (0.1,−1, 1, −1.1) and η' = (0.5, 0.5, −1.25, −1). Two values of ψ = 0, 1 are considered. The first three rows of results labeled ($\boldsymbol{\alpha}_{true}$, $\boldsymbol{\eta}_{true}$) in tables 1 and 2 correspond to an analysis that correctly specifies both baseline functions. The next three rows labeled ($\boldsymbol{\alpha}_{true}$, $\boldsymbol{\eta}_{false}$) correspond to an analysis that misspecifies $\rho_E(E,\mathbf{L};\boldsymbol{\eta})$ by using $\mathbf{L}'=(1,L_1,L_2,L_1L_2)$ in the regression model instead of $\mathbf{L}_E$, the following three rows labeled ($\boldsymbol{\alpha}_{false}$, η_{true}) report an analysis that misspecifies $\rho_G(G,\mathbf{L};\boldsymbol{\alpha})$ by using $\mathbf{L}$ in the regression model instead of $\mathbf{L}_G$, and the final three rows labeled ($\boldsymbol{\alpha}_{false}$, $\boldsymbol{\eta}_{false}$) correspond to an analysis where both baseline functions are incorrect. Under the correct model specification for the joint conditional density of E and G give $\mathbf{L}$ and Y = 1 (corresponding to the first three rows of results in tables 1 and 2); all four estimators have finite sample bias of comparable but negligible size. $\hat{\psi}_{eff}$ is surprisingly efficient relative to the other estimators, including the mle which suggests that in this simulation setting, little efficiency loss is incurred by our estimator in exchange for potential gain in robustness. Also under ($\boldsymbol{\alpha}_{true}$, $\boldsymbol{\eta}_{true}$), close to nominal type 1 error rates are achieved by all estimators in the null case ψ_0 = 0, and nearly nominal coverage rates are obtained in the case of ψ_0 = 1. Next, we find that as predicted by theory, $\hat{\psi}_{cmle,1}$ and $\hat{\psi}_{cmle,2}$ are both severely biased under misspecification of their respective baseline function as

nicely illustrated in the fourth row of the first table for n=1000, where the magnitude of the bias of $\hat{\psi}_{cmle,2}$ is over twenty-fold that of $\hat{\psi}_{eff}$ which remains negligible. Under model misspecification, the mle is both biased and yields CIs with incorrect coverage as illustrated by rows corresponding to settings ($\boldsymbol{\alpha}_{true}$, $\boldsymbol{\eta}_{false}$), ($\boldsymbol{\alpha}_{false}$, $\boldsymbol{\eta}_{true}$) and ($\boldsymbol{\alpha}_{false}$, $\boldsymbol{\eta}_{false}$) in both tables. Interestingly, the bias of the mle is always close to that of the misspecified cmle in rows corresponding to ($\boldsymbol{\alpha}_{true}$, $\boldsymbol{\eta}_{false}$) and ($\boldsymbol{\alpha}_{false}$, $\boldsymbol{\eta}_{true}$), with the bias, variance and coverage of $\hat{\psi}_{cmle,2}$ under ($\boldsymbol{\alpha}_{true}$, $\boldsymbol{\eta}_{false}$) always close to those of $\hat{\psi}_{cmle,1}$ under ($\boldsymbol{\alpha}_{false}$, $\boldsymbol{\eta}_{true}$) at both sample sizes and values of ψ_0. Also in the union model, we note that $\hat{\psi}_{eff}$ is surprisingly nearly as efficient as the correctly specified cmle with corresponding close to nominal coverage rates, as illustrated by the eighth and ninth rows of table two. Increasing sample size has the intended effect of reducing finite sample bias and variance of consistent estimators ($\hat{\psi}_{cmle,1}$, $\hat{\psi}_{eff}$) under ($\boldsymbol{\alpha}_{true}$, $\boldsymbol{\eta}_{false}$), and ($\hat{\psi}_{cmle,2}$, $\hat{\psi}_{eff}$) under ($\boldsymbol{\alpha}_{false}$, $\boldsymbol{\eta}_{true}$) for both null and alternative values of ψ_0. As expected, the bias of misspecified models is little affected by increasing sample size. Finally, no estimator gives valid results when both baseline functions are incorrect.

4 A Specification Test under the union model

Next we describe a simple specification test to detect which of the two baseline models ρ_E (E, **L; η**) and ρ_G (G, **L; α**) is correct under the union model. The approach is based on the following observation; if ρ_G (G, **L; α**) is correct, the corresponding cmle of ψ should be close (within sampling variability) to the doubly robust estimator whether or not ρ_E (E, **L; η**) holds, as both estimators should be consistent for the truth. In contrast, if ρ_G (G, **L; α**) is incorrect, the limiting value of the corresponding cmle should differ (beyond sampling variability) from that of the doubly robust estimator, since the former is biased, while the latter is consistent under the union model.

Thus, our test statistic is the suitably standardized (by its covariance) difference between the estimates of a subset of the cmle vector $\hat{\psi}_{cmle,1}$ obtained under model ρ_G (G, **L; α**) to those obtained from the proposed doubly robust method. This strategy essentially uses an idea due to Hausman (1978) and further developed by Newey (1985). In the simple case where ψ_0 is scalar, the statistic takes on the simple form $T=n\left(\widehat{\psi}_{dr}-\widehat{\psi}_{cmle,1}\right)^2/\left(\widehat{\sigma}^2_{dr}-\widehat{\sigma}^2_{cmle,1}\right)$, where $\widehat{\sigma}^2_{dr}$ and $\widehat{\sigma}^2_{\widehat{\psi},1}$ are consistent estimates of the variance of $n^{1/2}(\hat{\psi}_{dr}-\psi_0)$ and $n^{1/2}$ $(\hat{\psi}_{cmle,1}-\psi_0)$ respectively, and $\widehat{\sigma}^2_{dr}-\widehat{\sigma}^2_{cmle,1}$ is a consistent estimate of the variance of $n^{1/2}$ $(\hat{\psi}_{dr}-\hat{\psi}_{cmle,1})$ under the null hypothesis that both estimators are consistent for ψ_0. This estimator of the variance is of the simple form of a difference by virtue of $\hat{\psi}_{cmle,1}$ being an efficient estimator under the null semi-parametric model which specifies {γ (G, E, **L**, ψ), ρ_G (G, **L; α**)} (Hausman, 1978; Newey, 1985). Thus, under the null hypothesis, T has a χ^2_1 asymptotic distribution, and under the alternative it follows a non-central χ^2_1 asymptotic distribution, with non-centrality parameter solely determined by the direction of the asymptotic bias of $\hat{\boldsymbol{\psi}}_{cmle,1}$ (Newey; 1985). To guarantee that T is always positive, an analytical estimate of the variance of the numerator based on influence function arguments may be used, or alternately, the nonparametric bootstrap estimator of the variance of the numerator may replace the difference estimator in the denominator of the test statistic. A similar approach yields a specification test for ρ_E (E, **L; η**) under the union model. Outside of the union model, both cmles and the doubly robust estimator are converging to distinct values all of which are biased. Our specification test may still be used. However, the finite sample power of the test may be low under those rare alternatives where the two estimates being compared have similar bias, and thus are both wrong, and yet close to each other so that the proposed test statistic is unlikely to reject. The multivariate version of the test statistic is easily deduced from the previous description.

5 A Data Example

In this section, we illustrate the use of our methodology in an analysis of data from a population-based case-control study based on all ovarian cancer patients identified in Israel between 1 March 1994 and 30 June 1999 (Modan et al. 2001). Two controls per case were selected from the central population registry matching on age within two years, area of birth and place and length of residence. Blood samples was collected on both cases and controls and used to test for the presence of mutation in two major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. Additional data was collected on reproductive and gynecological history such as parity, number of years of oral contraceptive use and gynecological surgery. The main objective of the study was to examine the interplay of the BRCA1/2 genes and known reproductive/gynecological risk factors of ovarian cancer. To test for interactions between reproductive risk factors and BRCA1/2 in their effects on the risk of ovarian cancer, the authors performed the unadjusted case-only analysis of interaction of Piegorsch et al. (1994) under an assumption that genetic variants and environment factor are unconditionally independent in the population. Chatterjee and Carroll (2005) re-analyzed these data using a fully parametric logistic regression model for disease given the gene, environment and confounding factors, under the additional conditional independence assumption of gene and environment given a subset of measured covariates. However, their results are not directly comparable to ours for two reasons; first, because BRCA1/2 have high penetrance for the risk of ovarian cancer, the disease may not be rare within levels of G, E and $\mathbf{L}$, as a result, the interaction parameter on the logistic scale may differ from that of the risk ratio scale (in other words, assumption (a.1*) stated in the introduction may not hold). Second, even were the disease rare within all strata of G, E and $\mathbf{L}$, they assume a fully parametric model for the disease outcome regression model which may result in biased estimates of interactions if their specified working models for the functions $m_G(G, \mathbf{L})$, $m_E(E, \mathbf{L})$, $m_{\mathbf{L}}(\mathbf{L})$ are incorrect. As a consequence, our estimated interaction parameter is not directly comparable to that obtained by Chatterjee and Carroll (2005). In our re-analysis, we illustrate the case-only method developed in this paper by ignoring the data on controls. Specifically, using 832 cases who did not have bilateral oophorectomy, were interviewed for risk factor information and successfully tested for BRCA1/2 mutations. Our aim is to estimate the interaction between the dichotomous variable representing a person's BRCA1/2 mutation status and her use of oral contraceptives and parity. We dichotomize oral contraceptive use as use for over six years vs use for six years or less, while parity is dichotomized as having less than two children vs having two or more children. Thus we estimate (ψ_1, ψ_2) in the interaction model

$$\gamma(G, E, \mathbf{L}; \boldsymbol{\psi}) = \psi_1 \times I(BRCA1/2 = yes) \times I(OC\ use > 6yrs) + \psi_2 \times I(BRCA1/2 = yes) \times I(\#children > 1)$$

We assume conditional independence of gene and environmental factors given $\mathbf{L}$; consisting of age (categorical defined by decades), ethnic background (Ashkenazi or non-Ashkenazi), the presence of personal history of breast cancer, a history of gynecological 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). We specified a logistic model $\text{logit}[\rho_G\{1, \mathbf{L}; \alpha = (\alpha_0, \alpha_1^T)^T] = \alpha_0 + \mathbf{L}'\alpha_1$ for BRCA1/2 and a polytomous logistic model $\log\{\rho_E(e_k, \mathbf{L}; \boldsymbol{\eta}) / \rho_E(e_0, \mathbf{L}; \boldsymbol{\eta})\} = \eta_{0k} + \mathbf{L}'\boldsymbol{\eta}_{1,k}$, $k = 1, 2, 3$, where $\eta_k = (\eta_{0k}, \eta_k^T)^T$, $\eta = (\eta_1^T, \eta_2^T, \eta_3^T)^T$ and $e_0 = I(OC\ use \le 6yrs) \times I(\#children \le 1)$, $e_1 = I(OC\ use > 6yrs) \times I(\#children \le 1)$, $e_2 = I(OC\ use \le 6yrs) \times I(\#children > 1)$, $e_3 = I(OC\ use > 6yrs) \times I(\#children > 1)$. We then used these models to obtain four estimates $(\hat{\psi}_{1,cmle,1}, \hat{\psi}_{2,cmle,1}) = (0.99, 0.44)$ (bootstrap s.e.=0.43, 0.21), $(\hat{\psi}_{1,cmle,2}, \hat{\psi}_{2,cmle,2}) = (1.08; 0.44)$ (bootstrap

s.e.=0.42, 0.21), $(\hat{\psi}_{1,mle}, \hat{\psi}_{2,mle})$ = (1.05; 0.43) (bootstrap s.e.=0.40, 0.19) and $(\hat{\psi}_{1,eff}, \hat{\psi}_{2,eff})$ = (1.04,0.46) (bootstrap s.e.=0.42, 0.21) of the interaction parameters. All estimates indicate strong positive interactions between BRCA1/2 and oral contraceptive use and parity. These results suggest that after adjusting for background variables; the well known protective effects of oral contraceptive use and parity on ovarian cancer risk among BRCA1/2 non-carriers (Modan et al, 2001) may no longer apply to BRCA1/2 carriers. All point estimates are close to each other, providing convincing evidence that our models for both baseline functions are a reasonable fit to the data (in fact, the χ^2_2 specification test statistic using bootstrap covariance estimates yields p-values>0.99 in comparing each cmle to the doubly robust estimate). Interestingly, all estimators have very similar uncertainty, suggesting that as we also found in our simulation study, we do not necessarily give up much efficiency for the important gain in robustness afforded by the doubly robust approach.

Finally, despite not being directly comparable to our results, we apply the standard logistic regression approach using data both on cases and controls to estimate the interaction between the genetic variant and OC use/parity indicators in a regression model which also included main effects for BRCA1/2 indicator, OC use indicator, parity indicator, and **L**. We obtained the following imprecise estimates $(\hat{\psi}_{1,logistic}, \hat{\psi}_{2,logistic})$ = (0.32, 0.005) (s.e.=1.12, 0.80), which, in light of the improved accuracy of our proposed locally efficient approach, further illustrates the inefficiency of logistic regression for case-control data under gene-environment independence. The first three of our estimators can be obtained from standard software. The doubly robust locally efficient estimator is implemented in IML, SAS 9.1.3, and can be downloaded from the first author's website. The data can freely be obtained from Nilanjan Chatterjee's website at the Biostatistics Branch of the National Cancer Institute.

6 Discussion

We have developed a doubly robust locally-efficient estimator of interactions within a semi-parametric case-only framework. The proposed methodology is a flexible and efficient extension of the original crude and adjusted case-only estimators. We recommend its use particularly in settings where inferences on gene-environment or gene-gene interactions are sought, genetic information is only collected in cases, the factors of interest are known to be conditionally independent and as in most observational studies, it is necessary to further adjust for high dimensional confounders. However, if either the conditional independence assumption does not hold or it does hold but the disease is not rare and data on the unaffected is available, the approach of Vansteelandt S. et al. (2009) should be preferred to that presented herein. This is because the case-only estimator is no longer valid in the first setting, whereas in the second situation it is still valid, but does not make use of all of the available information and is therefore not efficient.

As mentioned in the introduction, when the target of inference is a parameter ψ_0 for a gene-environment interaction operating on the odds ratio scale, as in the logistic regression given in *a**.1) and assumption *a*.2) is satisfied, seriously flawed inferences may still result from a case-only approach, even if by common standards, the disease is considered to be rare within levels of (*G*, *E*, *L*) in the population. To illustrate this point, consider the following rare disease asymptotic analysis. Suppose for simplicity that there are no covariates **L** and (*G*, *E*) are both dichotomous, so that at sample size *n*, the observed data is generated under the model $\text{logit}\{P(Y=1|G,E;\boldsymbol{\psi}_0,\beta_0,p_n)\} = \log\{p_n/(1-p_n)\}+\beta_0(G+E)$ such that $p_n \ll 1$ is a sequence of positive numbers which converges to zero as *n* goes to infinity, $\boldsymbol{\psi}_0 = 0$ and $m_G(\cdot) = m_E(\cdot) = \beta_0 \times \cdot$. A Taylor series expansion around the limit point $\lim_{n\to\infty} p_n = 0$ shows that the case-only estimand, i.e. the log-odds ratio relating *G* to *E* in cases only (*Y* = 1), is to first-order equal to $r(\beta_0)\, p_n$ where $r(\beta_0) = 2\{\exp(2\beta_0)-1\} - \{\exp(\beta_0)-1\}$. This observation

implies that since, the case-only estimator $\hat{\psi}$ is asymptotically unbiased for the case only estimand, the case-only estimator $\hat{\psi}$, viewed as an estimator of the log prospective odds ratio parameter $\psi_0 = 0$, will have asymptotic bias equal to the large sample limit of $n^{1/2}r\,(\beta_0)p_n$. This in turn leads to the conclusion that whenever $\beta_0 \neq 0$ and p_n converges to zero at a rate no faster than $n^{-1/2}$, a case-only hypothesis test of the null that $\psi_0 = 0$, via say, either checking whether a case-only Wald-type $(1 - \alpha)$-confidence interval contains zero, or whether the value of a Wald-type case-only test statistic rejects under the null, will in general have incorrect coverage and type I error respectively. The practical implication of this asymptotic analysis is that the rare disease assumption may not be useful in settings where disease prevalence although small, is roughly of the order of magnitude $n^{-1/2}$ of the standard error of the case-only estimates, or of a larger order of magnitude (say $n^{-1/4} > n^{-1/2}$).

In the current paper, we have assumed that **L** includes all factors that one needs to adjust for in the gene-environment conditional independence model and in the disease-risk model. However, as pointed out by a referee, the case-only design adopted in this paper may not always be appropriate when there are variables **V** listed in **L** such that *G* and *E* are independent given **L**, but *G* and *E* are not independent given **L****V**; and the disease risk model of interest and thus the G–E interaction function of primary interest do not condition on **V** (see Chatterjee and Chen, 2007 for an example). The current methods would still apply in such a situation if **V** were also known not to be associated with disease given **L****V**, which implies but is not implied by the following assumption: $\gamma(G, E, \mathbf{L}; \boldsymbol{\psi}_0)$ does not depend on *V*. Thus the *G–E* conditional interaction parameter $\boldsymbol{\psi}_0$ would have the desired marginal interpretation, while the baseline densities $\rho_E(E, \mathbf{L})$ and $\rho_G(G, \mathbf{L})$ would remain functions of the entire vector **L**.

Acknowledgments

The author is grateful to the editor, the associate editor and two referees for their constructive and insightful comments.

References

1. Albert P, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the Case-only Design for Identifying Gene-Environment Interactions. American Journal of Epidemiology. 2001; 154(8):687–693. [PubMed: 11590080]
2. Breslow NE. Statistics in epidemiology: the case-control study. Journal of the American Statistical Association. 1996; 91:14–28. [PubMed: 12155399]
3. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. 2005; 92:399–418.
4. Chatterjee N, Chen YH. Maximum likelihood inference on a mixed conditionally and marginally specied regression model for genetic epidemiologic studies with two phase sampling. Journal of the Royal Statistical Society Series B-Statistical Methodology. 2007; 69:123–142.
5. Chen YH. A Semi-parametric Odds Ratio Model for Measuring Association. Biometrics. 2007; Volume 63(Number 2):413–421. 9.
6. Gatto N, Campbell U, Rundle A, Ahsa H. Further development of the case-only design for assessing gene-environmnent interaction: evaluation of and adjustment for bias. International Journal of Epidemiology. 2004; 33:1014–1024. [PubMed: 15358745]
7. Hausman JA. Specication tests in econometrics. Econometrica. 1978; 46:1251–1272.
8. Modan MD, Hartge P, et al. Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. New England Journal of Medicine. 2001; 345:235–240. [PubMed: 11474660]
9. Newey WK. Generalized Method of Moments Specication Testing. Journal of Econometrics. 1985; 29:229–256.

10. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Statistics in Medicine. 1994; 13:153–162. [PubMed: 8122051]

11. Schmidt S, Schaid D. Potential Misinterpretation of the Case-Only Study to Assess Gene-Environment Interaction. American Journal of Epidemiology. 1999; 150(8):878–885. [PubMed: 10522659]

12. Tchetgen Tchetgen E, Robins J, Rotnitzky A. On doubly robust estimation in a semi-parametric odds ratio model. Biometrika. 2009 In press.

13. Umbach DM, Weinberg CR. Designing and analyzing case-control studies to exploit independence of genotype and exposure. Statistics in Medicine. 1997; 16(15):1731–1743. [PubMed: 9265696]

14. Vanderweele, T.; Hernandez-Diaz, S.; Hernan, M. Technical Report. Harvard School of Public Health; 2009. Case-only gene-environment interaction studies: when does association imply mechanistic interactions?.

15. Vansteelandt S, VanderWeele TJ, Tchetgen E, Robins JM. Multiply robust inference for statistical interactions. Journal of the American Statistical Association. 2008; Vol. 103(No. 484) Theory and Methods.

16. Yang Q, Khoury MJ, Sun F, Flanders D. Case-Only Design to Measure Gene-Gene Interaction. Epidemiology. 1999; Vol. 10(No. 2):167–170. [PubMed: 10069253]

Table 1

Simulation Results for $\psi_0 = 0$

		n=200				n=1000			
		$\hat{\psi}_{cmle,1}$	$\hat{\psi}_{cmle,2}$	$\hat{\psi}_{mle}$	$\hat{\psi}_{eff}$	$\hat{\psi}_{cmle,1}$	$\hat{\psi}_{cmle,2}$	$\hat{\psi}_{mle}$	$\hat{\psi}_{eff}$
α_{true}, η_{true}	bias	0.004	0.008	0.004	0.007	0.007	0.005	0.005	0.007
	variance	0.123	0.113	0.125	0.113	0.023	0.022	0.021	0.023
	Coverage[#]	0.956	0.958	0.955	0.955	0.944	0.956	0.953	0.951
α_{true}, η_{false}	bias	0.004	−0.240	−0.239	0.009	0.007	−0.245	0.245	0.007
	variance	0.123	0.114	0.114	0.134	0.023	0.020	0.020	0.023
	Coverage	0.956	0.891	0.895	0.945	0.944	0.584	0.587	0.949
α_{false}, η_{true}	bias	−0.245	0.008	−0.227	0.007	−0.245	0.005	−0.233	0.005
	variance	0.108	0.113	0.098	0.125	0.020	0.022	0.019	0.023
	Coverage	0.875	0.958	0.882	0.952	0.593	0.956	0.618	0.953
α_{false}, η_{false}	bias	−0.245	−0.240	−0.229	−0.229	−0.245	−0.245	−0.252	−0.255
	variance	0.108	0.114	0.108	0.107	0.020	0.020	0.021	0.020
	Coverage	0.875	0.891	0.880	0.894	0.593	0.584	0.563	0.550

[#]Coverage of Wald confidence intervals using the second derivative of the score equations to estimate the variance of *cmles* and of the mle, and $\hat{\Sigma}$ to estimate the variance of $\hat{\psi}_{eff}$.

Table 2

Simulation Results for $\psi_0 = 1$

		n=200				n=1000			
		$\hat{\psi}_{cmle,1}$	$\hat{\psi}_{cmle,2}$	$\hat{\psi}_{mle}$	$\hat{\psi}_{eff}$	$\hat{\psi}_{cmle,1}$	$\hat{\psi}_{cmle,2}$	$\hat{\psi}_{mle}$	$\hat{\psi}_{eff}$
α_{true}, η_{true}	bias*	0.025	0.025	0.023	0.026	0.008	0.009	0.008	0.009
	variance*	0.140	0.124	0.120	0.143	0.030	0.027	0.027	0.030
	coverage#	0.971	0.971	0.972	0.963	0.951	0.938	0.935	0.948
α_{true}, η_{false}	bias	0.025	−0.234	−0.236	0.028	0.008	−0.253	−0.254	0.002
	variance	0.140	0.136	0.135	0.161	0.030	0.026	0.025	0.030
	coverage	0.971	0.892	0.942	0.893	0.951	0.627	0.945	0.620
α_{false}, η_{true}	bias	−0.245	0.025	−0.230	0.037	−0.268	0.009	−0.247	0.006
	variance	0.125	0.124	0.112	0.151	0.023	0.027	0.021	0.029
	coverage	0.897	0.971	0.906	0.958	0.593	0.938	0.613	0.942
α_{false}, η_{false}	bias	−0.245	−0.234	−0.251	−0.264	−0.268	−0.253	−0.251	−0.252
	variance	0.125	0.136	0.126	0.126	0.023	0.026	0.024	0.024
	coverage	0.897	0.892	0.879	0.916	0.593	0.627	0.580	0.617

#Coverage of Wald confidence intervals using the second derivative of the score equations to estimate the variance of cmles and of the mle, and $\hat{\Sigma}$ to estimate the variance of $\hat{\psi}_{eff}$.