Harvard University

Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 174

Control Function Assisted IPW Estimation with a Secondary Outcome in Case-Control Studies

Tamar Sofer[*] Marilyn C. Cornelis[†]

Peter Kraft[‡] Eric J. Tchetgen Tchetgen[**]

[*] Harvard School of Public Health, tsofer@hsph.harvard.edu

[†] Harvard School of Public Health, mcorneli@hsph.harvard.edu

[‡] Harvard School of Public Health, pkraft@hsph.harvard.edu

[**] Harvard School of Public Health, etchetge@hsph.harvard.edu

Control Function Assisted IPW Estimation with a Secondary Outcome in Case-Control Studies

Tamar Sofer[1,*], Marilyn C. Cornelis[2], Peter Kraft[1,3], and Eric J. Tchetgen Tchetgen[1,**]
[1]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, U.S.A.
[2]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, U.S.A.
[3]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, U.S.A.

July 15, 2014

Abstract

Case-control studies are designed towards studying associations between risk factors and a single, primary outcome. Information about additional, secondary outcomes is also collected, but association studies targeting such secondary outcomes should account for the case-control sampling scheme, or otherwise results may be biased. Often, one uses inverse probability weighted (IPW) estimators to estimate population effects in such studies. However, these estimators are inefficient relative to estimators that make additional assumptions about the data generating mechanism. We propose a class of estimators for the effect of risk factors on a secondary outcome in case-control studies, when the mean is modeled using either the identity or the log link. The proposed estimator combines IPW with a mean zero control function that depends explicitly on a model for the primary disease outcome. The efficient estimator in our class of estimators reduces to standard IPW when the model for the primary disease outcome is unrestricted, and is more efficient than standard IPW when the model is either parametric or semiparametric.

KEY WORDS: Case-control study; Gene association studies; Inverse probability weighting; Semiparametric inference.

1 Introduction

Case-control studies are designed to study associations between exposures and a traditionally-rare, primary outcome. Recently, genome-wide association studies (GWAS) are routinely conducted using a case-control study design, even when the primary disease outcome is relatively common, to increase power while maintaining relatively low cost. For instance, type 2 diabetes (T2D) is studied in a case-control GWAS study nested within the Nurses Health Study (NHS), and its prevalence in the cohort is estimated to be 8.4% (Cornelis et al., 2012). Such case-control studies typically collect information about additional, secondary outcomes, potentially associated with the primary disease. Specifically, Body Mass Index (BMI) measurements, which is well known to be associated with T2D, were collected in the T2D case-control study. We are interested in re-purposing the T2D GWAS data to study associations of Single Nucleotide Polymorphisms (SNPs) from the FTO gene, coding the Fat Mass and Obesity Protein, with BMI.

As Nagelkerke et al. (1995) pointed out, and others later demonstrated (Jiang et al., 2006; Richardson et al., 2007; Wang and Shete, 2011, for instance), applying standard regression methods to case-control data for analysis of a secondary outcome can bias inference, and therefore analysts need to adapt analysis schemes.

Several approaches have been proposed for the analysis of secondary outcomes from case-control studies. Nagelkerke et al. (1995) suggested that using solely the control group will be valid if it is fairly representative of the general population. This happens when the disease is rare, but may not hold otherwise. Richardson et al. (2007) and Monsees et al. (2009) discussed using Inverse Probability Weighting (IPW), in which the contribution of each subject for the estimating equation is weighted by the inverse of its selection probability into the sample. Using IPW is robust to the sampling bias, though it may be inefficient when, as typically the case in nested case-control studies, additional information can be obtained from the underlying cohort. In such settings, Augmented inverse-probability (AIPW) can be used for efficiency gain (Robins et al., 1994). However, this is potentially true for other estimators as well, and we subsequently discuss "stand-alone" case-control studies, i.e. assuming the underlying cohort from which cases and controls emanate, is not available to the analyst. Lin and Zeng (2009) proposed to estimate model parameters by maximizing the retrospective likelihood, taking into account case-control ascertainment. Li and Gail (2012) generalized their approach and suggested an adaptively weighted estimate of the association between the exposure and a binary secondary outcome, via a weighted sum of two retrospective likelihood-based estimators that differ in their assumed disease model. Chen et al. (2013) proposed a bias correction formula for an estimated odds ratio parameter, so that one can fit a regression model for the marginal or conditional analysis of the secondary outcome, and correct the estimate using the result from regressing the primary outcome on the secondary outcome and the exposure. Fewer methods are available for continuous secondary outcomes. Ghosh et al. (2013) also took a retrospective likelihood approach, extending the previous work mainly by incorporating auxiliary covariates. These likelihood based estimators rely heavily on distributional assumptions. Wei et al. (2013) modeled a continuous secondary outcome semiparametrically and relaxed the distributional assumptions, but assumed that the primary disease is rare, which does not apply in many situations, including the T2D case-control study introduced earlier. Tchetgen Tchetgen (2014) proposed a general model based on a nonparametric parameterization for the secondary outcome conditional on disease status and covariates. One can use this framework to compute an estimator under parametric, semiparametric or nonparametric models. The estimator is semiparametric locally efficient, i.e. it achieves the semiparametric efficiency bound in the absence of model misspecification, and remains consistent and asymptotically linear even if the error distribution for the outcome is incorrectly specified, provided the specified mean structure is correct. Bias may result from an incorrect mean model. The approach is developed for the identity, log and logit link functions. An implication of the parameterization proposed by Tchetgen Tchetgen (2014) is that adding a disease indicator to the regression design model (i.e., treating disease as a predictor or confounder) will bias effect estimates of the SNPs on the secondary outcome, unless either the conditional mean of the secondary outcome is almost identical in cases and in control (i.e. approximately no selection bias), or the primary disease is rare across all levels of the exposure, and the magnitude of the selection bias does not vary with covariates.

Current methodology (1) relies on distributional assumptions or (2) in the cases where fewer assumptions are made, proposed estimators are not necessarily efficient. Here, we use semiparametric theory to propose estimators for the population regression of the secondary outcome on covariates that are both robust and locally semiparametric efficient. We construct a control function in terms of a model for the primary disease risk conditional on covariates, and add it to the usual IPW estimating equation. We get a new estimating equation, which reduces to the usual IPW in the absence of any restriction on the model of disease risk given covariates. When this model is (semi)parametric, our proposed estimator is more efficient than IPW. Interestingly, we show that the parameterization proposed by Tchetgen Tchetgen (2014) is closely related to the new estimating equation. However, focusing on the identity and log links, our approach is more robust to certain forms of misspecification than the estimator of Tchetgen Tchetgen (2014). We emphasize that the proposed approach is crucially different from AIPW. Specifically, in contrast with AIPW, here only data available in the case-control sample contribute information, so that our estimators retain an IPW form. However, we also note that in a nested case-control study, one could in principle augment the estimating equations developed in this paper for additional efficiency gains using AIPW theory.

This paper is organized as follows. In Section 2 we describe the proposed class of estimators. In Section 3 we introduce semiparametric theory that forms the basis for our suggested estimators, provide the semiparametric locally efficient estimator in the class of estimators, and asymptotic properties. Throughout, we focus on the identity link (continuous outcome) and the log link (count, or positive outcome) for modeling the outcome mean. In Section 4 we present simulation results, empirically demonstrating the balance that our proposed estimators strike between robustness and efficiency, by comparing them to prevailing estimators in the literature. We use our proposed estimator in Section 5 in associating SNPs from the FTO gene with BMI, using the case-control, GWAS, T2D data set. Finally, in Section 6 we discuss our results.

2 Model

Suppose the case-control study has $i = 1, \ldots, n$ independent participants, with D_i an indicator for the primary disease, so that $D_i = 1$ if the ith participant is a case and $D_i = 0$ otherwise. Let Y_i denote the secondary outcome of interest, and $\mathbf{X}_i$ the $q \times 1$ vector of covariates of subject i. Let S_i be an indicator of inclusion in the case-control study.

We assume that the probability of selection into the study depends solely on the disease status, D_i, and is denoted by $p(S_i = 1|D_i, Y_i, \mathbf{X}_i) = \pi(D_i)$. Further, we assume that $\pi(D_i)$ is known by design. Equivalently, we assume that $p(D_i = 1)$ in the population is known. Denote by $p(\mathbf{X}_i) = p(D_i = 1|\mathbf{X}_i)$ the conditional probability of disease given covariates in the target population, and let

$$\mu(\mathbf{X}_i; \boldsymbol{\beta}) = g\{\mathbb{E}(Y_i|\mathbf{X}_i)\} \tag{1}$$

be the model for the mean after transformation using the link function $g(\cdot)$. In the case of a continuous outcome with the identity link, for instance, $\mu(\mathbf{X}_i; \boldsymbol{\beta}) = \mathbb{E}(Y_i|\mathbf{X}_i)$, and when the log link is used, $\exp\{\mu(\mathbf{X}_i; \boldsymbol{\beta})\} = \mathbb{E}(Y_i|\mathbf{X}_i)$, where expectations are taken over the entire population (rather than the case-control study population). Note that $\boldsymbol{\beta}$ is the $q \times 1$ vector of population regression coefficients that we wish to estimate. Let $\mathcal{M}$ denote the semiparametric model defined by the mean model specification (1) and the assumed model for $p(\mathbf{X})$.

Hereafter, unless otherwise stated, all expectations are taken with respect to the case control study population. Taking an estimating equations approach, parameter estimates are obtained by solving an equation of the form

$$\sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}) = 0 \tag{2}$$

for $\boldsymbol{\beta}$, where $\mathbf{U}_i(\boldsymbol{\beta})$ are $q \times 1$ functions, with $\mathbb{E}\{\mathbf{U}_i(\boldsymbol{\beta})\} = 0$, i.e. the estimating equation should be unbiased. A traditional approach for estimation in case-control studies, originating in the sample survey literature, is Inverse Probability Weighting of each equation according to its probability of selection into the study. IPW to estimate the population mean model entails solving for $\boldsymbol{\beta}$ equation (2) with

$$\mathbf{U}_{ipw,i}(\boldsymbol{\beta}) = \frac{S_i h(\mathbf{X}_i)}{\pi(D_i)}[Y_i - g^{-1}\{\mu(\mathbf{X}_i; \boldsymbol{\beta})\}], \tag{3}$$

where $h(\mathbf{X}_i)$ is a user specified $q \times 1$ function, such that $\mathbb{E}(\partial \mathbf{U}_{ipw}/\partial \boldsymbol{\beta})$ is invertible. It is straightforward to see that this equation is unbiased, using the law of iterated expectations.

Suppose that the probability of disease conditional on covariates $p(\mathbf{X})$ is known. We can use such knowledge to extend IPW estimating equations. Consider adding a general *control function* to the estimating equation to obtain:

$$\mathbf{U}_{cont}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{S_i}{\pi(D_i)}\Big(h_1(\mathbf{X}_i)[Y_i - g^{-1}\{\mu(\mathbf{X}_i; \boldsymbol{\beta})\}] - h_2(\mathbf{X}_i, D_i)\Big) = 0, \tag{4}$$

where $h_2(\mathbf{X}, D)$ is a $q \times 1$ vector control function that depends on the disease model and satisfies $\mathbb{E}\{S\,h_2(\mathbf{X}, D)/\pi(D)\big|\mathbf{X}\} = 0$. Control functions have been used in econometrics as a mean to control for bias due to specific forms of selection, see Wooldridge (2002); Petrin and Train (2010) for instance. Typically, a control function approach includes two stages of estimation, a first stage in which a subset of observed variables is employed to estimate the control function, which is subsequently used to augment a second stage regression model to identify the parameter of interest. In the present setting, we adopt a control function framework for efficiency improvement.

Note that the second term in (4) is inverse probability weighted, and has mean zero for all h_2, so that

$\mathbf{U}_{cont}(\boldsymbol{\beta})$ is unbiased. The choice $h_2(\mathbf{X}, D) = 0$ gives standard IPW. We aim to find $h_2(\mathbf{X}, D) \neq 0$ such that the resulting estimator is asymptotically at least as efficient as IPW for a fixed $h_1(\mathbf{X})$. We subsequently will characterize the optimal choice of $h_1(\mathbf{X})$.

For any choice of $h_2(\mathbf{X}, D)$, there exists a corresponding choice of $\tilde{h}_2(\mathbf{X})$ such that $h_2(\mathbf{X}, D) = \tilde{h}_2(\mathbf{X})\{D - p(\mathbf{X})\}$, that is, the set of functions $h_2(\mathbf{X}, D)$ satisfying the mean zero restriction is equivalent to the set of functions $\tilde{h}_2(\mathbf{X})\{D - p(\mathbf{X})\}$ (see Appendix, Lemma 1). Using the second parameterization, it is clear that $\mathbf{U}(\boldsymbol{\beta}, \tilde{h}_2)$ is unbiased for all $\tilde{h}_2(\mathbf{X})$. In practice, $p(\mathbf{X})$ is unknown and must be estimated. Here, we use semiparametric theory to study in a unified framework the semiparametric efficiency implications of positing a nonparametric, semiparametric or parametric model for $p(\mathbf{X})$.

3 Semiparametric theory

In this section, we develop the semiparametric framework that serves as a basis for our methods. We first provide definitions of Regular and Asymptotically Linear (RAL) estimators and tangent spaces, and provide some examples. We then characterize the RAL estimators corresponding to a given disease model $p(\mathbf{X})$, and subsequently continue to discuss inference.

3.1 Asymptotically linear estimators

An estimator $\widehat{\boldsymbol{\beta}}$ is said to be *asymptotically linear* if one can write

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = n^{-1/2} \sum_{i=1}^{n} \psi(Y_i, \mathbf{X}_i, D_i; \boldsymbol{\beta}) + o_p(1)$$

where $\psi(Y_i, \mathbf{X}_i, D_i; \boldsymbol{\beta})$ is a zero-mean function, called the ith *influence function* for $\boldsymbol{\beta}$. The asymptotic variance of the estimator $\widehat{\boldsymbol{\beta}}$ has a simple form and is given by

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \mathbb{E}\{\psi(Y, \mathbf{X}, D; \boldsymbol{\beta})\psi(Y, \mathbf{X}, D; \boldsymbol{\beta})^T\}.$$

The influence function is therefore very important, as the asymptotic distribution of an asymptotically linear estimator is completely identified from its influence function. Influence functions were first introduced by Huber (Huber, 1972) in the context of robust statistics. Later, they took on a different role in semiparametric theory, in the sense of Bickel et al. (1993). In either context, they represent the influence of a single observation on the estimator.

An *efficient* estimator is the one with associated influence function ψ^{opt} satisfying

$$\boldsymbol{a}^T[\mathbb{E}\{\psi^{opt}(Y,\mathbf{X},D;\boldsymbol{\beta})\psi^{opt}(Y,\mathbf{X},D;\boldsymbol{\beta})^T\} - \mathbb{E}\{\psi(Y,\mathbf{X},D;\boldsymbol{\beta})\psi(Y,\mathbf{X},D;\boldsymbol{\beta})^T\}]\boldsymbol{a} \geq 0$$

for all other influence functions ψ and $q \times 1$ vectors $\boldsymbol{a}$. In other words, the difference between the covariance matrices of the efficient estimator and another estimator is a positive semidefinite matrix.

In this paper we will restrict attention to the subset of asymptotically linear estimators that are also regular, i.e. that are locally uniformly consistent. Regularity is a desirable property for an estimator, since it ensures that its asymptotic behavior provides a good approximation for its finite sample behavior irrespective of the data generating mechanism within the model. For completeness, local uniform consistency is formally defined in the Appendix. A more technical definition of RAL estimators can be found in Bickel et al. (1993).

3.2 Tangent spaces

Let $\mathcal{L}_2^0$ be the space of square integrable mean zero functions. According to semiparametric theory, the *tangent space* of a parametric model is the finite dimensional linear subspace of $\mathcal{L}_2^0$, spanned by its scores. The definition of a tangent space of a parametric model generalizes to that of a semiparametric model, as the closed linear span in $\mathcal{L}_2^0$, of scores of its parametric submodels. The *nuisance tangent space* is the subspace of the tangent space containing all scores for *nuisance parameters*, i.e. any parameters indexing the observed data law, that are distinct from the parameter of scientific interest $\boldsymbol{\beta}$. We now consider the tangent space Λ_D of scores of $p(\mathbf{X})$, under parametric, semiparameteric, and nonparametric specifications. Denote the probability of disease for individuals in the case-control study, conditional on covariates, by $p_{cc}(\mathbf{X}) = p(D = 1|\mathbf{X}, S = 1)$. Let $\boldsymbol{\alpha}$ be any set of parameters indexing the probability model $p(\mathbf{X})$. In the nonparametric model in which $p(\mathbf{X})$ is unrestricted, $\Lambda_D = \Lambda_{D,npar}$, where

$$\Lambda_{D,npar} = \left\{\frac{S}{\pi(D)}h(\mathbf{X})\{D - p(\mathbf{X};\boldsymbol{\alpha})\}, \text{ for any function } h(\mathbf{X})\right\} \cap \mathcal{L}_2^0.$$

In the parametric model for $p(\mathbf{X};\boldsymbol{\alpha}) = \text{expit}(\boldsymbol{\alpha}^T\boldsymbol{x})$ with an unknown parameter $\boldsymbol{\alpha}$,

$$\Lambda_D = \Lambda_{D,\alpha} = \left\{\frac{S}{\pi(D)}\frac{p_{cc}(\mathbf{X})}{p(\mathbf{X};\boldsymbol{\alpha})}C^T\mathbf{X}\{D - p(\mathbf{X};\boldsymbol{\alpha})\}, \text{ for any conformable matrix } C\right\} \cap \mathcal{L}_2^0.$$

Consider now the semiparametric model $p(\mathbf{X};\boldsymbol{\alpha}) = \text{expit}\{\alpha_1(\boldsymbol{x}_1) + \boldsymbol{\alpha}_2^T\boldsymbol{x}_2\}$, in which the function $\alpha_1(\boldsymbol{x}_1)$ is unrestricted. Here

$$\Lambda_D = \Lambda_{D,\alpha_1,\alpha_2} = \Big\{\frac{S}{\pi(D)}\frac{p_{cc}(\mathbf{X})}{p(\mathbf{X};\boldsymbol{\alpha})}\{g(\mathbf{X}_1) + C^T\mathbf{X}_2\}\{D - p(\mathbf{X};\boldsymbol{\alpha})\},\ \text{for any conformable matrix } C \\ \text{and any function } g(\mathbf{X}_1)\Big\} \cap \mathcal{L}_2^0.$$

We show in the Appendix that the scaling factor $p_{cc}(\mathbf{X})/p(\mathbf{X})$ is required to appropriately account for retrospective sampling. In general, we will denote the tangent space of a parametric, semiparametric, or nonparametric submodel for $p(\mathbf{X})$ by $\boldsymbol{\Lambda}_{D,sub}$.

3.3 The RAL estimators for $\boldsymbol{\beta}$

Let $\Pi(\boldsymbol{v}|\Lambda)$ denote the orthogonal projection of the vector $\boldsymbol{v}$ on the subspace Λ of $\mathcal{L}_2^0$.

Theorem 1. *The set of influence function of $\boldsymbol{\beta}$ is given by*

$$\Gamma = \Big\{\frac{S}{\pi(D)}h_1(\mathbf{X})[Y - g^{-1}\{\mu(\mathbf{X},\boldsymbol{\beta})\}] - \frac{S\,h_2(\mathbf{X},D)}{\pi(D)} + \Pi\Big(\frac{S\,h_2(\mathbf{X},D)}{\pi(D)}\Big|\Lambda_{D,sub}\Big) : \\ \mathbb{E}\{\frac{S}{\pi(D)}h_2(\mathbf{X},D)|\mathbf{X}\} = 0\Big\} \cap \mathcal{L}_2^0$$

up to a multiplicative constant.

Theorem 1 characterizes all RAL estimators of $\boldsymbol{\beta}$ in a semiparametric model $\mathcal{M}$ defined by $\mu(\mathbf{X},\boldsymbol{\beta})$ and a choice of model for $p(\mathbf{X})$. The proof is in the Appendix. Interestingly, it states that if

$$\frac{S\,h_2(\mathbf{X},D)}{\pi(D)} = \Pi\Big(\frac{S\,h_2(\mathbf{X},D)}{\pi(D)}\Big|\Lambda_{D,sub}\Big),$$

then all influence functions for $\boldsymbol{\beta}$ are IPW influence functions. This equality holds, for instance, in the special case where the model $p(\mathbf{X})$ is saturated, or nonparametric. In other words, even if one uses the estimator (4), for any choice of $h_2(\mathbf{X},D)$ the asymptotic distribution of the estimator will mimic the IPW estimator and the estimator could not be made more efficient. The following Corollary 1 summarizes this observasion.

Corollary 1. *Consider the model for $\mathcal{M}$ with $p(\mathbf{X})$ unrestricted. For a fixed choice of $h_1(\mathbf{X})$ in (4), the optimal choice of function $h_2(\mathbf{X},D)$ is $h_2^{opt}(\mathbf{X},D) = 0$, and the most efficient estimator for $\boldsymbol{\beta}$ is the IPW* *estimator that solves the estimating equation*

$$\mathbf{U}_{ipw}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{S_i}{\pi(D_i)}\Big(h_1(\mathbf{X}_i)[Y_i - g^{-1}\{\mu(\mathbf{X}_i;\boldsymbol{\beta})\}]\Big) = 0.$$

In the following sections we restrict $p(\mathbf{X})$ by posing modeling assumptions. We first focus on finding the most efficient estimating equation for $\boldsymbol{\beta}$ in Γ for any fixed $h_1(\mathbf{X})$ in Section 3.4, and then provide the optimal $h_1(\mathbf{X})$ and the locally efficient estimator in Section 3.5.

3.4 Inference for a restricted model $p(\mathbf{X})$ and a fixed $h_1(\mathbf{X})$

Suppose that the model for $p(\mathbf{X})$ is restricted. For a fixed $h_1(\mathbf{X})$, we wish to find the optimal $h_2(\mathbf{X}, D)$, that minimizes the variance of the estimating equations in Γ defined in Theorem 1. This function is given in the following Theorem 2 and the proof is in the Appendix.

Theorem 2. *Suppose that $h_1(\mathbf{X})$ is fixed. The function $h_2^{opt}(\mathbf{X}, D)$ that minimizes the variance of $\widehat{\boldsymbol{\beta}}$ in model $\mathcal{M}$ is given by*

$$h_2^{opt}(\mathbf{X}, D) = h_1(\mathbf{X})[\mathbb{E}(Y|\mathbf{X}, D;\boldsymbol{\beta}) - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}].$$

Denote $\tilde{\mu}(\mathbf{X}, D;\boldsymbol{\beta}) = g\{\mathbb{E}(Y|\mathbf{X}, D;\boldsymbol{\beta})\}$, which satisfies $\mathbb{E}\{\mathbb{E}(Y|\mathbf{X}, D;\boldsymbol{\beta})\big|\mathbf{X}\} = g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}$. Then the ith influence function corresponding to $h_2^{opt}(\mathbf{X}, D)$, up to a multiplicative constant, is

$$\frac{S_i h_1(\mathbf{X}_i)}{\pi(D_i)}\left[Y_i - g^{-1}\{\tilde{\mu}(\mathbf{X}_i, D_i;\boldsymbol{\beta})\}\right].$$

Tchetgen Tchetgen (2014) provided parameterizations of $\tilde{\mu}(\mathbf{X}, D;\boldsymbol{\beta})$ in terms of $\mu(\mathbf{X};\boldsymbol{\beta})$ for the identity, log, and logit links. We use these parameterizations to construct feasible estimating equations $\mathbf{U}_{ident}^{opt}$ and $\mathbf{U}_{log}^{opt}$ based on Theorem 2. Consider first the identity link function. As was shown in Tchetgen Tchetgen (2014), $\mathbb{E}(Y|\mathbf{X}, D;\boldsymbol{\beta})$ can be parameterized as $\mathbb{E}(Y|\mathbf{X}, D;\boldsymbol{\beta}) = \mu(\mathbf{X};\boldsymbol{\beta}) + \gamma(\mathbf{X})\{D - p(\mathbf{X})\}$, where $\gamma(\mathbf{X}) = \mathbb{E}(Y|D = 1, \mathbf{X}) - \mathbb{E}(Y|D = 0, \mathbf{X})$ is the "selection bias function", resulting from sampling according to disease status.

We have that

$$\mathbf{U}^{opt}_{ident}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{S_i h_1(\mathbf{X}_i)}{\pi(D_i)} \Big[Y_i - \mu(\mathbf{X}_i;\boldsymbol{\beta}) - \gamma(\mathbf{X}_i)\{D_i - p(\mathbf{X}_i)\} \Big].$$

For the log link, it was shown in Tchetgen Tchetgen (2014) that

$$\tilde{\mu}(\mathbf{X}, D;\boldsymbol{\beta}) = \mathbb{E}(Y|\mathbf{X}, D;\boldsymbol{\beta}) = \exp\left(\mu(\mathbf{X};\boldsymbol{\beta}) + \nu(\mathbf{X}, D) - \log \mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]\right),$$

where the selection bias function $\nu(\mathbf{X}, D)$ is defined as

$$\nu(\mathbf{X}, D) = \log\left\{ \frac{\mathbb{E}(Y|\mathbf{X}, D)}{\mathbb{E}(Y|\mathbf{X}, D = 0)} \right\}$$

and reflects the log multiplicative association between D and Y given $\mathbf{X}$, and note that the expectation in $\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]$ is taken over the population. Therefore, we have that

$$\mathbf{U}^{opt}_{log}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{S_i h_1(\mathbf{X}_i)}{\pi(D_i)} \Big\{ Y_i - \exp\left(\mu(\mathbf{X}_i;\boldsymbol{\beta}) + \nu(\mathbf{X}_i, D_i) - \log \mathbb{E}[\exp\{\nu(\mathbf{X}_i, D_i)\}|\mathbf{X}_i]\right) \Big\}.$$

Note that these estimating equations are robust, in the sense that even if the selection bias functions γ and ν are misspecified, the estimating equations would remain unbiased as long as $\mu(\mathbf{X};\boldsymbol{\beta})$ and $p(\mathbf{X})$ are correctly modeled.

3.5 The semiparametric locally efficient estimator

An estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is called *locally efficient* at a submodel for $f(Y, \mathbf{X}, D, S)$ in a semiparametric model $\mathcal{M}$ if its asymptotic variance achieves the *semiparametric efficiency bound* for $\mathcal{M}$, and remains consistent and asymptotically normal (CAN) outside of the submodel (Bickel et al., 1993). For instance, in the identity link case, $\mathcal{M}$ may be a model that specifies parametric models for $p(\mathbf{X};\boldsymbol{\alpha})$, $\mu(\mathbf{X};\boldsymbol{\beta})$, and its parameters could be estimated using an estimating equation $\mathbf{U}^{opt}_{ident}(\boldsymbol{\beta}) = 0$. Different choices of $h_1(\mathbf{X})$ will lead to estimators of $\widehat{\boldsymbol{\beta}}$, that are CAN but with different asymptotic variances. The *semiparametric efficiency bound* is the smallest variance that can be obtained by a regular estimator in $\mathcal{M}$. This RAL estimator has the *efficient influence function*, which is the projection of any influence function of $\boldsymbol{\beta}$ in $\mathcal{M}$ onto the tangent space for the model (Bickel et al., 1993).

Theorem 3 below characterizes the efficient influence function of $\boldsymbol{\beta}$ in model $\mathcal{M}$ and the corresponding estimating equation, by giving the optimal $h_1^{opt}(\mathbf{X})$.

Theorem 3. *The semiparametric efficient influence function for $\boldsymbol{\beta}$ in model $\mathcal{M}$ is given by*

$$\frac{S_i h_1^{opt}(\mathbf{X}_i)}{\pi(D_i)}\Big[Y_i - g^{-1}\{\tilde{\mu}(\mathbf{X}_i, D_i;\boldsymbol{\beta})\}\Big],$$

with

$$h_1^{opt}(\mathbf{X}) = \mathbb{E}\Big\{\frac{1}{\pi(D)} var(Y|D,\mathbf{X})\Big|\mathbf{X}\Big\}^{-1}\frac{\partial}{\partial\boldsymbol{\beta}}\left[g^{-1}\{\tilde{\mu}(\mathbf{X},D;\boldsymbol{\beta})\}\right].$$

The corresponding estimator $\widehat{\boldsymbol{\beta}}$ is locally efficient in the submodel of $\mathcal{M}$ in which $h_1(\mathbf{X})$ and $h_2(\mathbf{X}, D)$ are correctly modeled. If these functions are misspecified, $\widehat{\boldsymbol{\beta}}$ will still be CAN, but less efficient. Below we use the efficient influence function to define an estimating equation by substituting empirical estimates of all unknown nuisance parameters. The asymptotic distribution of the resulting estimator allowing for model misspecification is given in Section 3.6.

3.6 Asymptotic properties

We saw that $\widehat{\boldsymbol{\beta}}$ is a RAL estimator in model $\mathcal{M}$ in which $p(\mathbf{X})$ is correctly specified. We compute $\widehat{\boldsymbol{\beta}}$ by solving the estimating equation $\widehat{\mathbf{U}}_{cont}^{opt}(\boldsymbol{\beta}) = 0$, defined as $\mathbf{U}_{cont}^{opt}(\boldsymbol{\beta})$ with $\widehat{h}_1(\mathbf{X}), \widehat{h}_2(\mathbf{X}, D)$, and $\widehat{p}(\mathbf{X})$.

Let $\boldsymbol{\delta}$ denote the parameters for the selection bias function, i.e. either $\nu(\mathbf{X}, D;\boldsymbol{\delta})$ (log link) or $\gamma(\mathbf{X};\boldsymbol{\delta})$ (identity link). Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$. It is convenient to estimate $\boldsymbol{\theta}$ jointly, by modifying the estimating equation $\mathbf{U}_{cont}^{opt}(\boldsymbol{\beta})$ to define $\mathbf{U}_{cont}^{opt}(\boldsymbol{\theta})$ by taking

$$h_1^{opt}(\mathbf{X}) = \mathbb{E}\Big\{\frac{1}{\pi(D)}\mathrm{var}(Y|D,\mathbf{X})\Big|\mathbf{X}\Big\}^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}[g^{-1}\{\mu(\mathbf{X},D;\boldsymbol{\theta})\}].$$

In the Appendix, we describe how to compute the estimator $\widehat{\boldsymbol{\theta}}$. To find its asymptotic distribution (and calculate standard errors), we need to know its influence function. Its influence function is found from the first order Taylor expansion of the estimating equation around the limiting value of $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\delta}}^T)^T$. We provide this derivation in Appendix. Let $\mathbf{V}(\boldsymbol{\alpha})$ be the estimating equation for $\boldsymbol{\alpha}$. The influence function for $\boldsymbol{\theta}$ is given by

$$\psi(\boldsymbol{\theta};\boldsymbol{\alpha}) = -\Big[\mathbb{E}\frac{\partial}{\partial\boldsymbol{\theta}}\Big\{\mathbf{U}_{cont}(\boldsymbol{\theta};\boldsymbol{\alpha})\Big\}\Big]^{-1} \times \\ \left[\mathbf{U}_{cont}(\boldsymbol{\theta};\boldsymbol{\alpha}) - \mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\alpha}}\mathbf{U}_{cont}(\boldsymbol{\theta};\boldsymbol{\alpha})\right\}\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\alpha}}\mathbf{V}(\boldsymbol{\alpha})\right\}^{-1}\mathbf{V}(\boldsymbol{\alpha})\right].$$

A consistent estimator of the covariance matrix of the estimator $\widehat{\theta}$ is given by

$$\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}_i\left(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\alpha}}\right)\widehat{\psi}_i^T\left(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\alpha}}\right),$$

where $\widehat{\psi}_i$ is the influence function evaluated at the ith subject, with all expectations in the expression $\psi(\boldsymbol{\theta};\boldsymbol{\alpha})$ estimated by the corresponding sample means.

Corollary 2. *The estimator $\widehat{\boldsymbol{\theta}}$ that solves $\mathbf{U}_{cont}(\boldsymbol{\beta},\boldsymbol{\delta};\widehat{\boldsymbol{\alpha}})$ under $\mathcal{M}$ is asymptotically normally distributed with asymptotic mean $\boldsymbol{\theta}$ and covariance*

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbb{E}\left\{\psi(\boldsymbol{\theta};\boldsymbol{\alpha})\psi(\boldsymbol{\theta};\boldsymbol{\alpha})^T\right\}.$$

Further more, in the submodel where $\widehat{h}_1^{opt}(\mathbf{X}) \rightarrow p\lim_{n\rightarrow\infty} h_1^{opt}(\mathbf{X})$, and $\widehat{h}_2^{opt}(\mathbf{X},D) \rightarrow p\lim_{n\rightarrow\infty} h_2^{opt}(\mathbf{X},D)$, $\widehat{\boldsymbol{\beta}}$ is locally efficient.

Note that $\widehat{\boldsymbol{\theta}}$ will be asymptotically normal with covariance matrix $\mathbb{E}\left\{\psi(\boldsymbol{\theta}^*;\widehat{\boldsymbol{\alpha}})\psi(\boldsymbol{\theta}^*;\widehat{\boldsymbol{\alpha}})^T\right\}$, where $\boldsymbol{\theta}^*$ is $p\lim_{n\rightarrow\infty}\widehat{\boldsymbol{\theta}}$, even if one of $p(\mathbf{X}), \mu(\mathbf{X};\boldsymbol{\beta})$, or both, are misspecified. In the case of misspecification, $\boldsymbol{\theta}^*$ is likely a biased estimate of the true $\boldsymbol{\theta}$.

4 Simulations

In this section, we demonstrate the robustness and efficiency of our proposed estimators compared to the prevailing estimators, when modeling the mean via the identity link. We simulate case-control studies with continuous secondary outcomes in two sets of simulations. The goal of the first set was to investigate the robustness and efficiency of the proposed control function estimator (‘cont’), compared to multiple other prevailing estimators: the estimator that conditions on disease status, using disease indicator in the regression of the secondary outcome on covariates, (denoted by Dind), the estimator that treats all observations equally, ignoring disease status (pooled), and the IPW estimator (IPW). The goal of the second set was to compare the performance of 'cont' to the estimators proposed by Ghosh et al. (2013) and Lin and Zeng (2009). In each section below, we describe the simulations and provide results, where for cont, we provide two sets of results: when the model for $\mathbb{E}(Y|\mathbf{X}, D)$ is correctly specified, and when it is misspecified. For each scenario, we calculated the mean bias of the estimates $\frac{1}{n.sim}\sum_{k=1}^{n.sim}\widehat{\beta}_k - \beta$, the mean squared error (MSE) $\frac{1}{n.sim}\sum_{k=1}^{n.sim}(\widehat{\beta}_k - \beta)^2$, the sample standard deviation of the estimator $\{\frac{1}{n.sim}\sum_{k=1}^{n.sim}(\widehat{\beta}_k - \bar{\widehat{\beta}})^2\}^{1/2}$, the mean of the estimated standard deviations in the simulations $\frac{1}{n.sim}\sum_{k=1}^{n.sim}\widehat{\mathrm{sd}(\beta_k)}$, and the Wald coverage probability. Due to limited space, only some of the simulation results are presented in the main manuscript. Additional extensive simulation results are delegated to the Appendix, including all summaries pertaining to the performance of the estimator "Dind" and "pooled".

The proposed cont estimators and the estimated standard deviations were calculated as described in the Appendix. The IPW estimator and the estimated standard deviations were calculated using Newton-Raphson iterations of the estimating function $\mathbf{U}_{ipw}$, with $h_1(\mathbf{X}_i) = \mathbf{X}_i$, with the robust (sandwich) covariance matrix. The naïve estimators Dind and pooled were calculated from linear regression.

All simulation scenarios included 500 cases and 500 controls, and were run 1000 times. The prevalence of the disease D in the population (the primary case-control outcome) was fixed at 0.12, i.e. the disease is relatively common.

We conducted other simulation studies, under a variety of plausible scenarios. First, we performed a simulation study for the identity link with a single exposure variable, in which we also considered the estimator proposed by Tchetgen Tchetgen (2014). Second, we performed simulations for the log link, and lastly, we carried out another identity link simulation study, closely mimicking the observed data distribution in the T2D sample. Results for these additional scenarios are provided in the Appendix. In general, they support the conclusions of the simulations presented here.

4.1 Simulation set 1 - studying robustness and efficiency

To design the simulations, we first note that we need to sample data from the distribution $f(Y, D|\mathbf{X})$, in such a way that the parameter of interest $\mathbb{E}[Y|\mathbf{X}]$ is defined. We consider the decomposition $f(Y, D|\mathbf{X}) = f(Y|D, \mathbf{X})p(D|\mathbf{X})$, and generate the data according to the two parts of the likelihood, $p(D|\mathbf{X})$, and $f(Y|D, \mathbf{X})$. Note that this decomposition always holds, and makes no assumption on the underlying model. We use the nonparametric decomposition of $\mathbb{E}[Y|\mathbf{X}, D]$ proposed by Tchetgen Tchetgen (2014), that allows specifying the two parts of the likelihood. First, exposures/covariates variables $\mathbf{X}$ were sampled. Then, disease probabilities were calculated for each subject, based on exposure values. The intercept for the disease model $p(\mathbf{X})$ was set so that disease prevalence was 0.12. Disease statuses were obtained from disease probabilities, and the secondary outcomes Y were generated based on exposure values and disease status.

In more details, we simulated two covariates, X_1 and X_2 where $X_1 \sim \mathcal{N}(2, 4)$, and $X_2 \sim$ Binary(0.1). The primary disease probability was calculated by

$$\text{logit}\{p(D = 1|\mathbf{X})\} = -3.2 + 0.3X_1 + X_2,$$

and disease status was sampled. The conditional mean of the secondary outcome was:

$$\mathbb{E}(Y|\mathbf{X}, D) = 50 + 4X_1 + 3X_2 + 3X_1X_2 + \{D - p(\mathbf{X})\}(3 + 2X_1 + 2X_2 + 2X_1X_2),$$

so that $\mu(\mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}^T\boldsymbol{\beta}$ with $\mathbf{X} = (1, X_1, X_2, X_1X_2)^T$ and $\boldsymbol{\beta} = (50, 4, 3, 3)^T$, and $\gamma(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\alpha}$ with $\boldsymbol{\alpha} = (3, 2, 2, 2)^T$. The residuals were sampled by $\epsilon \sim \mathcal{N}(0, 4)$. The design matrix for $\gamma(\mathbf{X})$ was $\mathbf{X} = (1, X_1, X_2, X_1X_2)^T$ when the model was correctly specified. We studied the following forms of misspecification of the design matrix of $\gamma(\mathbf{X})$. The estimator 'cont-mis1' had the design matrix $\mathbf{X} = (1, X_1, X_2)^T$ (no interaction term), 'cont-mis2' had $\mathbf{X} = (1, X_1)^T$, 'cont-mis3' had $\mathbf{X} = (1, X_2)^T$, and 'cont-mis4' accounted only for an intercept, i.e. design matrix $\mathbf{X} = 1$.

A table providing comprehensive simulation results for this simulation is provided in the Appendix. Figure 1 compares between the estimated bias, MSE, and coverage probabilities of the cont estimators (correctly specified and misspecified), and the usual IPW. We do not provide graphic results for the estimators that where heavily biased: Dind, and pooled.

As shown in Figure 1 and in the Appendix, the estimated bias of the cont estimator is usually smaller than that of the IPW estimator and is very small when the model for $\gamma(\mathbf{X})$ is correctly specified, and slightly larger when the model for $\gamma(\mathbf{X})$ is misspecified.

Both the MSE and the empirical standard deviation of the cont estimator were higher when the model for $\gamma(\mathbf{X})$ was misspecified, yet interestingly, the MSE of cont was always smaller than that of the IPW. In fact, the relative efficiency of cont-cor was from 17% (β_2) to 71% (β_3) lower than that of the IPW. Even cont-

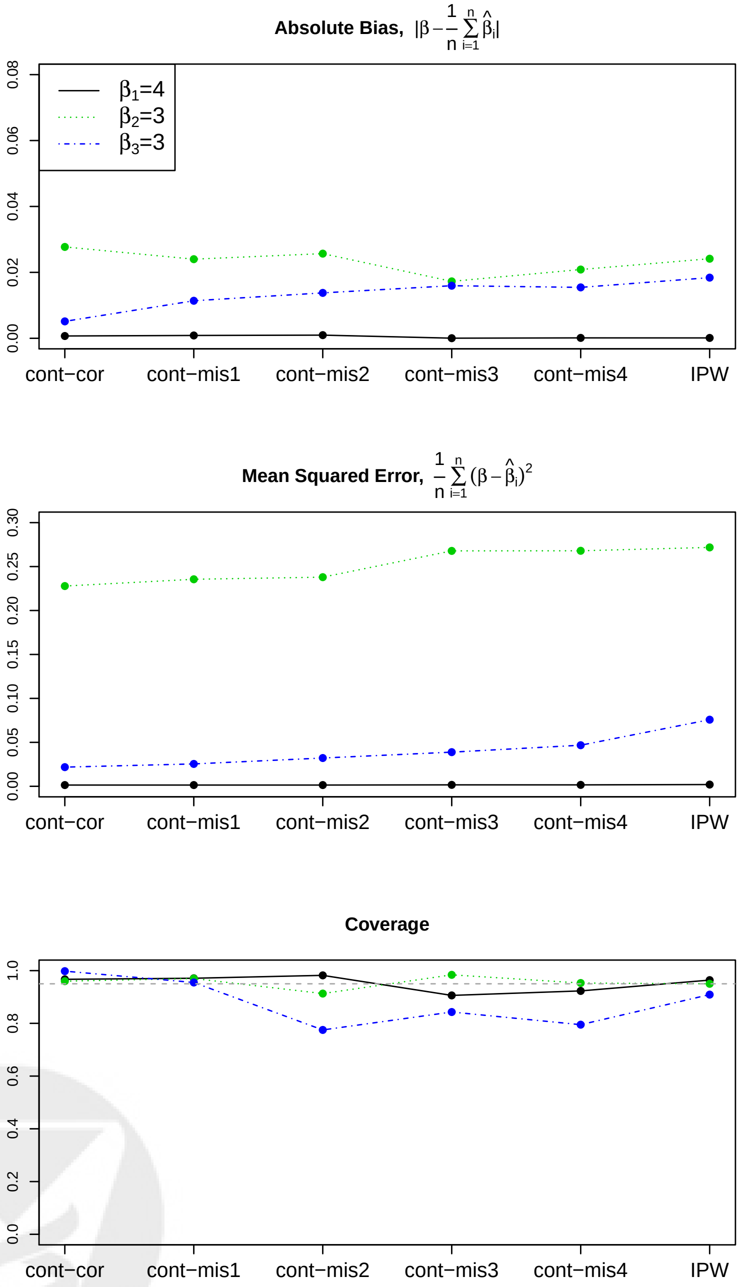


Figure 1: Results from Identity link simulations set 2, second settings, two covariates. Estimated bias, MSE, and coverage probability of the control function under correct and misspecification of the selection bias function (cont-cor, cont-mis1, . . ., cont-mis4), and IPW, in estimating population effects of X_1, X_2 and their interaction.

mis4, the estimator with only intercept used in the selection bias function, had relative efficiency from 1.5% to 38% lower than that of the IPW. Coverage probabilities were always correct for the coefficients of X_1 and X_2, but sometimes too small for the coefficient of the interaction term X_1X_2 when $\gamma(\mathbf{X})$ was misspecified (more specifically, when X_1 was not in the design matrix of $\gamma(\mathbf{X})$). In comparison, the coverage probability of the IPW estimator was always very close to the desired 95%. Dind and the pooled estimator again yielded biased estimates with, usually, very low coverage probability. Interestingly, the bias of pooled was usually lower than the bias of Dind.

4.2 Simulation set 2 - comparison to another recently proposed method

Here we compare our estimator 'cont', and the IPW, to the pseudo-likelihood estimator proposed by Ghosh et al. (2013), and the retrospective likelihood estimator proposed by Lin and Zeng (2009). We followed the simulation scenario performed in Ghosh et al. (2013), using code shared by the authors. We also adapted our simulations from Section 4.1 to their assumed data structure.

First, we ran 1000 simulations in Ghosh et al. (2013) simulation settings and compared the estimators. In their simulations, they focused on a single coefficient, namely the effect of a single nucleotide polymorphism (SNP) G on the outcome Y. G had a minor allele frequency (MAF) 0.25. There were two covariates Z, one continuous and one binary, with probability 0.45. The disease and the secondary outcome were modeled by a bivariate normal distribution and thresholding, so that the disease model is dependent on G and $\boldsymbol{Z}$ via a logistic model. However, it is unclear how to specify correctly $\gamma(\mathbf{X})$. We use a linear model of the form $\gamma(\mathbf{X}) = \mathbf{X}\delta$, though this is likely incorrect. The outcome Y had variance 1, and disease prevalence was 0.05. The effect of interest was 0.1. We used 500 cases and 500 controls. More details can be found in Ghosh et al. (2013). The results of these simulations are presented at the top part of Table 1.

Then, we ran 1000 simulations in settings adapted from our simulations from Section 4.1. Here, we had the same $G, \boldsymbol{Z}$ variables, with Z_1 continuous and Z_2 binary. $Z_1 \sim \mathcal{N}(0, 4)$, and $Z_1 \sim \text{Binary}(0.2)$. The primary disease probability was calculated by

$$\text{logit}\{p(D = 1|\mathbf{X})\} = -3.8 + 0.3X_1 + X_2,$$

and disease status was sampled. Note that the intercept value was selected to that disease prevalence was roughly 0.05, as in Ghosh et al. (2013). The SNP G has minor allele frequency 0.3. The conditional mean model was:

$$\mathbb{E}(Y|\mathbf{X}, D) = 3 + 0.7Z_1 + 0.5Z_2 + 0.3G + \{D - p(\mathbf{X})\}(1 + 0.5Z_1 + 0.3Z_2).$$

500 cases and 500 controls were sampled from the simulated population. We compared the estimation of the effect of G on Y. The results of these simulations are presented at the bottom part of Table 1.

In the first simulation set, the estimators Ghosh2013, IPW and cont were unbiased, and achieved the nominal coverage level, while Lin2009 was heavily biased. Note that cont likely misspecified the model $\gamma(\mathbf{X})$. The estimator of Ghosh et al. (2013) had slightly lower MSE than the IPW and control function estimators, as expected, since this estimator is based on the same model used to produce the simulated data. In the second set of simulations, in which the data were sampled by specifying models for $p(\mathbf{X}), \gamma(\mathbf{X})$, and $\mu(\mathbf{X}; \boldsymbol{\beta})$, both estimators Ghosh2013 and Lin2009 were biased (both biases about -0.4) and had low coverage of 0.67% (Ghosh2013) and 50% (Lin2009). In contrast, the IPW and cont estimators performed similarly, with low bias and correct coverage probability. These results demonstrate the robustness of IPW and cont, which use fewer modeling assumptions.

The estimator Lin2009 was biased under both simulation scenarios. As other likelihood-based estimators, it assumes a certain probability model, and it is biased when this model is fails to hold. The model proposed by Ghosh et al. (2013) is different than the one proposed by Lin and Zeng (2009).

5 Analysis of Type 2 diabetes GWAS

We analyzed the case-control GWAS study of T2D, with the goal of identifying SNPs in the FTO gene region, associated with BMI. There were 3080 female participants in this data, genotyped on the affymetrix 6.0 array, with 1326 cases and 1754 controls (Cornelis et al., 2012). There were 152 genotyped SNPs from the region on chromosome 16 spanning the FTO variants. There are a few SNPs from the FTO gene associated with BMI (Speliotes et al., 2010), and validated on large cohorts. In particular, the SNP rs1558902 has the strongest association with log-BMI. This SNP is not in the data, but other SNPs in high Linkage Disequilabrium (LD) with it are. The population prevalence of T2D was 8.4% (Cornelis et al., 2012). We compared the usual IPW, the control function estimator ‘cont’, the pooled estimator ignoring disease status, the estimator Dind with disease indicator in the design matrix, and the estimator of Lin and Zeng (2009) dubbed Lin2009.

All analyses were adjusted to age, binary smoking status (current versus past or never), binary alcohol

Table 1: Simulation results for estimating the effect of a SNP on a normally distributed secondary outcome. We compare results for the usual IPW estimator, the proposed control function estimator (‘cont’), the pseudo-likelihood estimator of Ghosh et al. (2013) (‘Ghosh 2013), and the retrospective likelihood estimator of Lin and Zeng (2009) (‘Lin2009’). The top part of the table provides results of simulations in the settings in Ghosh et al. (2013), and the bottom part is of simulations designed according to the conditional mean model $\mathbb{E}(Y|\mathbf{X}, D)$. For each estimator and each estimated parameter the table reports the estimator’s mean bias, MSE, empirical standard deviation over all simulations, mean estimated standard deviation using the appropriate formula, and coverage probability.

estimator/value	bias	MSE	emp sd	est sd	coverage
Settings 1 (Ghosh, 2013). $\beta = 0.1$					
Ghosh2013	0.000	0.003	0.056	0.057	0.961
cont	0.000	0.004	0.067	0.067	0.952
IPW	0.000	0.004	0.067	0.067	0.953
Lin2009	-0.765	0.588	0.049	0.055	0.000
Settings 2. $\beta = 0.7$					
Ghosh2013	−0.402	0.228	0.258	0.261	0.666
cont	0.002	0.002	0.043	0.042	0.946
IPW	0.002	0.002	0.043	0.042	0.948
Lin2009	-0.394	0.194	0.197	0.200	0.505

intake measure according to less or more than 10 grams a day, physical activity (above or under the median) and to the first four principal components of the genetic data. The outcome, BMI, was log transformed, as is usually done with BMI. For the analysis using the estimator cont, all models, i.e. the mean model of BMI, the model for disease probability $p(D = 1|\mathbf{X})$, and the selection bias model $\gamma(\mathbf{X})$ used the same covariates. All SNPs were analyzed in the additive mode of inheritance.

Figure 2 compares between the estimated effect sizes and their respective standard errors (SEs), of all 152 SNPs in the FTO gene, between cont, and the other estimators under consideration. The cont estimator yielded roughly identical results to that of the IPW. This is in agreement with the simulation study imitating the effect sizes in the T2D data set (see Appendix), and is expected since both T2D and BMI are complex traits, and no single SNP highly affects them. Thus, incorporating the disease and selection bias models in the estimation cannot improve it much. Although the standard errors appear to be “the same” when looking at the plot, in fact that are small differences, such that the p-values and adjusted p-values of the cont estimates are smaller than those of the IPW, as is seen in Table 2. Effect estimates of other estimators are quite different than those of cont, while their SEs are usually smaller. That is since these estimators make more assumptions on the data distribution, resulting in lower SEs.

There were ten SNPs with Holm's adjusted p-value ≤0.05 by the pooled estimator, which yielded the lowest p-values. As they were all in high LD, we selected the SNP that is in highest LD with rs1558902, namely, rs1421085 (Johnson et al., 2008). Table 2 compares between the various analyses results on this SNP. As the effects are relatively low (~ -0.02), all estimates are within a rang of 0.04 of each other. Consistent with the plot, cont and IPW gave identical effect estimates (after rounding) while other estimates were usually different. The effect estimate is largest (in absolute value) in the pooled estimator. Since pooled and Dind are likely biased estimators (as supported by the simulations mimicking the T2D diabetes, reported in the Appendix), we now consider Lin2009. This estimator properly accounts for case-control sampling, but assumes that the outcome is normally distributed around the population mean. To study the appropriateness of this assumption, we compared the density of the residuals of log-BMI after removing the population mean estimated by IPW. Figure 3 provides this comparison, suggesting that the normality assumption does not hold and that the estimator is potentially biased.

Table 2: Effect estimates, and their respective SEs and p-values for the SNP rs1421085 from the FTO gene. The values were obtained by the control function estimator ('cont'), the usual IPW, the pooled estimator ignoring disease status, and the disease with disease indicator in the design matrix ('Dind').

Estimator	effect	SE	p-value (raw)	p-value (adj)
		rs1421085		
cont	−0.017	0.0054	1.7e-3	0.247
IPW	−0.017	0.0054	1.9e-3	0.273
pooled	−0.021	0.0050	4.2e-5	0.006
Dind	−0.018	0.0046	9.3e-5	0.014
Lin2009	−0.019	0.0047	4.7e-5	0.007

6 Discussion

In this work we provide semiparametric, efficient and robust estimators for the population mean effects of covariates on secondary outcomes in case-control studies. The main idea behind the proposed estimators is the addition of an inverse probability weighted, control function that preserves unbiasedness of the estimating equation when a model for disease probability given covariates is correctly specified. Additional required assumptions are correct specification of the population mean model and known sampling fractions for the case-control study. No other distributional assumptions are made about the outcome. We propose

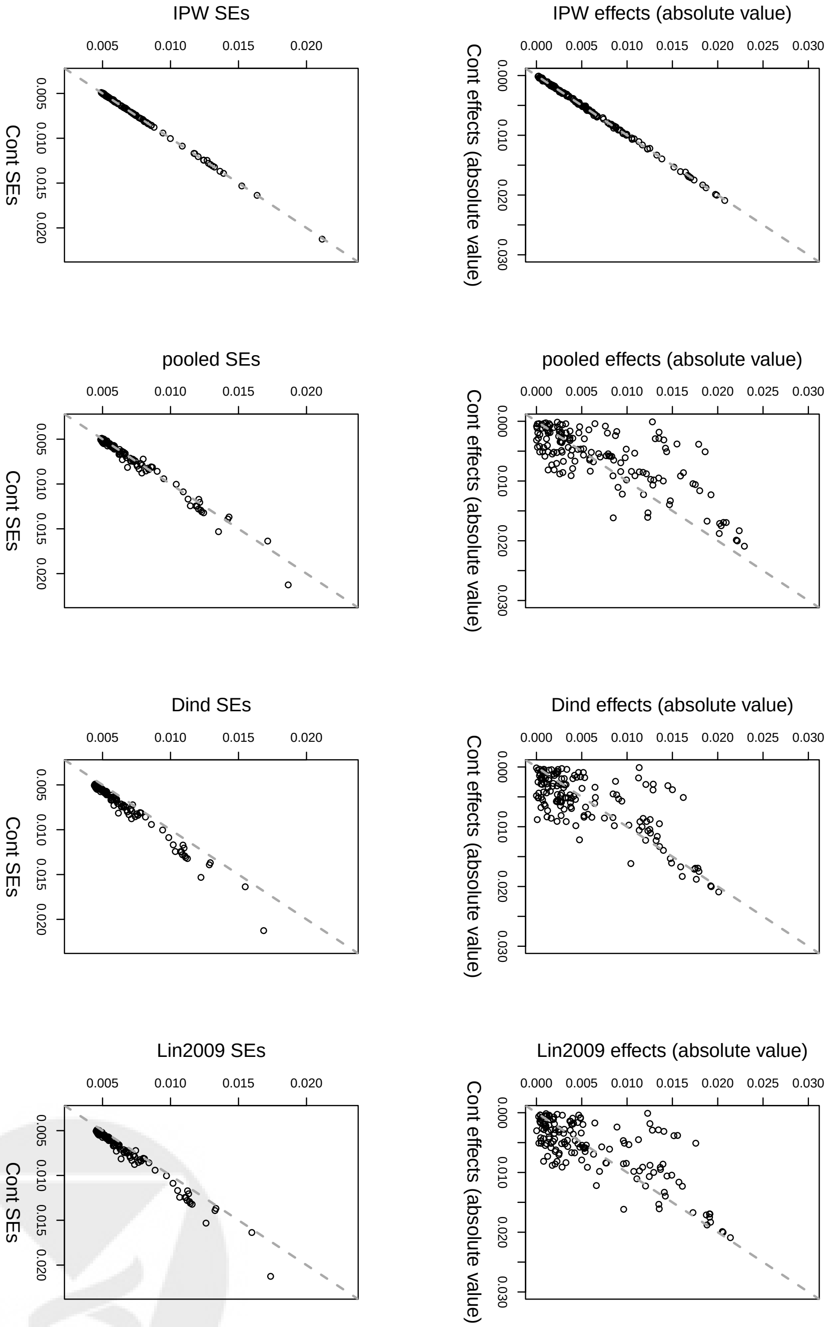


Figure 2: Comparison of effect estimates for the SNPs in the FTO gene on log-BMI, and their standard errors. Estimates of the control function estimator (‘cont’) and their SE were compared to the usual IPW, the estimator ignoring disease status (pooled), the estimator using disease indicator in its design matrix (‘Dind’) and the estimator of Lin and Zeng (2009) (‘Lin2009’). Every point in the plot represent a SNP. If a point falls on the diagonal - its associated effect (SE) estimate is equal in cont and the compared estimator. If it falls below the diagonal, its estimated effect (SE) is smaller in cont compared to the other estimator.

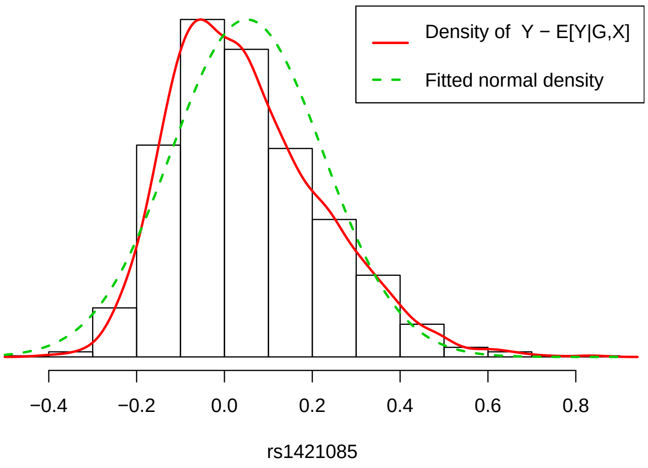


Figure 3: Histogram, and overlaid empirical and fitted normal densities to the residuals of log-BMI after removing estimated population mean.

estimators that may be used with identity and log links. This approach could potentially extend to the logit link, which presents a challenge for future research.

The control function estimator is unbiased under correct specification of the disease model given covariates, even if the model for the selection bias function is misspecified. We recommend evaluating the disease model fit with respect to the model predictions (estimated disease probabilities). One can use Area Under the operating Curve (AUC) and cross validation as measures that give indications of fit due to good or poor prediction. For a comprehensive review of such methods see Harrell et al. (1996). It is also useful to compare the control function effect estimate to the IPW, as the IPW is robust to misspecification of the disease model. Under correct specification of the disease model we expect to see similar effect estimates for both IPW and control function estimators, with smaller standard errors for the later.

Robustness, as is here attributed to the IPW and control function estimator, is somewhat different than the robustness property in robust statistics. Operating in the semiparametric theory framework, we carefully define the modeling assumptions required to produce unbiased estimators. For instance, correct specification of the population mean, and of the disease model. Robustness here refers to the fact that our estimator will be consistent for all data generating mechanisms in which these assumption hold (e.g. normal errors, but also t-distributed errors, or even non-symmetric errors). In comparison, in robust statistics the focus is on estimators that are protected against outliers, which is (conceptually) more oriented towards smaller sample sizes.

In recent work, especially that relying on the retrospective likelihood (Lin and Zeng, 2009; Li and Gail, 2012; Chen et al., 2013; Ghosh et al., 2013), the primary disease probability is modeled in a logistic regression, with both the exposure and the secondary outcome, and sometimes their interaction, as predictors. Our formulation does not explicitly use the secondary outcome in the disease model. However, efficient control function estimator incorporates a selection bias function, which encodes the association between the secondary outcome and the case control status conditional on covariates. Hence, as in any likelihood based approach, this association is accounted for, while more general specifications of this association are readily applied. Although the control function estimator is far more general and relies on fewer assumptions, it is guaranteed to be most efficient if all models are correctly specified. We also note that in many settings, the secondary outcome may occur on the causal pathway between the exposure and the primary outcome (e.g. mammographic density and breast cancer, or smoking and lung cancer) in which case the model for the D adjusting for X and Y is difficult to interpret.

Acknowledgements

This research was supported by NIH grants R21ES019712 and R01ES020337.

References

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore.

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.

Cornelis, M. C., Tchetgen Tchetgen, E. J., Liang, L., Qi, L., Chatterjee, N., Hu, F. B. and Kraft, P. (2012). Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *American Journal of Epidemiology* **175** 191–202.

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

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15** 361–387.

URL http://dx.doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Huber, P. J. (1972). The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics* **43** 1041–1067.

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

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J. and de Bakker, P. I. (2008). Snap: a web-based tool for identification and annotation of proxy snps using hapmap. *Bioinformatics* **24** 2938–2939.

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

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

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

Nagelkerke, N. J., 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** 769–775.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research* **47** 3–13.

Richardson, D. B., Rzehak, P., Klenk, J. and Weiland, S. K. (2007). Analyses of casecontrol 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.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R. et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42** 937–948.

Tchetgen Tchetgen, E. J. (2014). A general regression framework for a secondary outcome in case-control studies. *Biostatistics* **15** 117–128.

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.

Wei, J., Carroll, R. J., Müller, U. U., Van Keilegom, I. and Chatterjee, N. (2013). Robust estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of the Royal Statistical Society: Series B (Methodological)* **75** 185–206.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. The MIT press, Cambridge Massachusetts.

Appendix

7 What are regular estimators?

According to the definition in **?**, it is assumed that there is a true distribution generating the data, indexed by a parameter θ. In practice, a sampled data set is distributed according to θ_n where θ_n is $\sqrt{n}$-consistent for the true θ. This process by which data are sampled from a $\sqrt{n}$ perturbation of the truth is called a local data generating process. Regularity, or "local uniform consistency", means that the estimator for $\boldsymbol{\beta}$ (some parameter of the distribution indexed by θ) does not depend on the local data generating process.

8 Mathematical derivations.

8.1 The tangent space of a model for $p(\mathbf{X})$ in a case-control study

We here show that the tangent space, or the collection of scores, for $p(\mathbf{X})$ (disease probability in the general population) in a case-control study is related to the tangent space for $p_{cc}(\mathbf{X})$ (disease probability in the case-control study population) via the "scaling factor" $p_{cc}(\mathbf{X})/p(\mathbf{X})$. Or in other words, a score for $p_{cc}(\mathbf{X})$ evaluated in a case-control study is multiplied by this scaling factor to obtain a score for $p(\mathbf{X})$. A general score for the disease probability $p_{cc}(\mathbf{X})$ in the case control study is given by:

$$Sh(\mathbf{X})\{D - p_{cc}(\mathbf{X})\}.$$

Recall the identity

$$\begin{aligned} \text{logit}p(\mathbf{X}) &= \text{logit}p_{cc}(\mathbf{X}) + \log\left[\frac{p(D=1)\{1-p(D=1|S=1)\}}{p(D=1|S=1)\{1-p(D=1)\}}\right] \\ \Downarrow & \\ \frac{p(\mathbf{X})}{1-p(\mathbf{X})} &= \frac{p_{cc}(\mathbf{X})}{1-p_{cc}(\mathbf{X})}\left[\frac{p(D=1)\{1-p(D=1|S=1)\}}{p(D=1|S=1)\{1-p(D=1)\}}\right]. \end{aligned}$$

We make a few transformations in order to write this score in terms of $\{D - p(\mathbf{X})\}$:

$$\begin{aligned} Sh(\mathbf{X})\{D - p_{cc}(\mathbf{X})\} &= Sh(\mathbf{X})\frac{\{D-p_{cc}(\mathbf{X})\}}{p_{cc}(\mathbf{X})\{1-p_{cc}(\mathbf{X})\}}p_{cc}(\mathbf{X})\{1-p_{cc}(\mathbf{X})\} \\ &= Sh(\mathbf{X})\frac{(-1)^{1-D}}{p_{cc}^{D}(\mathbf{X})\{1-p_{cc}(\mathbf{X})\}^{1-D}}p_{cc}(\mathbf{X})\{1-p_{cc}(\mathbf{X})\} \\ &= Sh(\mathbf{X})\frac{(-1)^{1-D}p_{cc}(\mathbf{X})}{\left\{\frac{p_{cc}(\mathbf{X})}{1-p_{cc}(\mathbf{X})}\right\}^{D}} \\ &= Sh(\mathbf{X})\frac{(-1)^{1-D}p_{cc}(\mathbf{X})}{\left(\left[\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\right]\left[\frac{p(D=1|S=1)\{1-p(D=1)\}}{p(D=1)\{1-p(D=1|S=1)\}}\right]\right)^{D}} \end{aligned}$$

Since:

$$\frac{(-1)^{(1-D)}}{p(\mathbf{X})^{D-1}\{1-p(\mathbf{X})\}^{-D}} = D - p(\mathbf{X}),$$

we get:

$$
\begin{aligned}
S\,h(\mathbf{X})\{D - p_{cc}(\mathbf{X})\} &= S\,h(\mathbf{X})\frac{p_{cc}(\mathbf{X})}{p(\mathbf{X})}\{D - p(\mathbf{X})\}\left\{\frac{p(D=1)p(D=0|S=1)}{p(D=1|S=1)p(D=0)}\right\}^{D} \\
\propto \quad S\,h(\mathbf{X})\frac{p_{cc}(\mathbf{X})}{p(\mathbf{X})}\{D - p(\mathbf{X})\}&\left\{\frac{p(D=1)p(D=0|S=1)}{p(D=1|S=1)p(D=0)}\right\}^{D} \cdot \frac{p(D=0)}{p(D=0|S=1)p(S=1)} \\
&= \frac{S}{p(S=1|D)}\frac{p_{cc}(\mathbf{X})}{p(\mathbf{X})}h(\mathbf{X})\{D - p(\mathbf{X})\}.
\end{aligned}
$$

As required.

In the main manuscript, we showed that scores of the tangent space in the nonparametric model are

$$\frac{S}{p(S=1|D)}h(\mathbf{X})\{D - p(\mathbf{X})\},$$

and this holds since the function $h(\mathbf{X})$ can be written as $\tilde{h}(\mathbf{X})p_{cc}(\mathbf{X})/p(\mathbf{X})$ for some $\tilde{h}(\mathbf{X}) = h(\mathbf{X})p(\mathbf{X})/p_{cc}(\mathbf{X})$. However, in the parametric and nonparametric cases, $C^{T}\mathbf{X} \neq p_{cc}(\mathbf{X})/p(\mathbf{X})\tilde{C}^{T}\mathbf{X}$, since $p_{cc}(\mathbf{X})/p(\mathbf{X})$ is not fixed.

8.2 Proof of Theorem 1

Before approaching this proof, Lemma 1 provides the form of $h_2(\mathbf{X}, D)$ in each of the link functions under consideration.

Lemma 1

(a) Any function $h_2(\mathbf{X}, D)$ such that $\mathbb{E}\{h(\mathbf{X}, D)|\mathbf{X}\} = 0$, where the expectation is taken in the general population, can be written as

$$h_2(\mathbf{X}, D) = \gamma(\mathbf{X})\{D - p(\mathbf{X})\},$$

for any function $\gamma(\mathbf{X})$. This parametrization will be used in the linear link ca se.

(b) $h_2(\mathbf{X}, D)$ can equivalently be written in the form

$$h_2(\mathbf{X}, D) = h(\mathbf{X})[1 - \exp(\nu(\mathbf{X}, D) - log\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}])],$$

where $h(\mathbf{X})$ is any function of $\mathbf{X}$, and $\nu(\mathbf{X}, D)$ is such that $\nu(\mathbf{X}, 0) = 0$. We will use this parametrization in the log link case.

Proof of Lemma 1

1. Define the two sets $\mathcal{A}_1 = \{h_2(\mathbf{X}, D) : \mathbb{E}\{h(\mathbf{X}, D)|\mathbf{X}\} = 0\}$ (where the expectation is taken in the general population) and $\mathcal{A}_2 = \{\gamma(\mathbf{X})\{D - p(\mathbf{X})\} : \gamma(\mathbf{X})$ any function of $\mathbf{X}\}$. We show that the two sets are equal. The first direction, $\mathcal{A}_2 \subseteq \mathcal{A}_1$ is trivial, by noting that $\mathbb{E}(D|\mathbf{X}) = p(\mathbf{X})$. To show that $\mathcal{A}_1 \subseteq \mathcal{A}_2$, let $h_2(\mathbf{X}, D)$ be an element of $\mathcal{A}_1$. We show that it is also an element of $\mathcal{A}_2$. Choose $\gamma(\mathbf{X}) = h_2(\mathbf{X}, 1) - h_2(\mathbf{X}, 0)$. Then we can verify that for this choice of $\gamma(\mathbf{X})$, indeed $h_2(\mathbf{X}, D) = \gamma(\mathbf{X})\{D - p(\mathbf{X})\} = \{h_2(\mathbf{X}, 1) - h_2(\mathbf{X}, 0)\}\{D - p(\mathbf{X})\}$.

   For $D = 1$, we have that $h_1(\mathbf{X}, 1) = \gamma(\mathbf{X})\{1 - p(\mathbf{X})\}$ yields $h_2(\mathbf{X}, 0) = -\{h_2(\mathbf{X}, 1) - h_2(\mathbf{X}, 0)\}p(\mathbf{X})$, and for $D = 0$, $h_1(\mathbf{X}, 0) = \gamma(\mathbf{X})\{0 - p(\mathbf{X})\}$ also gives $h_2(\mathbf{X}, 0) = -\{h_2(\mathbf{X}, 1) - h_2(\mathbf{X}, 0)\}p(\mathbf{X})$. This equality is true: $\mathbb{E}\{h_2(\mathbf{X}, D)|\mathbf{X}\} = 0 = h_2(\mathbf{X}, 0)\{1 - p(\mathbf{X})\} + h_2(\mathbf{X}, 1)p(\mathbf{X})$.

2. First, rewrite

$$h(\mathbf{X})\left\{1 - \exp\left(\nu(\mathbf{X}, D) - \log\mathbb{E}\left[\exp\{\nu(\mathbf{X}, D)\}\middle|\mathbf{X}\right]\right)\right\} = \frac{h(\mathbf{X})}{\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]}\left(\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}] - \exp\{\nu(\mathbf{X}, D)\}\right).$$

   We show that

$$\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}] - \exp\{\nu(\mathbf{X}, D)\} = \{p(\mathbf{X}) - D\}[\exp\{\nu(\mathbf{X}, 1)\} - \exp\{\nu(\mathbf{X}, 0)\}],$$

   and therefore $\gamma(\mathbf{X}) = h(\mathbf{X})/(\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}][\exp\{\nu(\mathbf{X}, 1)\} - \exp\{\nu(\mathbf{X}, 0)\}])$. To show the required equality, notice that since D is binary:

$$\exp\{\nu(\mathbf{X}, D)\} = D[\exp\{\nu(\mathbf{X}, 1)\} - \exp\{\nu(\mathbf{X}, 0)\}] + \exp\{\nu(\mathbf{X}, 0)\}.$$

   Writing $\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]$ using simple algebra, the results follows. ■

Proof of the theorem. Recall

$$\mathbf{U}_{cont}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{S_i}{\pi(D_i)}\Big(h_1(\mathbf{X}_i)[Y_i - g^{-1}\{\mu(\mathbf{X}_i; \boldsymbol{\beta})\}] - h_2(\mathbf{X}_i, D_i)\Big).$$

Consider the parametric submodel $f_t(O) = f_t(Y|D, S = 1, \mathbf{X})f(S = 1|D)f_t(D|\mathbf{X})f_t(\mathbf{X})$, where $f_{t=0}(O) = f(O)$

is the true law. Denote by $\mathbb{S}^{sub}(O) = \mathbb{S}^{sub}(Y|D, S = 1, \mathbf{X}) + \mathbb{S}^{sub}(D|\mathbf{X}) + \mathbb{S}^{sub}(\mathbf{X})$ the scores in the submodel (e.g. $\mathbb{S}^{sub}(O) = \partial/\partial t \log\{f_t(O)\}$, etc.)

Let:

$$\Psi_t(\boldsymbol{\beta}, h_1, h_2) = \mathbb{E}_t\Big\{\frac{S}{\pi(D)}\Big(h_1(\mathbf{X})[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - h_2(\mathbf{X}, D)\Big)\Big\}$$

under the submodel, where $\boldsymbol{\beta}$ may not be the true value $\boldsymbol{\beta}_0$. Then, (assuming that integration and differentiation are exchangeable),

$$\begin{aligned}
\frac{\partial\Psi_t(\boldsymbol{\beta}, h_1, h_2)}{\partial t}\Big|_{t=0} &= \frac{\partial}{\partial t}\mathbb{E}_t\Big\{\frac{S}{\pi(D)}\Big(h_1(\mathbf{X})[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - h_{2,t}(\mathbf{X}, D)\Big)\Big\} \\
&= \mathbb{E}\Big\{\mathbb{S}(O)\frac{S}{\pi(D)}\Big(h_1(\mathbf{X})[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - h_2(\mathbf{X}, D)\Big)\Big\} \\
&\quad + \frac{\partial}{\partial t}\mathbb{E}\Big\{\frac{S}{\pi(D)}\Big(h_1(\mathbf{X})[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - h_{2,t}(\mathbf{X}, D)\Big)\Big\}.
\end{aligned}$$

Consider the second argument.

$$\frac{\partial}{\partial t}\mathbb{E}\Big\{\frac{S}{\pi(D)}\Big(h_1(\mathbf{X})[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - h_{2,t}(\mathbf{X}, D)\Big)\Big\} = -\frac{\partial}{\partial t}\mathbb{E}\Big\{\frac{S}{\pi(D)}h_{2,t}(\mathbf{X}, D)\Big\}.$$

From Lemma 1 (a), with the log link function we have:

$$\begin{aligned}
&\frac{\partial}{\partial t}\mathbb{E}\Big\{\frac{S}{\pi(D)}\Big(h_1(\mathbf{X})[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - h_{2,t}(\mathbf{X}, D)\Big)\Big\} = \\
&-\frac{\partial}{\partial t}\mathbb{E}\Big\{\frac{S}{\pi(D)}h_{2,t}(\mathbf{X}, D)\Big\} = -\frac{\partial}{\partial t}\mathbb{E}\Big\{h(\mathbf{X})\exp\left(\nu(\mathbf{X}, D) - \log\mathbb{E}_t[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]\right)\Big\} \\
&= -\mathbb{E}\Big[h(\mathbf{X})\exp\{\nu(\mathbf{X}, D)\}\frac{\partial}{\partial t}\mathbb{E}_t[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]^{-1}\Big] \\
&= \mathbb{E}\Big(h(\mathbf{X})\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}\big|\mathbf{X}]\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]^{-2}\frac{\partial}{\partial t}\mathbb{E}_t[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]\Big) \\
&= \mathbb{E}\Big(h(\mathbf{X})\mathbb{E}\left[\exp\{\nu(\mathbf{X}, D)\}\big|\mathbf{X}\right]^{-1}\mathbb{E}\left[\exp\{\nu(\mathbf{X}, D)\}\mathbb{S}^{sub}(D|\mathbf{X})\big|\mathbf{X}\right]\Big) \\
&= \mathbb{E}\Big[h(\mathbf{X})\mathbb{E}\left\{\exp\left(\nu(\mathbf{X}, D) - \log\mathbb{E}\left[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}\right]\right)\mathbb{S}^{sub}(D|\mathbf{X})\big|\mathbf{X}\right\}\Big] \\
&= \mathbb{E}\Big\{h(\mathbf{X})\exp\left(\nu(\mathbf{X}, D) - \log\mathbb{E}[\exp\{\nu(\mathbf{X}, D)\}|\mathbf{X}]\right)\mathbb{S}^{sub}(D|\mathbf{X})\Big\} \\
&= \mathbb{E}\Big\{h_2(\mathbf{X}, D)\mathbb{S}^{sub}(D|\mathbf{X})\Big\} = \mathbb{E}\Big\{\frac{S}{\pi(D)}h_2(\mathbf{X}, D)\mathbb{S}^{sub}(D|\mathbf{X})\Big\}.
\end{aligned}$$

We now show the same result for the identity link. We use Lemma 1 (b) and get:

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}\Big\{\frac{S}{\pi(D)}\Big(h_1(\mathbf{X})[Y-g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}]-h_{2,t}(\mathbf{X},D)\Big)\Big\} &= -\frac{\partial}{\partial t}\mathbb{E}\Big\}\frac{S}{\pi(D)}h_{2,t}(\mathbf{X},D)\Big\}\\
=-\frac{\partial}{\partial t}\mathbb{E}\Big[\frac{S}{\pi(D)}\gamma(\mathbf{X})\{D-p_t(\mathbf{X})\}\Big]=\frac{\partial}{\partial t}\mathbb{E}\Big\{\frac{S}{\pi(D)}\gamma(\mathbf{X})p_t(\mathbf{X})\Big\}\\
=\mathbb{E}\Big[\frac{S}{\pi(D)}\gamma(\mathbf{X})\mathbb{E}\{D\mathbb{S}^{sub}(D|\mathbf{X})\Big|\mathbf{X}\}\Big]=\mathbb{E}\Big(\frac{S}{\pi(D)}\gamma(\mathbf{X})\mathbb{E}[\{D-p(\mathbf{X})\}\mathbb{S}^{sub}(D|\mathbf{X})\Big|\mathbf{X}]\Big)\\
=\mathbb{E}\Big[\frac{S}{\pi(D)}\gamma(\mathbf{X})\{D-p(\mathbf{X})\}\mathbb{S}^{sub}(D|\mathbf{X})\Big]=\mathbb{E}\Big\{\frac{S}{\pi(D)}h_2(\mathbf{X},D)\mathbb{S}^{sub}(D|\mathbf{X})\Big\}.\end{aligned}$$

Recall that $p(\mathbf{X})$ is restricted via some nonparametric, semiparameteric or parameteric model, and denote its tangent space by $\Lambda_{D,sub}\subseteq\Lambda_{D,npar}$ where $\Lambda_{D,npar}$ is the tangent space in the unrestricted model for $p(\mathbf{X})$. The score $\mathbb{S}^{sub}(D|\mathbf{X})$ satisfies $\mathbb{S}^{sub}(D|\mathbf{X})\in\Lambda_{D,sub}$, since this tangent space is spanned by all scores of in the submodel for $p(D|\mathbf{X})$. Therefore, $\mathbb{S}^{sub}(D|\mathbf{X})$ is orthogonal to the orthocomplement of the submodel tangent space $\Lambda^{\perp}_{D,sub}$. Denote the projection of a vector v on a space $\mathbf{U}$ by $\Pi(v\Big|\mathbf{U})$. We can decompose

$$\frac{S}{\pi(D)}h_2(\mathbf{X},D) = \Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big|\Lambda_{D,sub}\Big)+\Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big|\Lambda^{\perp}_{D,sub}\Big)$$

and the latter term is orthogonal to $\mathbb{S}^{sub}(D|\mathbf{X})$. Thus,

$$\mathbb{E}\Big\{\frac{S}{\pi(D)}h_2(\mathbf{X},D)\mathbb{S}^{sub}(D|\mathbf{X})\Big\}=\mathbb{E}\Big\{\Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big|\Lambda_{D,sub}\Big)\mathbb{S}^{sub}(D|\mathbf{X})\Big\}.$$

It follows that

$$\begin{aligned}\frac{\partial}{\partial t}\Psi(\boldsymbol{\beta},h_1,h_2)\Big|_{t=0} &= \mathbb{E}\Big\{\mathbb{S}^{sub}(O)\frac{S}{\pi(D)}\Big(h_1(\mathbf{X})[Y-g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}]-h_2(\mathbf{X},D)\Big)\Big\}\\
&\quad-\mathbb{E}\Big\{\Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big|\Lambda_{D,sub}\Big)\mathbb{S}^{sub}(D|\mathbf{X})\Big\}. \qquad \text{(B. 1)}\end{aligned}$$

To complete, it suffices to note that

$$\mathbb{E}\Big[\Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big|\Lambda_{D,sub}\Big)\Big\{\mathbb{S}^{sub}(\mathbf{X})+\mathbb{S}^{sub}(Y|D,\mathbf{X})\Big\}\Big]=0. \qquad \text{(B. 2)}$$

Combining identities (B. 1) and (B. 2), and since every influence functions ψ in the restricted model satisfies the following equation:

$$\frac{\partial \Psi_t(\boldsymbol{\beta}, h_1, h_2)}{\partial t}\bigg|_{t=0} = \mathbb{E}\{\mathbb{S}^{sub}(O)\psi^T\},$$

it follows that every influence function in the restricted model is of the form

$$\frac{S h_1(\mathbf{X})}{\pi(D)}\Big[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\Big] - \frac{S}{\pi(D)}h_2(\mathbf{X}, D) + \Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X}, D)\Big|\Lambda_{D,sub}\Big). \qquad \blacksquare$$

8.3 Proof of Corollary 1

Corollary 1 follows since if $p(\mathbf{X})$ is unrestricted, then so is the tangent space unrestricted $\Lambda_{d,sub} = \Lambda_{D,npar}$, and the projection of a vector on the submodel tangent space does not change the vector, i.e.

$$\Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X}, D)\Big|\Lambda_{D,sub}\Big) = \frac{S}{\pi(D)}h_2(\mathbf{X}, D),$$

so that the influence function has to be

$$\frac{S h_1(\mathbf{X})}{\pi(D)}\Big[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\Big],$$

the IPW influence function. $\blacksquare$

8.4 Proof of Theorem 2

Suppose that $h_1(\mathbf{X})$ is fixed, and $p(\mathbf{X})$ is known. We here find the function $h_2(\mathbf{X}, D)$ that minimizes the variance over all functions in the submodel tangent space. First, note that we can write the influence functions for $\boldsymbol{\beta}$ in the form:

$$\begin{aligned}
\psi(\boldsymbol{\beta}) &= \frac{S h_1(\mathbf{X})}{\pi(D)}\Big[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\Big] - \Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X}, D)\Big|\Lambda^{\perp}_{D,sub} \cap \Lambda_{D,npar}\Big) \\
&= \Pi\Big(\frac{S h_1(\mathbf{X})}{\pi(D)}\Big[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\Big]\Big|\Lambda^{\perp}_{D,sub} \cap \Lambda_{D,npar}\Big) - \Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X}, D)\Big|\Lambda^{\perp}_{D,sub} \cap \Lambda_{D,npar}\Big) \\
&\quad +\Pi\Big(\frac{S h_1(\mathbf{X})}{\pi(D)}\Big[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\Big]\Big|\Lambda_{D,sub}\Big),
\end{aligned}$$

so that minimizing the variance of $\psi(\boldsymbol{\beta})$ over functions $h_2(\mathbf{X}, D)$ is equivalent to minimizing the variance of the first two terms (since the third term is orthogonal to the term involving $h_2(\mathbf{X}, D)$). Consider finding

$h_2^{opt}(\mathbf{X}, D)$ that satisfies the normal equations:

$$\begin{aligned} 0 &= \mathbb{E}\Big\{\frac{S}{\pi(D)}h_2(\mathbf{X}, D)\Big(\frac{S\,h_1(\mathbf{X})}{\pi(D)}[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - \frac{S}{\pi(D)}h_2^{opt}(\mathbf{X}, D)\Big)\Big\} \\ &= \mathbb{E}\Big\{\frac{S}{\pi(D)}h_2(\mathbf{X}, D)\Big(\frac{S\,h_1(\mathbf{X})}{\pi(D)}[\mathbb{E}(Y|\mathbf{X}, D) - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - \frac{S}{\pi(D)}h_2^{opt}(\mathbf{X}, D)\Big)\Big\}. \end{aligned}$$

This equality is satisfied by

$$h_2^{opt}(\mathbf{X}, D) = h_1(\mathbf{X})[\mathbb{E}(Y|\mathbf{X}, D) - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}],$$

as required. ∎

8.5 Proof of Theorem 3

We here find the function $h_1^{opt}(\mathbf{X})$ that using it in the estimating equation $\mathbf{U}_{cont}(\boldsymbol{\beta})$ yields the most efficient (with minimal variance) estimator of $\widehat{\boldsymbol{\beta}}$. According to the generalized information equality (**?**), for every function $h_1(\mathbf{X})$:

$$-\mathbb{E}\Big[\frac{\partial \mathbf{U}_{cont}^{opt}\{\boldsymbol{\beta}; h_1(\mathbf{X})\}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\Big] = \mathbb{E}\Big[\mathbf{U}_{cont}^{opt}\{\boldsymbol{\beta}; h_1(\mathbf{X})\}\mathbf{U}_{cont}\{\boldsymbol{\beta}; h_1^{opt}(\mathbf{X})\}^T\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\Big].$$

Then:

$$\begin{aligned} \mathbb{E}\Big[\frac{S}{\pi(D)}h_1(\mathbf{X})\frac{\partial g^{-1}\{\tilde{\mu}(\mathbf{X}, D;\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\Big] &= \mathbb{E}\Big[h_1(\mathbf{X})\frac{\partial g^{-1}\{\tilde{\mu}(\mathbf{X}, D;\boldsymbol{\beta}_0)\}}{\partial \boldsymbol{\beta}}\Big] \\ &= \mathbb{E}\Big\{h_1(\mathbf{X})h_1^{opt}(\mathbf{X})\mathbb{E}\left(\frac{1}{\pi(D)}[Y - g^{-1}\{\tilde{\mu}(\mathbf{X}, D;\boldsymbol{\beta}_0)\}]^2\Big|\mathbf{X}\right)\Big\}. \end{aligned}$$

This equation is satisfied by:

$$\frac{\partial g^{-1}\{\tilde{\mu}(\mathbf{X}, D;\boldsymbol{\beta}_0)\}}{\partial \boldsymbol{\beta}} = h_1^{opt}(\mathbf{X})\mathbb{E}\Big(\frac{1}{\pi(D)}[Y - g^{-1}\{\tilde{\mu}(\mathbf{X}, D;\boldsymbol{\beta}_0)\}]^2\Big|\mathbf{X}\Big).$$

Recall that $\tilde{\mu}(\mathbf{X}, D; \boldsymbol{\beta}) = g\{\mathbb{E}(Y|\mathbf{X}, D)\}$. We can then write:

$$
\begin{aligned}
h_1^{opt}(\mathbf{X}) &= \mathbb{E}\Big[\frac{1}{\pi(D)}\{Y - \mathbb{E}(Y|\mathbf{X}, D)\}^2\Big|\mathbf{X}\Big]^{-1} \frac{\partial g^{-1}\{\tilde{\mu}(\mathbf{X}, D; \boldsymbol{\beta}_0)\}}{\partial \boldsymbol{\beta}} \\
&= \mathbb{E}\Big\{\frac{1}{\pi(D)}\mathrm{Var}(Y\Big|\mathbf{X}, D)\Big|\mathbf{X}\Big\}^{-1} \frac{\partial g^{-1}\{\tilde{\mu}(\mathbf{X}, D; \boldsymbol{\beta}_0)\}}{\partial \boldsymbol{\beta}},
\end{aligned}
$$

as required. ∎

8.6 Deriving the locally semiparametric efficient influence function

We first derive the estimating equation for $\boldsymbol{\theta}$ accounting for the estimation of $\boldsymbol{\alpha}$, and then provide the corresponding influence function for $\boldsymbol{\theta}$. Denote the true value of $\boldsymbol{\alpha}$ by $\boldsymbol{\alpha}_0$, and recall that $\mathbf{V}(\boldsymbol{\alpha})$ is the estimating equation for $\boldsymbol{\alpha}$, and denote for simplicity $\mathbf{U}(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \mathbf{U}_{cont}^{opt}(\boldsymbol{\theta})$. (In fact, the following derivation holds for any estimating equation $\mathbf{U}(\boldsymbol{\theta}; \boldsymbol{\alpha})$, in particular to any functions $h_1(\mathbf{X}), h_2(\mathbf{X}, D)$, not just the optimal ones). Let $\mathbf{V}_i(\boldsymbol{\alpha}), \mathbf{U}_i(\boldsymbol{\theta}; \boldsymbol{\alpha})$ be the contributions of the ith subject to the estimating equations. To estimate $\boldsymbol{\alpha}, \boldsymbol{\theta}$, one solves $\mathbb{P}_n\mathbf{V}_i(\boldsymbol{\alpha}) = 0,\ \mathbb{P}_n\mathbf{U}_i(\boldsymbol{\theta}; \boldsymbol{\alpha}) = 0$, where $\mathbb{P}_n(x_i) = 1/n\sum_{i=1}^n x_i$.

Consider the following expansions of the estimating equations around $\boldsymbol{\alpha}_0$:

$$
\begin{aligned}
\sqrt{n}\mathbb{P}_n\mathbf{U}_i(\boldsymbol{\theta}; \widehat{\boldsymbol{\alpha}}) &= \sqrt{n}\mathbb{P}_n\mathbf{U}_i(\boldsymbol{\theta}; \boldsymbol{\alpha})\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} + \sqrt{n}\mathbb{P}_n\frac{\partial}{\partial \boldsymbol{\alpha}}\mathbf{U}_i(\boldsymbol{\theta}; \widehat{\boldsymbol{\alpha}})\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1) \\
&= \sqrt{n}\mathbb{P}_n\mathbf{U}_i(\boldsymbol{\theta}; \boldsymbol{\alpha}_0) + E\left\{\frac{\partial}{\partial \boldsymbol{\alpha}}\mathbf{U}(\boldsymbol{\theta}; \boldsymbol{\alpha}_0)\right\}\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1).
\end{aligned}
$$

Similarly,

$$
\sqrt{n}\mathbb{P}_n\mathbf{V}_i(\widehat{\boldsymbol{\alpha}}) = \sqrt{n}\mathbb{P}_n\mathbf{V}_i(\boldsymbol{\alpha}_0) + E\left\{\frac{\partial}{\partial \boldsymbol{\alpha}}\mathbf{V}(\boldsymbol{\alpha}_0)\right\}\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1).
$$

From the estimation procedure, we have that by definition $\sqrt{n}\mathbb{P}_n\mathbf{V}_i(\widehat{\boldsymbol{\alpha}}) = 0$. Therefore, combining these two equations we get:

$$
\sqrt{n}\mathbb{P}_n\mathbf{U}_i(\boldsymbol{\theta}; \widehat{\boldsymbol{\alpha}}) = \sqrt{n}\mathbb{P}_n\Big[\mathbf{U}_i(\boldsymbol{\theta}; \boldsymbol{\alpha}_0) - E\left\{\frac{\partial}{\partial \boldsymbol{\alpha}}\mathbf{U}(\boldsymbol{\theta}; \boldsymbol{\alpha}_0)\right\}E\left\{\frac{\partial}{\partial \boldsymbol{\alpha}}\mathbf{V}(\boldsymbol{\alpha}_0)\right\}^{-1}\mathbf{V}_i(\boldsymbol{\alpha}_0)\Big] + o_p(1).
$$

So that there is an additional term, namely $\sqrt{n}\mathbb{P}_n E\{\frac{\partial}{\partial \boldsymbol{\alpha}}\mathbf{U}(\boldsymbol{\theta}; \boldsymbol{\alpha}_0)\}E\{\frac{\partial}{\partial \boldsymbol{\alpha}}\mathbf{V}(\boldsymbol{\alpha}_0)\}^{-1}\mathbf{V}_i(\boldsymbol{\alpha}_0)$, that accounts for the estimation of $\boldsymbol{\alpha}$. Notice that in order to estimate $\boldsymbol{\theta}$ we do not in fact need to use this estimating equation, since $\sqrt{n}\mathbb{P}_n\mathbf{V}_i(\boldsymbol{\alpha}_0)$ is estimated by $\sqrt{n}\mathbb{P}_n\mathbf{V}_i(\widehat{\boldsymbol{\alpha}}) = 0$. However, for the purpose of variance estimation, it is important to use this estimating equation and account for the estimation of $\boldsymbol{\alpha}$.

Using the same technic, we obtain

$$\sqrt{n}\mathbb{P}_n\mathbf{U}_i(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\alpha}}) \quad = \quad \sqrt{n}\mathbb{P}_n\mathbf{U}_i(\boldsymbol{\theta}_0;\widehat{\boldsymbol{\alpha}}) + E\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta}_0;\widehat{\boldsymbol{\alpha}})\right\}\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + o_p(1),$$

and since $\sqrt{n}\mathbb{P}_n\mathbf{U}_i(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\alpha}}) = 0$, we get:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = \sqrt{n}\mathbb{P}_n\Big[E\Big\{\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta}_0;\widehat{\boldsymbol{\alpha}})\Big\}^{-1}\mathbf{U}_i(\boldsymbol{\theta}_0;\widehat{\boldsymbol{\alpha}})\Big] + o_p(1),$$

and we see that $\widehat{\boldsymbol{\theta}}$ is an asymptotically linear estimator with the ith influence function given by

$$\psi_i(\boldsymbol{\theta};\boldsymbol{\alpha}) = E\Big\{\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}_i(\boldsymbol{\theta}_0;\widehat{\boldsymbol{\alpha}})\Big\}^{-1}\mathbf{U}_i(\boldsymbol{\theta}_0;\widehat{\boldsymbol{\alpha}}).$$

Notice that

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}_i(\boldsymbol{\theta}_0;\widehat{\boldsymbol{\alpha}}) \quad &= \quad \frac{\partial}{\partial\boldsymbol{\theta}}\Big[\mathbf{U}_i(\boldsymbol{\theta};\boldsymbol{\alpha}_0) - E\Big\{\frac{\partial}{\partial\boldsymbol{\alpha}}\mathbf{U}(\boldsymbol{\theta};\boldsymbol{\alpha}_0)\Big\}E\Big\{\frac{\partial}{\partial\boldsymbol{\alpha}}\mathbf{V}(\boldsymbol{\alpha}_0)\Big\}^{-1}\mathbf{V}_i(\boldsymbol{\alpha}_0)\Big] \\
&= \quad \frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}_i(\boldsymbol{\theta};\boldsymbol{\alpha}_0) - E\Big\{\frac{\partial}{\partial\boldsymbol{\alpha}}\mathbf{U}(\boldsymbol{\theta};\boldsymbol{\alpha}_0)\Big\}E\Big\{\frac{\partial}{\partial\boldsymbol{\alpha}}\mathbf{V}(\boldsymbol{\alpha}_0)\Big\}^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{V}_i(\boldsymbol{\alpha}_0) \\
&= \quad \frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}_i(\boldsymbol{\theta};\boldsymbol{\alpha}_0),
\end{aligned}$$

since $\mathbf{V}_i(\boldsymbol{\alpha}_0)$ does not depend on $\boldsymbol{\theta}$.

8.7 Proof of Corollary 2

Under the standard regularity conditions found in **?**, the asymptotic normality of $\widehat{\boldsymbol{\theta}}$ follows from the central limit theorem, and its mean and covariance are as indicated since we assume that the models for $\widehat{\boldsymbol{\theta}} = (\boldsymbol{\beta},\boldsymbol{\delta})$ are correctly specified, so that $\psi(\boldsymbol{\theta};\boldsymbol{\alpha})$ has mean zero. Local efficiency follows from Theorem 3, in which we provide the efficient influence function, and from the definition of local efficiency (**?**). ∎

9 Computation of the control function estimator

Here we describe how to compute estimators of $\boldsymbol{\beta}$ for the identity and log links, when $p(\mathbf{X})$ is modeled parametrically with $p(\mathbf{X};\boldsymbol{\alpha})$. In general, to find the estimator $\widehat{\boldsymbol{\beta}}$ we need to solve the estimating equation $\widehat{\mathbf{U}}^{opt}_{cont}(\boldsymbol{\beta}) = 0$, defined as $\mathbf{U}^{opt}_{cont}(\boldsymbol{\beta})$ with $\widehat{h}_1(\mathbf{X}), \widehat{h}_2(\mathbf{X}, D)$, and $\widehat{p}(\mathbf{X})$. This can be performed using the Newton-Raphson (NR) algorithm.

Let $\boldsymbol{\delta}$ denote the parameters for either $\nu(\mathbf{X}, D; \boldsymbol{\delta})$ (log link) or $\gamma(\mathbf{X}; \boldsymbol{\delta})$ (identity link). Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$. It is convenient to estimate $\boldsymbol{\theta}$ jointly, by modifying the estimating equation $\mathbf{U}^{opt}_{cont}(\boldsymbol{\beta})$ to define $\mathbf{U}^{opt}_{cont}(\boldsymbol{\theta})$ by taking

$$h_1^{opt}(\mathbf{X}) = \mathbb{E}\left\{\frac{1}{\pi(D)}\text{var}(Y|D, \mathbf{X})\middle|\mathbf{X}\right\}^{-1} \frac{\partial}{\partial\boldsymbol{\theta}}[g^{-1}\{\mu(\mathbf{X}, D; \boldsymbol{\theta})\}].$$

The estimation procedure takes the following steps:

1. Estimate the parameters of $p(D = 1|\mathbf{X}, S = 1)$, the probability of disease conditional on covariates in the case-control study population, using logistic regression with an offset, by exploiting the known relationship between disease probability in the population to disease probability in the case-control sample:

$$\text{logit}p(\mathbf{X}) = \text{logit}p(D = 1|\mathbf{X}, S = 1) + \log\left[\frac{p(D = 1)\{1 - p(D = 1|S = 1)\}}{p(D = 1|S = 1)\{1 - p(D = 1)\}}\right] \qquad \text{(C. 1)}$$

   where $p(D = 1)$ is the disease prevalence in the general population, and $p(D = 1|S = 1)$ is the fraction of cases in the case-control sample.

2. Obtain *starting values* for $\boldsymbol{\theta}$ according to the specifics given below.

3. Plug $\widehat{\boldsymbol{\alpha}}$ into $\mathbf{U}^{opt}_{cont}(\boldsymbol{\theta})$ and solve $\widehat{\mathbf{U}}^{opt}_{cont}(\widehat{\boldsymbol{\theta}}) = 0$ using NR.

This procedure is implemented in the R package RECSO (**?**). Note that the estimating equation $\mathbf{U}^{opt}_{cont}(\boldsymbol{\theta})$ is geared towards increasing the efficiency of the estimator for $\boldsymbol{\beta}$, so $\widehat{\boldsymbol{\delta}}$ may not be an efficient estimator.

Next we detail the estimation procedure for $p(D = 1|\mathbf{X}, S = 1)$, and $\boldsymbol{\theta}$ for each choice of link function.

9.1 Computation of $\widehat{p}(D = 1|\mathbf{X}, S = 1)$ using a parametric model

Let $\mathbf{V}(\boldsymbol{\alpha})$ be the estimating equation for parameters $\boldsymbol{\alpha}$ of $p(D = 1|\mathbf{X}, S = 1; \boldsymbol{\alpha})$. In the simple logistic model, it is given by:

$$\mathbf{V}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} S_i \boldsymbol{x}_i \{D_i - p(D_i = 1|\mathbf{X}_i, S_i = 1)\}$$

where $p(D_i = 1|\boldsymbol{x}_i, S_i = 1)$ is modeled through the inverse of the logit transformation, i.e. $p(D_i = 1|\boldsymbol{x}_i, S_i = 1) = \exp(\boldsymbol{\alpha}^T \boldsymbol{x}_i)/\{1 + \exp(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\}$. Here, we do not correct for the sampling bias resulting from the case-control ascertainment (e.g. we do not use IPW), but to later obtain estimates of the disease probability $\widehat{p}(\mathbf{X})$ we use the correction (C. 1).

9.2 Identity link

To solve the estimating equation $\widehat{\mathbf{U}}_{ident}(\boldsymbol{\beta}) = 0$ for $\boldsymbol{\beta}$, we follow the steps described above. First, we estimate $\widehat{\boldsymbol{\alpha}}$. Second, we calculate staring values for $\boldsymbol{\theta}$. For staring values of $\boldsymbol{\delta}$, we can estimate $\widehat{\gamma}(\mathbf{X})$ by regressing Y on the covariates $\mathbf{X}$ in the cases and control groups separately, calculating the predicted means for each subject under the two models, and taking the difference. Initial estimators of $\boldsymbol{\delta}$ can then be obtained by regressing the calculated differences on a given design matrix, say, if a linear model is assumed. Starting value for $\boldsymbol{\beta}$ could be obtained as the IPW estimator. At the third step we solve

$$0 = \sum_{i=1}^{n} \frac{h_1^{opt}(\mathbf{X}_i) S_i}{\pi(D_i)} \Big[\{Y_i - \mu(\mathbf{X}_i; \boldsymbol{\beta})\} - \{D_i - \widehat{p}(\mathbf{X}_i)\} \gamma(\mathbf{X}_i; \boldsymbol{\delta}) \Big]$$

using NR iterations, by which we update the estimated $\widehat{\boldsymbol{\theta}}$ until convergence.

Note that for $h_1^{opt}(\mathbf{X})$ we need to estimate $\mathrm{Var}(Y \big| D, \mathbf{X})$. When the outcome is continuous, it is convenient to assume homoscedasticity , in which case $h_1^{opt}(\mathbf{X}_i)$ can be chosen

$$\sum_{d \in \{0,1\}} \left[\frac{1}{\pi(d)} \frac{1}{n} \sum_{j=1}^{n} (y_i - \mu(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}) - \widehat{\gamma}(\mathbf{X}_i) \{d - \widehat{p}(\mathbf{X}_i)\}^2 \right] p(D_i = d)$$

The estimate $\widehat{\boldsymbol{\beta}}$ will remain consistent even if the homoscedasticity assumption does not hold.

9.3 Log link

As in the identity link case, we start by estimating $\widehat{\boldsymbol{\alpha}}$, as described earlier. At the second step, we calculate starting values for $\boldsymbol{\theta}$. $\widehat{\nu}(\mathbf{X}, D)$ could be estimated, for instance, by estimating the parameters of a generalized linear model with the log link function, of Y on the covariates $\mathbf{X}$ in the cases and controls separately, calculating the predicted means $\mathbb{E}(Y_i|\mathbf{X}_j, D_i = 0)$ and $\mathbb{E}(Y_i|\mathbf{X}_j, D_i = 1)$ for every subject i, and plugging-in to the equation for $\nu(\mathbf{X}, D)$ for each subject. We can then estimate an initial $\widehat{\boldsymbol{\delta}}$ based on a model. A starting value for $\boldsymbol{\beta}$ could be the IPW estimator. We can proceed to the third step and solve

$$0 = \sum_{i=1}^{n} \frac{h_1^{opt}(\mathbf{X}_i) S_i}{\pi(D_i)} \Big(Y_i - \exp\left[\mu(\mathbf{X}_i;\boldsymbol{\beta}) + \nu^{opt}(\mathbf{X}_i, D_i;\boldsymbol{\delta}) - \bar{\nu}\{\mathbf{X}_i;\boldsymbol{\delta},\widehat{p}(\mathbf{X}_i)\}\right]\Big)$$

for $\boldsymbol{\theta}$ using NR iterations.

Note that at the kth iteration, we also need to estimate $h_1^{opt}(\mathbf{X})$. We can either update the estimate of $h_1^{opt}(\mathbf{X})$ at the kth iteration, using the estimated $\widehat{\boldsymbol{\theta}}$ from the $(k-1)$th iteration, or we can use a plug-in estimator based on the initial estimator of $\boldsymbol{\theta}$. Usually the latter option is more stable (updating $h_1^{opt}(\mathbf{X})$ may lead to convergence problems). Note that for $h_1^{opt}(\mathbf{X})$ one needs an estimate of

$$\mathbb{E}\left\{\frac{1}{\pi(D_i)}\text{Var}(Y_i|\mathbf{X}_i, D_i)\middle|\mathbf{X}_i\right\} = \sum_{d\in\{0,1\}} \left\{\frac{1}{\pi(D_i)}\text{Var}(Y_i|\mathbf{X}_i, D_i)p(D_i = d|\mathbf{X}_i)\right\}$$

for each subject $i, i = 1, \ldots, n$. In the case of a Poisson model, we can simply use the predicted means, as $\widehat{\text{Var}}(Y|\mathbf{X}, D) = \widehat{\mathbb{E}}(Y|\mathbf{X}, D_i) = \exp\{\mu(\mathbf{X};\widehat{\boldsymbol{\beta}}) + \nu^{opt}(\mathbf{X}, D;\widehat{\boldsymbol{\delta}}) - \bar{\nu}(\mathbf{X};\widehat{\boldsymbol{\delta}})\}$. As before, these predicted means could be updated at each iteration or be based on the initial estimators (the more stable option).

10 Identity link simulations - additional information

10.1 Simulation study with a single exposure variable

The simulation study described here, is similar to the identity link simulation study presented in the manuscript (Section 4.1), but simpler, so that only a single exposure variable is used. In this simulation we implemented and compared the estimator TT of Tchetgen Tchetgen (2014), which this estimator is not presented in the more complex simulation studies in the manuscript, as it then suffered from convergence problems. The TT estimator was calculated using maximum likelihood, and the robust standard error estimators. Results are provided under correct specification of the selection bias function $\gamma(\mathbf{X})$ (TT-cor) and under misspecification (TT-mis).

The simulation was generated as follows. As in the simulation study presented in the main manuscript, first, an exposures variables X_1 was sampled with distribution $X_1 \sim \mathcal{N}(2, 4)$. Then, disease probabilities were calculated for each subject, from the model

$$\operatorname{logit}\{p(D = 1|\mathbf{X})\} = -3.2 + 0.3X_1,$$

and disease status was sampled. Then, the conditional mean of the secondary outcome was set to

$$\mathbb{E}(Y|\mathbf{X}, D) = 50 + 4X_1 + \{D - p(\mathbf{X})\}(3 + 2X_1),$$

so that the population mean is $\mu(\mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}^T\boldsymbol{\beta}$ with $\mathbf{X} = (1, X_1)^T$ and $\boldsymbol{\beta} = (50, 4)^T$, and $\gamma(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\alpha}$ with $\boldsymbol{\alpha} = (3, 2)^T$. Finally, the residuals were normally distributed, so that Y_i was sampled from:

$$Y_i = \mathbb{E}(Y|\mathbf{X}_i, D_i) + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, 4).$$

All estimators estimated the sample mean based on the full design matrix, i.e. with $\mathbf{X} = (1, X_1)^T$. TT and the control function estimator estimated $\gamma(\mathbf{X})$. When the model was correctly specified, the design matrix in the model for $\gamma(\mathbf{X})$ was taken to include all the terms $\mathbf{X} = (1, X_1)^T$, but when the model was incorrectly specified, it only had the intercept, i.e. $\mathbf{X} = 1$.

Table 3 provides comprehensive simulation results (i.e. all summary statistics for all estimators under investigations), while Figure 4 provides graphical results, comparing the bias, MSE and coverage of the unbiased estimators cont-mis, cont-cor, IPW and TT-cor.

10.2 Table summarizing the identity link simulations provided in Section 4.1 in the manuscript

The following Table 4 provide comprehensive simulation results for the simulation study described in Section 4.1 in the paper.

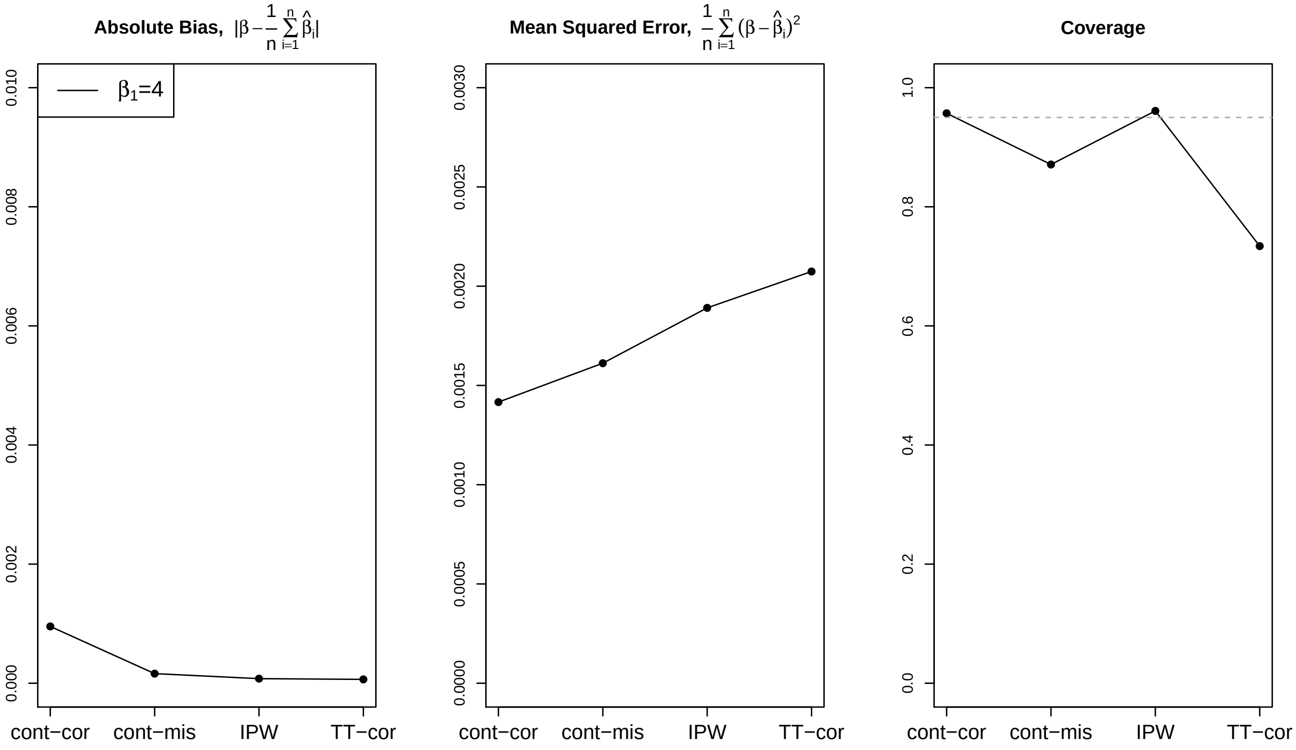


Figure 4: Results from Identity link simulations in the simple settings with a single covariate. Estimated bias, MSE, and coverage probability of the control function under correct and misspecification of the selection bias function (cont-cor, cont-mis, respectively), IPW and TT (correctly specified) estimators, in estimating the population effect of X_1.

Table 3: Simulation results for estimating the effect of covariates on a normally distributed secondary outcome using the identity link function, in the first, simple settings (a single covariate). We report results for the usual IPW estimator, the proposed estimator with the control function, when the model for $\nu(\mathbf{X}, D)$ is correctly specific ('cont-cor') and when the model is misspecified ('cont-mis1'), the naïve conditional and pooled estimators (Dind and pooled) with and without disease status in the regression model, respectively, and the estimator proposed by Tchetgen Tchetgen (2014) (TT).

Estimator	bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 50$					
cont-cor	0.000	0.018	0.136	0.133	0.942
cont-mis	−0.001	0.018	0.136	0.142	0.957
IPW	−0.002	0.019	0.137	0.134	0.939
pooled	1.640	2.711	0.145	0.200	0.000
Dind	−1.450	2.126	0.151	0.165	0.000
TT-cor	0.000	0.018	0.135	0.130	0.942
TT-mis	−0.930	0.889	0.157	0.158	0.000
$X_1, \beta_1 = 4$					
cont-cor	−0.001	0.001	0.038	0.039	0.957
cont-mis	0.000	0.002	0.040	0.032	0.871
IPW	0.000	0.002	0.044	0.045	0.961
pooled	0.753	0.568	0.036	0.036	0.000
Dind	0.149	0.024	0.041	0.030	0.009
TT-cor	0.000	0.002	0.046	0.026	0.734
TT-mis	0.497	0.249	0.039	0.035	0.000

Table 4: Simulation results for estimating the effect of covariates on a normally distributed secondary outcome using the identity link function, in the second settings (two covariates, interaction term in the population regression and selection bias models). We report results for the usual IPW estimator, the proposed estimator with the control function, when the model for $\nu(\mathbf{X}, D)$ is correctly specific ('cont-cor') and when the model is misspecified ('cont-mis1') and the naïve conditional and pooled estimators (Dind and pooled) with and without disease status in the regression model, respectively.

Estimator	bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 50$					
cont-cor	0.007	0.019	0.138	0.139	0.958
cont-mis1	0.007	0.019	0.138	0.139	0.959
cont-mis2	0.007	0.019	0.138	0.141	0.961
cont-mis3	0.006	0.019	0.139	0.150	0.971
cont-mis4	0.006	0.019	0.139	0.153	0.973
IPW	0.006	0.019	0.139	0.141	0.964
pooled	1.520	2.332	0.150	0.227	0.000
Dind	−1.577	2.515	0.165	0.183	0.000
$X_1, \beta_1 = 4$					
cont-cor	−0.001	0.001	0.038	0.042	0.967
cont-mis1	−0.001	0.001	0.038	0.044	0.971
cont-mis2	−0.001	0.001	0.038	0.047	0.982
cont-mis3	0.000	0.002	0.040	0.034	0.906
cont-mis4	0.000	0.002	0.040	0.035	0.923
IPW	0.000	0.002	0.045	0.047	0.964
pooled	0.724	0.526	0.038	0.041	0.000
Dind	0.077	0.008	0.042	0.034	0.398
$X_2, \beta_2 = 3$					
cont-cor	0.028	0.228	0.477	0.491	0.960
cont-mis1	0.024	0.236	0.485	0.526	0.970
cont-mis2	0.026	0.238	0.487	0.431	0.913
cont-mis3	0.017	0.268	0.517	0.656	0.984
cont-mis4	0.021	0.268	0.517	0.536	0.953
IPW	0.024	0.272	0.521	0.521	0.950
pooled	2.256	5.461	0.608	0.648	0.051
Dind	−0.061	0.419	0.645	0.479	0.852
$X_1X_2, \beta_3 = 3$					
cont-cor	0.005	0.022	0.148	0.207	0.998
cont-mis1	0.011	0.025	0.159	0.164	0.955
cont-mis2	0.014	0.032	0.179	0.116	0.775
cont-mis3	0.016	0.039	0.197	0.146	0.843
cont-mis4	0.015	0.047	0.216	0.146	0.795
IPW	0.018	0.076	0.275	0.247	0.909
pooled	0.317	0.126	0.160	0.117	0.297
Dind	0.366	0.156	0.148	0.086	0.085

11 Simulation study: log link

We compared the control function estimator to pooled and Dind, that were calculated using generalized linear models in standard software. We simulated two covariates, X_1 and X_2, with $X_1 \sim \mathcal{N}(1, 0.2)$ and $X_2 \sim \mathcal{N}(1.5, 0.2)$. Primary disease probability was calculated by

$$\text{logit}\{p(D = 1|\mathbf{X})\} = -2.12 + 0.3X_1 + X_2,$$

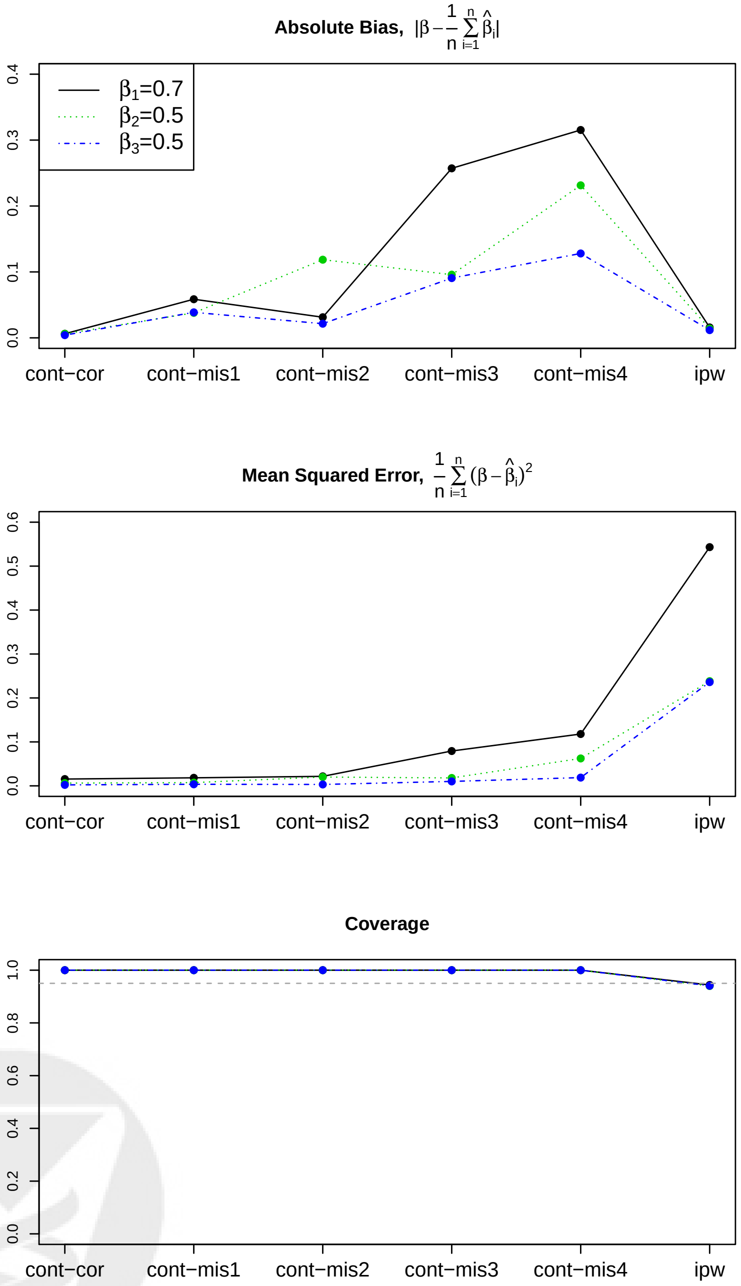
so that disease prevalence is 0.12. Disease statuses were sampled from the calculated probabilities. The secondary outcome mean was calculated by:

$$\begin{aligned}\mathbb{E}(Y|\mathbf{X}, D) &= \exp\{3 + 0.7X_1 + (0.3 + 0.5X_1 + 0.5X_1X_2)D\} \\ &\quad \times \exp\left[-\log\{\exp(0.5 + 0.3X_1 + 0.3X_2 + 0.3X_1X_2)p(D = 1|\mathbf{X}) + p(D = 0|\mathbf{X})\}\right],\end{aligned}$$

so that the population mean is $\exp\{\mu(\mathbf{X}, \boldsymbol{\beta})\} = \exp(\mathbf{X}^T\boldsymbol{\beta})$ with $\mathbf{X} = (1, X_1, X_2, X_1X_2)^T$ and $\boldsymbol{\beta} = (3, 0.7, 0.5, 0.5)^T$, and $\nu(\mathbf{X}, D) = D\mathbf{X}^T\boldsymbol{\alpha}$ with $\boldsymbol{\alpha} = (0.5, 0.3, 0.3, 0.3)^T$. Then Y was sampled from Poisson distributed, i.e. $Y \sim \text{Poisson}\{\mathbb{E}(Y|\mathbf{X}, D)\}$. 1000 cases and controls were sampled from the generated population.

All estimators estimated the sample mean based on the full design matrix, i.e. with $\mathbf{X} = (1, X_1, X_2, X_1X_2)^T$. The control function estimator estimated $\nu(\mathbf{X}, D)$. When the model was correctly specified, the design matrix was taken to include all of $\mathbf{X}$. To study the effect of misspecification, we implemented the control function estimator with the following misspecifications of the selection bias function $\nu(\mathbf{X}, D)$: cont-mis1 had design matrix $\mathbf{X} = (1, X_1, X_2)$. cont-mis2 had design matrix $\mathbf{X} = (1, X_1)^T$, cont-mis3 had $\mathbf{X} = (1, X_2)^T$, and cont-mis4 had only intercept.

Figure 5, provides the bias, MSE and coverage probabilities of the IPW and the control function estimators, calculated over the 1000 simulations. Table 13 reports, for each estimator and each estimated parameter, the estimator's mean bias, MSE, empirical standard deviation over all simulations, mean estimated standard deviation, and coverage probability. The bias of the control function estimator is small under correct specification of the selection bias function, but increases as more information is lost in various forms of misspecification. For instance, consider the estimator β_1, the coefficient of X_1. When the interaction term, or both interaction term and X_2, are not included in the design matrix for $\nu(\mathbf{X}, D)$ (cont-mis1, cont-mis2), it becomes slightly biased. When both interaction term and X_1, or all covariates, are not included in the design matrix (cont-mis3, cont-mis4), its bias more than doubles. However, surprisingly, the MSE of the control function estimators is superior to the IPW, and performs well even when the model for $\nu(\mathbf{X}, D)$ is misspecified. When the model for $\nu(\mathbf{X}, D)$ is misspecified, the bias, the MSE and the empirical standard deviation of the control function estimator were higher than under correct specification. Coverage probability was inflated and very close to 1, both when the model for $\nu(\mathbf{X}, D)$ was correctly specified and when it was misspecified. In comparison, the coverage probability of the IPW estimator was accurate. Finally, as in the identity link simulations, the naïve estimators Dind and pooled yielded biased estimators with, substantially lower than nominal, coverage probability.



Absolute Bias, $|\beta - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_i|$

Mean Squared Error, $\frac{1}{n}\sum_{i=1}^{n}(\beta - \hat{\beta}_i)^2$

Coverage

Figure 5: Results from log link simulations. Estimated bias, MSE, and coverage probability of the control function under correct and misspecification of the selection bias function (cont-cor, cont-mis1, . . ., cont-mis4), and IPW, in estimating population means.

Table 5: Simulation results for estimating the effect of covariates on a Poisson distributed secondary outcome using the log link function. We report results for the usual IPW estimator, the proposed estimator with the control function, when the model for $v(\mathbf{X}, D)$ is correctly specific ('cont-cor') and when the model is misspecified, under four forms of misspecification ('cont-mis1', ..., 'cont-mis4'), and the naïve conditional and pooled estimators (Dind and pooled) with and without disease status in the regression model, respectively.

Estimator	bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3$					
ours-cor	−0.009	0.023	0.151	0.645	1.000
ours-mis1	0.057	0.025	0.148	0.642	1.000
ours-mis2	−0.181	0.059	0.163	0.637	1.000
ours-mis3	−0.272	0.101	0.165	0.638	1.000
ours-mis4	−0.482	0.264	0.178	0.642	1.000
IPW	−0.020	0.546	0.739	0.730	0.944
pooled	0.025	0.444	0.666	0.064	0.135
Dind	−0.484	0.245	0.104	0.064	0.002
$X_1, \beta_1 = 0.7$					
ours-cor	0.006	0.015	0.124	0.641	1.000
ours-mis1	−0.059	0.018	0.122	0.639	1.000
ours-mis2	0.031	0.022	0.144	0.635	1.000
ours-mis3	0.257	0.079	0.115	0.627	1.000
ours-mis4	0.315	0.118	0.136	0.628	1.000
IPW	0.016	0.543	0.737	0.725	0.944
pooled	0.095	0.447	0.662	0.059	0.150
Dind	−0.078	0.016	0.097	0.060	0.623
$X_2, \beta_2 = 0.5$					
ours-cor	0.006	0.006	0.079	0.424	1.000
ours-mis1	−0.038	0.007	0.077	0.423	1.000
ours-mis2	0.119	0.020	0.078	0.416	1.000
ours-mis3	0.096	0.018	0.094	0.421	1.000
ours-mis4	0.231	0.062	0.094	0.419	1.000
IPW	0.014	0.238	0.488	0.481	0.940
pooled	0.044	0.193	0.437	0.041	0.148
Dind	−0.348	0.126	0.067	0.041	0.002
$X_1X_2, \beta_3 = 0.5$					
ours-cor	−0.004	0.002	0.047	0.422	1.000
ours-mis1	0.039	0.004	0.046	0.420	1.000
ours-mis2	−0.021	0.003	0.052	0.415	1.000
ours-mis3	−0.091	0.010	0.044	0.413	1.000
ours-mis4	−0.128	0.019	0.049	0.409	1.000
IPW	−0.012	0.236	0.486	0.477	0.941
pooled	−0.049	0.190	0.433	0.038	0.141
Dind	−0.002	0.004	0.063	0.038	0.758

12 Simulation study mimicking the T2D case-control study data set

The goal of these simulations was to study the performance of the control function estimator in simulations mimicking the T2D data set, by using the same variable types, as well as effect sizes, as seen in the data. We considered a few forms of misspecification of the selection bias function, to glean into the plausible effects of misspecification on estimation.

First, we took two SNPs that were found to be significantly associated with log-BMI an entire GWAS data analysis. These SNPs, dubbed SNP1 and SNP2, had very low Minor Allele Frequency (MAF), about 3%. We estimated the logistic disease model with the predictors: smoking status, alcohol measure, physically active status, and SNP1 and SNP2. We also estimated the regression model $\mathbb{E}[Y|\mathbf{X}]$ of log-BMI with age, smoking status, physically active status, SNP1, SNP2, and the interaction between SNP1 and physical activity status as predictors. In addition, we estimated a regression model for the selection bias function with smoking status and SNP1 as predictors. Note that for simplicity, we did not adjust for the principal components of the genetic data in these analysis. We used the estimated effects, rounded to the third digit, as effect values in the simulations. We then employed a few variations. We now describe the sampling and generation of the simulated data, and then the different variations of the simulation study.

12.1 Data sampling and generation:

We simulated a super population of 15,000 individuals. Then sampled cases and controls from this population, based only on disease status. For each of 1000 simulations, the super population was simulated as follows:

- SNP1 and SNP2 were sampled with replacement from the true SNP data.
- Binary smoking status as well as physically active status were sampled from a binary distribution, with parameter p estimated from the diabetes data set (for simplicity, ignoring case-control sampling).
- Alcohol measures and age were sampled form the case-control study data, with replacement.
- Disease probability was calculated by the inverse of the logistic model with parameters as estimated from the data, with adaptation of the intercept to have disease prevalence of about 8.4%, and possible variation as described later.

- Log-BMI values were simulated from a normal distribution, using the mean and variance parameters estimated from the diabetes data set, with possible variations as described later.

We sampled 500 cases and 500 controls from the super population.

12.2 Variations of the simulation

To study the effect of some properties of the data on the estimators, we applied the following variations, so that the simulations were ran with all combinations of the following options:

1. SNP1 and SNP2 where either the SNPs with very low MAF used to estimate the model parameters, or other two SNPs with high MAF (closer to 50%).

2. The effect of SNP1 on disease was set to a 'high' effect of 1.3 (instead of -0.04).

3. The effect of SNP1 on the selection bias function was set to a 'high' effect of -1 (instead of -0.053).

12.3 Misspecification of the selection bias function

We studied the control function estimator when the selection bias function is correctly specified, and also when it is misspecified, in the following ways. Recall that a correct specification refers to a linear model with an intercept, SNP1, and smoking status. The effect sizes were:
$\alpha_{intercept} = -0.158$
$\alpha_{smoke} = 0.022$
$\alpha_{snp1} = -0.053$ or (if set to 'high') $\alpha_{snp1} = -0.2$.
We allowed for the following misspecifications of the selection bias function:

1. cont-mis1: no SNP1 effect (just intercept and smoking status).

2. cont-mis2: no smoking status effect (just intercept and SNP1).

3. cont-mis3: neither SNP1 nor smoking status (just intercept).

12.4 Conclusions

In the following, figures and tables provide the simulations results. The figures focus on the various control-function estimates, and IPW (which can also be thought of as type of control-function estimator with the selection bias misspecified and equal to zero), and compare between the bias and MSE of the SNP effects. The tables provide comprehensive simulation results for all measures and estimators used.

1. The control function estimator improves over IPW when the effect of SNP1 (or more generally, covariates or exposures) on either the disease model or the selection bias model is high, and it is in fact included in the disease/selection bias model. In other words, cont-mis2 performs better than cont-mis1 and cont-mis3, that do not include effects of SNP1. Also, its performance is almost identical to the cont-cor and better than the usual IPW.

2. The improvement seen in the control function estimator was in the effect (bias or MSE) estimate of SNP1 and the interaction SNP1 and being physically active. The various control function estimators (i.e. under the different forms of misspecification) had similar behavior with respect to the estimation of SNP2 effect.

3. The control function estimators were never worse than IPW in terms of MSE.

4. When the MAF of the SNPs was low (rare SNP), coverage probabilities of all estimators were reduced, compared to when the MAF was relatively high (common SNP).

12.5 Figures and tables summarizing the results

Figure 6: Comparison between the estimated bias of SNP1 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations in which SNP1 and SNP2 have either low or high MAF, the effect of SNP1 on the disease model is either low or high, and the effect of SNP1 on the selection bias model ($\gamma(\mathbf{X})$) is either low or high.

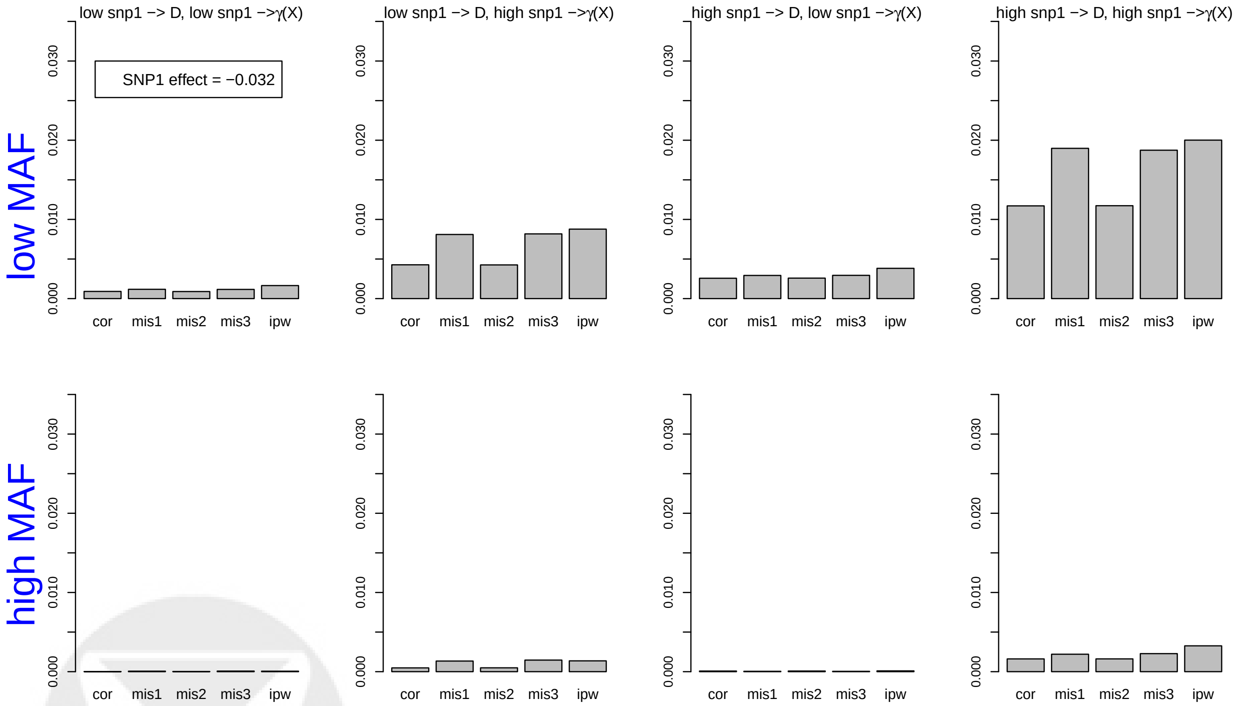


Figure 7: Comparison between the Mean Square Error (MSE) of SNP1 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations in which SNP1 and SNP2 have either low or high MAF, the effect of SNP1 on the disease model is either low or high, and the effect of SNP1 on the selection bias model ($\gamma(\mathbf{X})$) is either low or high.

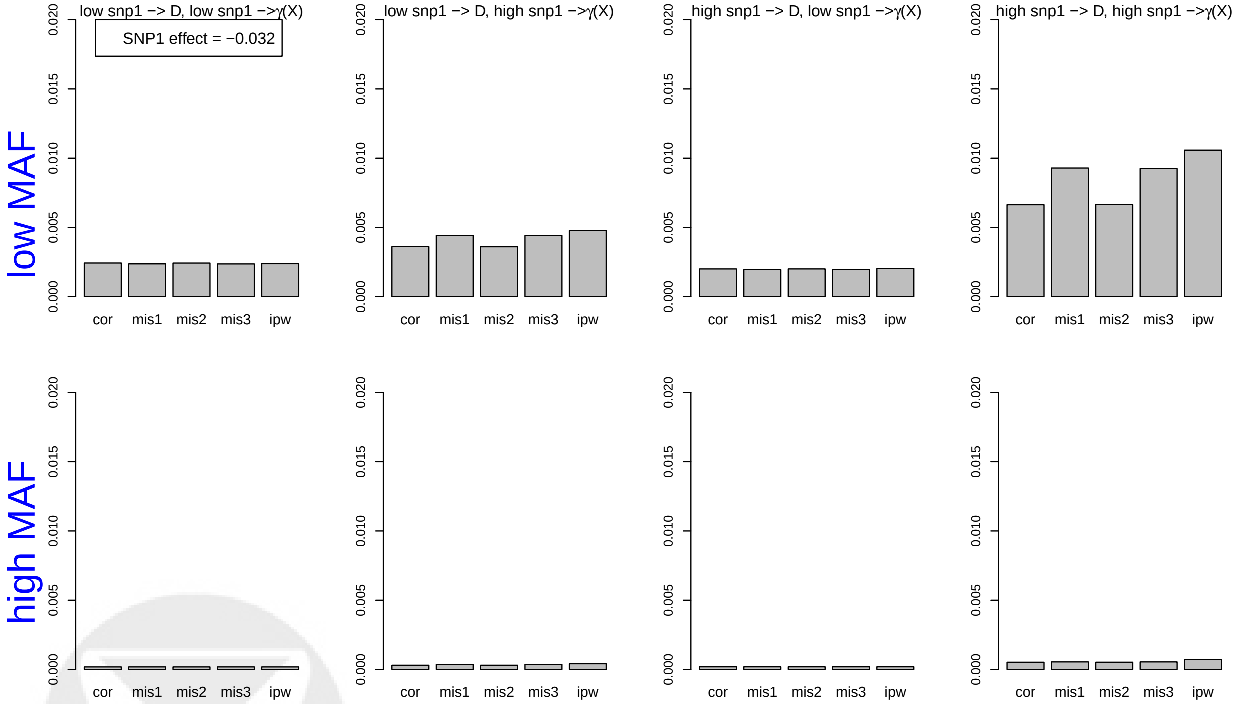


Figure 8: Comparison between the estimated bias of the effect of the interaction Active×SNP1 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations in which SNP1 and SNP2 have either low or high MAF, the effect of SNP1 on the disease model is either low or high, and the effect of SNP1 on the selection bias model ($\gamma(\mathbf{X})$) is either low or high.

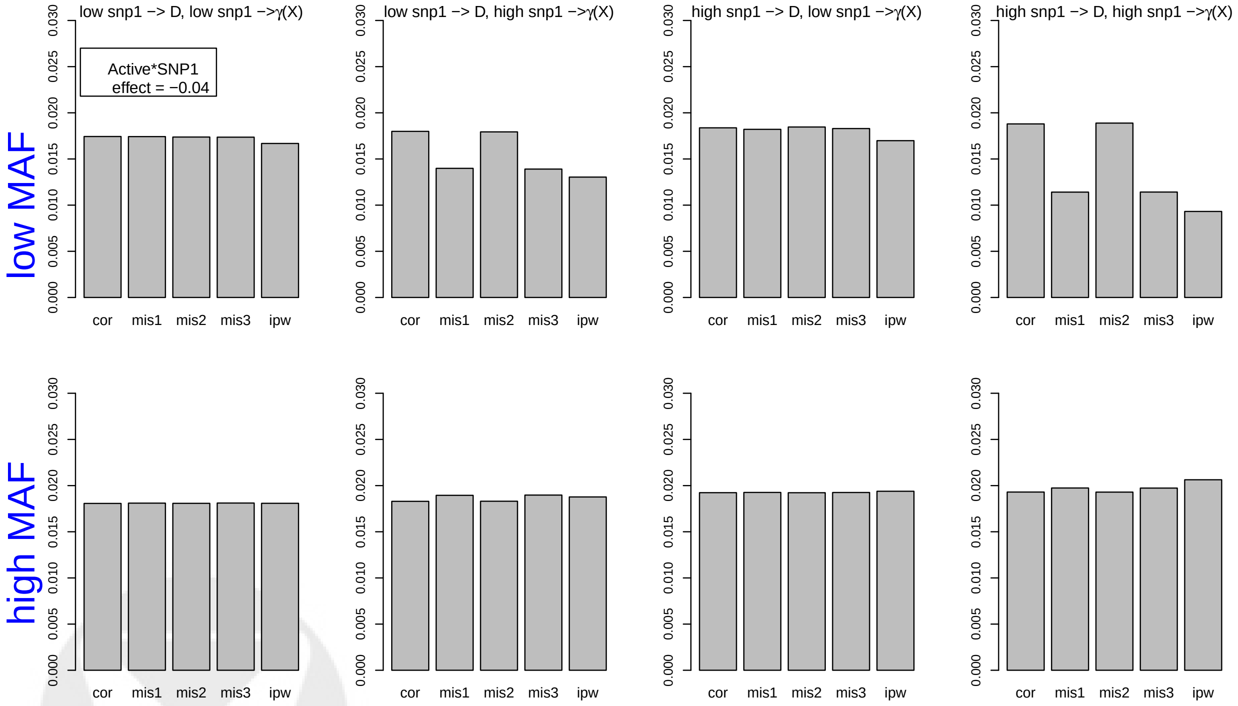


Figure 9: Comparison between the Mean Square Error (MSE) of the interaction Active×SNP1 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations in which SNP1 and SNP2 have either low or high MAF, the effect of SNP1 on the disease model is either low or high, and the effect of SNP1 on the selection bias model ($\gamma(\mathbf{X})$) is either low or high.

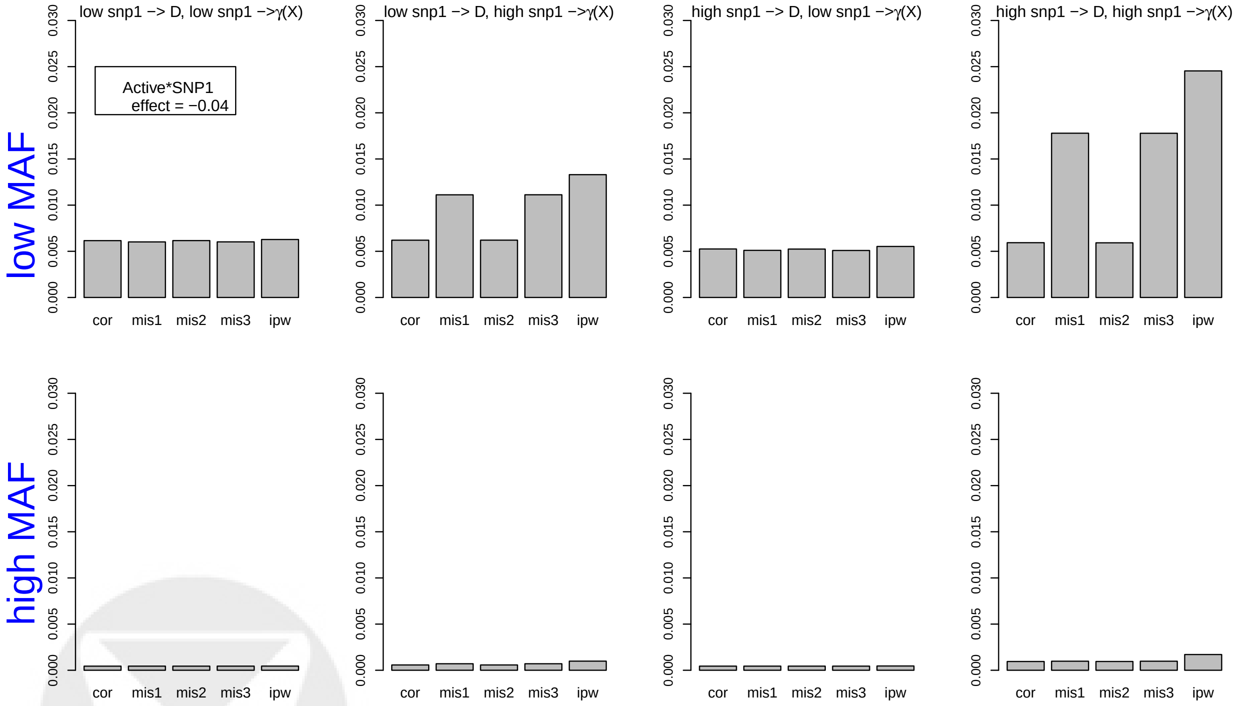


Figure 10: Comparison between the estimated bias of SNP2 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations in which SNP1 and SNP2 have either low or high MAF, the effect of SNP1 on the disease model is either low or high, and the effect of SNP1 on the selection bias model ($\gamma(\mathbf{X})$) is either low or high.

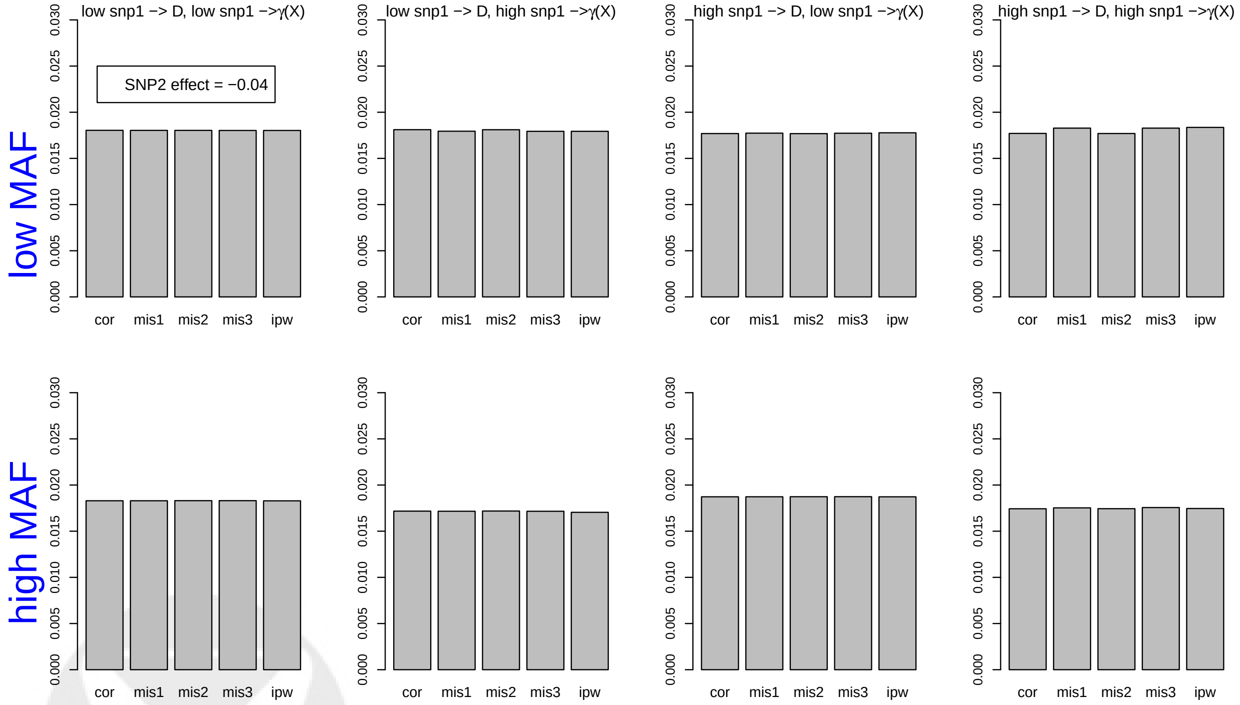


Figure 11: Comparison between the Mean Square Error (MSE) of SNP2 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations in which SNP1 and SNP2 have either low or high MAF, the effect of SNP1 on the disease model is either low or high, and the effect of SNP1 on the selection bias model ($\gamma(\mathbf{X})$) is either low or high.

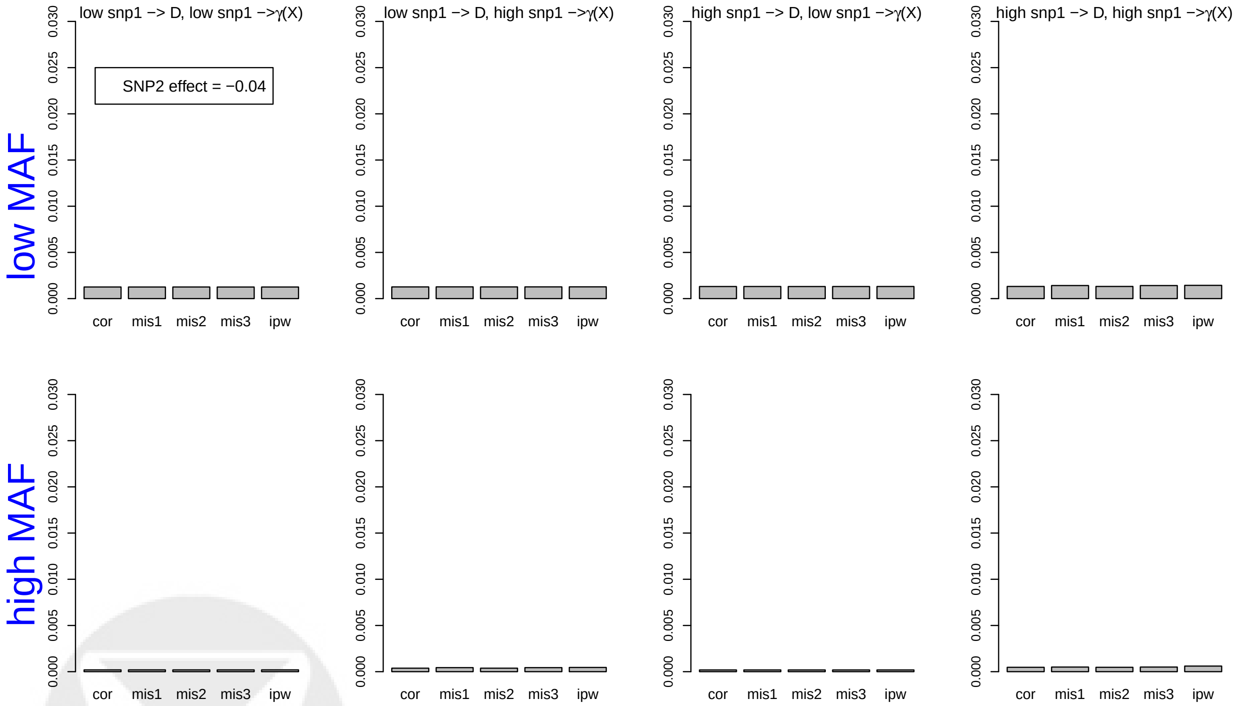


Table 6: Simulation results, averaged over 1000 simulations, for estimating the effect of covariates on a the simulated log(BMI) outcomes. The SNPs used had **low** MAF, the effect of SNP1 on the disease distribution was **low**, and its effect on the selection bias function was **low**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
ours-cor	0.001	0.002	0.047	0.048	0.952
ours-mis1	0.001	0.002	0.046	0.048	0.951
ours-mis2	0.001	0.002	0.046	0.048	0.951
ours-mis3	0.001	0.002	0.046	0.048	0.951
ipw	0.001	0.002	0.047	0.048	0.949
pooled	−0.061	0.005	0.040	0.040	0.667
dind	0.012	0.002	0.037	0.037	0.945
Age, $\beta_1 = 0.002$					
ours-cor	0.000	0.000	0.001	0.001	0.943
ours-mis1	0.000	0.000	0.001	0.001	0.944
ours-mis2	0.000	0.000	0.001	0.001	0.943
ours-mis3	0.000	0.000	0.001	0.001	0.943
ipw	0.000	0.000	0.001	0.001	0.939
pooled	0.000	0.000	0.001	0.001	0.949
dind	0.000	0.000	0.001	0.001	0.953
Smoker, $\beta_2 = -0.012$					
ours-cor	0.001	0.000	0.017	0.017	0.933
ours-mis1	0.001	0.000	0.017	0.017	0.932
ours-mis2	0.001	0.000	0.017	0.017	0.933
ours-mis3	0.001	0.000	0.017	0.017	0.934
ipw	0.001	0.000	0.017	0.017	0.932
pooled	−0.003	0.000	0.013	0.013	0.944
dind	0.018	0.000	0.012	0.012	0.684
Physically active, $\beta_3 = -0.032$					
ours-cor	0.000	0.000	0.015	0.015	0.942
ours-mis1	0.000	0.000	0.015	0.015	0.942
ours-mis2	0.000	0.000	0.015	0.015	0.942
ours-mis3	0.000	0.000	0.015	0.015	0.943
ipw	0.000	0.000	0.015	0.015	0.943
pooled	0.006	0.000	0.013	0.013	0.929
dind	−0.003	0.000	0.012	0.012	0.954
SNP1, $\beta_4 = -0.032$					
ours-cor	−0.001	0.002	0.049	0.045	0.916
ours-mis1	−0.001	0.002	0.049	0.045	0.910
ours-mis2	−0.001	0.002	0.049	0.045	0.915
ours-mis3	−0.001	0.002	0.049	0.045	0.913
ipw	−0.002	0.002	0.049	0.045	0.914
pooled	−0.027	0.002	0.040	0.039	0.883
dind	−0.023	0.002	0.035	0.036	0.902
SNP2, $\beta_5 = -0.040$					
ours-cor	0.001	0.001	0.035	0.034	0.931
ours-mis1	0.001	0.001	0.035	0.034	0.930
ours-mis2	0.001	0.001	0.035	0.034	0.930
ours-mis3	0.001	0.001	0.035	0.034	0.929
ipw	0.001	0.001	0.035	0.034	0.930
pooled	0.006	0.001	0.031	0.031	0.942
dind	−0.001	0.001	0.028	0.028	0.946
Active×SNP1, $\beta_6 = -0.021$					
ours-cor	−0.001	0.006	0.078	0.072	0.926
ours-mis1	−0.001	0.006	0.078	0.071	0.926
ours-mis2	−0.001	0.006	0.079	0.072	0.926
ours-mis3	−0.001	0.006	0.078	0.071	0.926
ipw	−0.002	0.006	0.079	0.071	0.926
pooled	0.001	0.004	0.067	0.063	0.940
dind	0.001	0.003	0.059	0.058	0.948

Table 7: Simulation results, averaged over 1000 simulations, for estimating the effect of covariates on a the simulated log(BMI) outcomes. The SNPs used had **high** MAF, the effect of SNP1 on the disease distribution was **low**, and its effect on the selection bias function was **low**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, β_0 = 3.077					
ours-cor	−0.001	0.003	0.056	0.055	0.943
ours-mis1	−0.001	0.003	0.056	0.055	0.942
ours-mis2	−0.001	0.003	0.056	0.055	0.943
ours-mis3	−0.001	0.003	0.056	0.055	0.942
ipw	−0.001	0.003	0.057	0.055	0.946
pooled	−0.069	0.007	0.050	0.049	0.699
dind	0.043	0.004	0.044	0.043	0.834
Age, β_1 = 0.002					
ours-cor	0.000	0.000	0.001	0.001	0.944
ours-mis1	0.000	0.000	0.001	0.001	0.944
ours-mis2	0.000	0.000	0.001	0.001	0.944
ours-mis3	0.000	0.000	0.001	0.001	0.944
ipw	0.000	0.000	0.001	0.001	0.941
pooled	0.000	0.000	0.001	0.001	0.939
dind	0.000	0.000	0.001	0.001	0.941
Smoker, β_2 = −0.012					
ours-cor	0.000	0.000	0.017	0.017	0.950
ours-mis1	0.000	0.000	0.017	0.017	0.953
ours-mis2	0.000	0.000	0.017	0.017	0.953
ours-mis3	0.000	0.000	0.017	0.017	0.953
ipw	0.000	0.000	0.017	0.017	0.949
pooled	−0.009	0.000	0.014	0.014	0.918
dind	0.021	0.001	0.013	0.012	0.613
Physically active, β_3 = −0.032					
ours-cor	0.001	0.001	0.027	0.027	0.953
ours-mis1	0.001	0.001	0.027	0.027	0.952
ours-mis2	0.001	0.001	0.027	0.027	0.952
ours-mis3	0.001	0.001	0.027	0.027	0.951
ipw	0.001	0.001	0.027	0.027	0.959
pooled	0.006	0.001	0.024	0.025	0.953
dind	−0.006	0.001	0.022	0.021	0.947
SNP1, β_4 = −0.032					
ours-cor	0.000	0.000	0.014	0.013	0.944
ours-mis1	0.000	0.000	0.014	0.013	0.944
ours-mis2	0.000	0.000	0.014	0.013	0.944
ours-mis3	0.000	0.000	0.014	0.013	0.944
ipw	0.000	0.000	0.014	0.013	0.942
pooled	−0.026	0.001	0.012	0.012	0.426
dind	−0.021	0.001	0.010	0.010	0.490
SNP2, β_5 = −0.040					
ours-cor	0.000	0.000	0.014	0.014	0.943
ours-mis1	0.000	0.000	0.014	0.014	0.943
ours-mis2	0.000	0.000	0.014	0.014	0.943
ours-mis3	0.000	0.000	0.014	0.014	0.944
ipw	0.001	0.000	0.014	0.014	0.945
pooled	0.006	0.000	0.012	0.012	0.919
dind	−0.003	0.000	0.010	0.010	0.937
Active×SNP1, β_6 = −0.021					
ours-cor	−0.001	0.000	0.021	0.021	0.949
ours-mis1	−0.001	0.000	0.021	0.021	0.949
ours-mis2	−0.001	0.000	0.021	0.021	0.949
ours-mis3	−0.001	0.000	0.021	0.021	0.949
ipw	−0.001	0.000	0.021	0.021	0.951
pooled	0.002	0.000	0.019	0.019	0.949
dind	0.002	0.000	0.016	0.016	0.936

Table 8: Simulation results, averaged over 1000 simulations, for estimating the effect of covariates on a the simulated log(BMI) outcomes. The SNPs used had **low** MAF, the effect of SNP1 on the disease distribution was **high**, and its effect on the selection bias function was **low**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
ours-cor	−0.003	0.002	0.049	0.048	0.947
ours-mis1	−0.003	0.002	0.049	0.048	0.948
ours-mis2	−0.003	0.002	0.049	0.048	0.947
ours-mis3	−0.003	0.002	0.049	0.048	0.948
ipw	−0.003	0.002	0.050	0.048	0.947
pooled	−0.064	0.006	0.040	0.040	0.650
dind	0.008	0.001	0.037	0.037	0.945
Age, $\beta_1 = 0.002$					
ours-cor	0.000	0.000	0.001	0.001	0.944
ours-mis1	0.000	0.000	0.001	0.001	0.945
ours-mis2	0.000	0.000	0.001	0.001	0.945
ours-mis3	0.000	0.000	0.001	0.001	0.946
ipw	0.000	0.000	0.001	0.001	0.939
pooled	0.000	0.000	0.001	0.001	0.945
dind	0.000	0.000	0.001	0.001	0.951
Smoker, $\beta_2 = -0.012$					
ours-cor	0.000	0.000	0.017	0.017	0.947
ours-mis1	0.000	0.000	0.017	0.017	0.947
ours-mis2	0.000	0.000	0.017	0.017	0.948
ours-mis3	0.000	0.000	0.017	0.017	0.947
ipw	0.000	0.000	0.017	0.017	0.944
pooled	−0.003	0.000	0.013	0.013	0.944
dind	0.018	0.000	0.012	0.012	0.690
Physically active, $\beta_3 = -0.032$					
ours-cor	0.000	0.000	0.015	0.015	0.945
ours-mis1	0.000	0.000	0.015	0.015	0.945
ours-mis2	0.000	0.000	0.015	0.015	0.945
ours-mis3	0.000	0.000	0.015	0.015	0.945
ipw	0.000	0.000	0.015	0.015	0.947
pooled	0.006	0.000	0.013	0.013	0.925
dind	−0.003	0.000	0.012	0.012	0.929
SNP1, $\beta_4 = -0.032$					
ours-cor	−0.003	0.002	0.045	0.043	0.924
ours-mis1	−0.003	0.002	0.044	0.041	0.921
ours-mis2	−0.003	0.002	0.045	0.043	0.926
ours-mis3	−0.003	0.002	0.044	0.041	0.918
ipw	−0.004	0.002	0.045	0.043	0.924
pooled	−0.044	0.003	0.028	0.029	0.676
dind	−0.002	0.001	0.026	0.027	0.951
SNP2, $\beta_5 = -0.040$					
ours-cor	0.001	0.001	0.036	0.035	0.927
ours-mis1	0.001	0.001	0.036	0.035	0.927
ours-mis2	0.001	0.001	0.036	0.035	0.927
ours-mis3	0.001	0.001	0.036	0.035	0.927
ipw	0.001	0.001	0.036	0.035	0.930
pooled	0.005	0.001	0.031	0.031	0.943
dind	−0.002	0.001	0.029	0.028	0.954
Active×SNP1, $\beta_6 = -0.021$					
ours-cor	0.000	0.005	0.073	0.067	0.908
ours-mis1	−0.001	0.005	0.071	0.064	0.907
ours-mis2	0.000	0.005	0.072	0.067	0.906
ours-mis3	0.000	0.005	0.071	0.064	0.906
ipw	−0.002	0.006	0.074	0.067	0.896
pooled	−0.003	0.002	0.047	0.047	0.946
dind	−0.001	0.002	0.043	0.044	0.954

Table 9: Simulation results, averaged over 1000 simulations, for estimating the effect of covariates on a the simulated log(BMI) outcomes. The SNPs used had **high** MAF, the effect of SNP1 on the disease distribution was **high**, and its effect on the selection bias function was **low**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
ours-cor	0.001	0.003	0.056	0.056	0.956
ours-mis1	0.001	0.003	0.056	0.056	0.956
ours-mis2	0.001	0.003	0.056	0.056	0.956
ours-mis3	0.001	0.003	0.056	0.056	0.956
ipw	0.001	0.003	0.056	0.056	0.954
pooled	−0.023	0.003	0.048	0.050	0.923
dind	0.017	0.002	0.042	0.043	0.939
Age, $\beta_1 = 0.002$					
ours-cor	0.000	0.000	0.001	0.001	0.953
ours-mis1	0.000	0.000	0.001	0.001	0.954
ours-mis2	0.000	0.000	0.001	0.001	0.953
ours-mis3	0.000	0.000	0.001	0.001	0.954
ipw	0.000	0.000	0.001	0.001	0.953
pooled	0.000	0.000	0.001	0.001	0.958
dind	0.000	0.000	0.001	0.001	0.967
Smoker, $\beta_2 = -0.012$					
ours-cor	0.000	0.000	0.017	0.017	0.945
ours-mis1	0.000	0.000	0.017	0.017	0.946
ours-mis2	0.000	0.000	0.017	0.017	0.945
ours-mis3	0.000	0.000	0.017	0.017	0.948
ipw	0.000	0.000	0.017	0.017	0.943
pooled	−0.003	0.000	0.014	0.014	0.950
dind	0.023	0.001	0.012	0.012	0.527
Physically active, $\beta_3 = -0.032$					
ours-cor	−0.001	0.001	0.027	0.028	0.948
ours-mis1	−0.001	0.001	0.027	0.028	0.949
ours-mis2	−0.001	0.001	0.027	0.028	0.947
ours-mis3	−0.001	0.001	0.027	0.028	0.949
ipw	−0.001	0.001	0.027	0.028	0.943
pooled	0.006	0.001	0.026	0.028	0.968
dind	−0.002	0.001	0.024	0.024	0.948
SNP1, $\beta_4 = -0.032$					
ours-cor	0.000	0.000	0.014	0.014	0.938
ours-mis1	0.000	0.000	0.014	0.013	0.939
ours-mis2	0.000	0.000	0.014	0.014	0.939
ours-mis3	0.000	0.000	0.014	0.013	0.938
ipw	0.000	0.000	0.014	0.014	0.940
pooled	−0.058	0.003	0.012	0.012	0.002
dind	0.006	0.000	0.011	0.011	0.919
SNP2, $\beta_5 = -0.040$					
ours-cor	0.000	0.000	0.014	0.014	0.951
ours-mis1	0.000	0.000	0.014	0.014	0.953
ours-mis2	0.000	0.000	0.014	0.014	0.954
ours-mis3	0.000	0.000	0.014	0.014	0.954
ipw	0.000	0.000	0.014	0.014	0.955
pooled	0.004	0.000	0.012	0.012	0.925
dind	−0.004	0.000	0.010	0.010	0.926
Active×SNP1, $\beta_6 = -0.021$					
ours-cor	0.000	0.000	0.021	0.021	0.948
ours-mis1	0.000	0.000	0.021	0.021	0.947
ours-mis2	0.000	0.000	0.021	0.021	0.946
ours-mis3	0.000	0.000	0.021	0.021	0.946
ipw	0.001	0.000	0.022	0.021	0.943
pooled	0.000	0.000	0.019	0.019	0.961
dind	−0.002	0.000	0.017	0.017	0.943

Table 10: Simulation results, averaged over 1000 simulations, for estimating the effect of covariates on a the simulated log(BMI) outcomes. The SNPs used had **low** MAF, the effect of SNP1 on the disease distribution was **low**, and its effect on the selection bias function was **high**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
ours-cor	0.001	0.002	0.047	0.048	0.952
ours-mis1	0.001	0.002	0.047	0.049	0.948
ours-mis2	0.001	0.002	0.047	0.048	0.952
ours-mis3	0.001	0.002	0.047	0.049	0.947
ipw	0.001	0.002	0.048	0.049	0.946
pooled	−0.060	0.006	0.047	0.047	0.753
dind	0.031	0.003	0.043	0.043	0.885
Age, $\beta_1 = 0.002$					
ours-cor	0.000	0.000	0.001	0.001	0.944
ours-mis1	0.000	0.000	0.001	0.001	0.949
ours-mis2	0.000	0.000	0.001	0.001	0.944
ours-mis3	0.000	0.000	0.001	0.001	0.949
ipw	0.000	0.000	0.001	0.001	0.944
pooled	0.000	0.000	0.001	0.001	0.944
dind	0.000	0.000	0.001	0.001	0.946
Smoker, $\beta_2 = -0.012$					
ours-cor	0.001	0.000	0.017	0.017	0.934
ours-mis1	0.001	0.000	0.017	0.017	0.942
ours-mis2	0.001	0.000	0.017	0.017	0.934
ours-mis3	0.001	0.000	0.017	0.017	0.943
ipw	0.001	0.000	0.017	0.017	0.940
pooled	−0.006	0.000	0.015	0.015	0.938
dind	0.020	0.001	0.014	0.014	0.694
Physically active, $\beta_3 = -0.032$					
ours-cor	0.000	0.000	0.015	0.015	0.942
ours-mis1	0.000	0.000	0.015	0.015	0.947
ours-mis2	0.000	0.000	0.015	0.015	0.942
ours-mis3	0.000	0.000	0.015	0.015	0.947
ipw	0.000	0.000	0.015	0.015	0.945
pooled	0.006	0.000	0.013	0.015	0.970
dind	−0.005	0.000	0.012	0.013	0.960
SNP1, $\beta_4 = -0.032$					
ours-cor	−0.004	0.004	0.060	0.064	0.966
ours-mis1	−0.008	0.004	0.066	0.055	0.900
ours-mis2	−0.004	0.004	0.060	0.064	0.965
ours-mis3	−0.008	0.004	0.066	0.055	0.902
ipw	−0.009	0.005	0.069	0.065	0.945
pooled	−0.445	0.212	0.121	0.045	0.004
dind	−0.440	0.204	0.104	0.041	0.001
SNP2, $\beta_5 = -0.040$					
ours-cor	0.001	0.001	0.036	0.035	0.932
ours-mis1	0.001	0.001	0.036	0.035	0.935
ours-mis2	0.001	0.001	0.036	0.035	0.933
ours-mis3	0.001	0.001	0.036	0.035	0.936
ipw	0.001	0.001	0.036	0.035	0.934
pooled	0.008	0.001	0.035	0.035	0.943
dind	−0.001	0.001	0.032	0.032	0.947
Active×SNP1, $\beta_6 = -0.021$					
ours-cor	−0.001	0.006	0.079	0.098	0.975
ours-mis1	−0.005	0.011	0.105	0.087	0.918
ours-mis2	−0.001	0.006	0.079	0.098	0.978
ours-mis3	−0.005	0.011	0.105	0.087	0.918
ipw	−0.006	0.013	0.115	0.101	0.941
pooled	0.034	0.041	0.200	0.073	0.497
dind	0.035	0.031	0.172	0.066	0.509

Table 11: Simulation results, averaged over 1000 simulations, for estimating the effect of covariates on a the simulated log(BMI) outcomes. The SNPs used had **high** MAF, the effect of SNP1 on the disease distribution was **low**, and its effect on the selection bias function was **high**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
ours-cor	−0.003	0.004	0.061	0.073	0.984
ours-mis1	−0.003	0.005	0.068	0.081	0.981
ours-mis2	−0.003	0.004	0.061	0.073	0.984
ours-mis3	−0.003	0.005	0.068	0.081	0.982
ipw	−0.005	0.007	0.082	0.079	0.927
pooled	−0.092	0.039	0.175	0.175	0.904
dind	0.574	0.340	0.101	0.094	0.000
Age, $\beta_1 = 0.002$					
ours-cor	0.000	0.000	0.001	0.001	0.985
ours-mis1	0.000	0.000	0.001	0.002	0.980
ours-mis2	0.000	0.000	0.001	0.001	0.985
ours-mis3	0.000	0.000	0.001	0.002	0.980
ipw	0.000	0.000	0.002	0.002	0.937
pooled	0.000	0.000	0.004	0.003	0.938
dind	0.000	0.000	0.002	0.002	0.931
Smoker, $\beta_2 = -0.012$					
ours-cor	−0.001	0.001	0.023	0.024	0.964
ours-mis1	−0.001	0.001	0.025	0.026	0.959
ours-mis2	−0.001	0.001	0.023	0.025	0.968
ours-mis3	−0.001	0.001	0.025	0.026	0.964
ipw	−0.001	0.001	0.025	0.027	0.958
pooled	−0.100	0.012	0.049	0.051	0.502
dind	0.075	0.006	0.028	0.027	0.224
Physically active, $\beta_3 = -0.032$					
ours-cor	0.001	0.001	0.027	0.029	0.967
ours-mis1	0.000	0.001	0.031	0.046	0.996
ours-mis2	0.001	0.001	0.027	0.029	0.967
ours-mis3	0.000	0.001	0.031	0.046	0.996
ipw	0.000	0.001	0.031	0.031	0.958
pooled	0.004	0.003	0.055	0.088	0.995
dind	−0.067	0.008	0.062	0.047	0.653
SNP1, $\beta_4 = -0.032$					
ours-cor	0.000	0.000	0.017	0.019	0.968
ours-mis1	−0.001	0.000	0.019	0.021	0.967
ours-mis2	0.000	0.000	0.017	0.019	0.968
ours-mis3	−0.001	0.000	0.019	0.021	0.967
ipw	−0.001	0.000	0.020	0.021	0.961
pooled	−0.442	0.197	0.041	0.042	0.000
dind	−0.411	0.170	0.026	0.022	0.000
SNP2, $\beta_5 = -0.040$					
ours-cor	0.002	0.000	0.019	0.019	0.940
ours-mis1	0.002	0.000	0.021	0.020	0.945
ours-mis2	0.002	0.000	0.019	0.019	0.940
ours-mis3	0.002	0.000	0.021	0.020	0.945
ipw	0.002	0.000	0.021	0.021	0.944
pooled	0.037	0.003	0.042	0.042	0.867
dind	−0.020	0.001	0.023	0.023	0.861
Active×SNP1, $\beta_6 = -0.021$					
ours-cor	0.000	0.001	0.024	0.029	0.984
ours-mis1	0.000	0.001	0.027	0.031	0.983
ours-mis2	0.000	0.001	0.024	0.029	0.984
ours-mis3	0.000	0.001	0.027	0.031	0.983
ipw	0.000	0.001	0.031	0.031	0.953
pooled	0.039	0.006	0.070	0.068	0.893
dind	0.039	0.004	0.049	0.036	0.735

Table 12: Simulation results, averaged over 1000 simulations, for estimating the effect of covariates on a the simulated log(BMI) outcomes. The SNPs used had **low** MAF, the effect of SNP1 on the disease distribution was **high**, and its effect on the selection bias function was **high**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, β_0 = 3.077					
ours-cor	−0.003	0.002	0.050	0.048	0.947
ours-mis1	−0.003	0.003	0.051	0.050	0.941
ours-mis2	−0.003	0.002	0.050	0.048	0.947
ours-mis3	−0.003	0.003	0.051	0.050	0.943
ipw	−0.003	0.003	0.052	0.050	0.935
pooled	−0.064	0.006	0.049	0.048	0.729
dind	0.031	0.003	0.044	0.044	0.885
Age, β_1 = 0.002					
ours-cor	0.000	0.000	0.001	0.001	0.945
ours-mis1	0.000	0.000	0.001	0.001	0.944
ours-mis2	0.000	0.000	0.001	0.001	0.945
ours-mis3	0.000	0.000	0.001	0.001	0.942
ipw	0.000	0.000	0.001	0.001	0.934
pooled	0.000	0.000	0.001	0.001	0.947
dind	0.000	0.000	0.001	0.001	0.952
Smoker, β_2 = −0.012					
ours-cor	0.000	0.000	0.017	0.017	0.946
ours-mis1	0.000	0.000	0.018	0.018	0.948
ours-mis2	0.000	0.000	0.017	0.017	0.948
ours-mis3	0.000	0.000	0.018	0.018	0.949
ipw	0.000	0.000	0.018	0.018	0.949
pooled	−0.002	0.000	0.016	0.016	0.950
dind	0.026	0.001	0.014	0.015	0.568
Physically active, β_3 = −0.032					
ours-cor	0.000	0.000	0.015	0.015	0.945
ours-mis1	0.000	0.000	0.015	0.015	0.946
ours-mis2	0.000	0.000	0.015	0.015	0.945
ours-mis3	0.000	0.000	0.015	0.015	0.947
ipw	0.000	0.000	0.015	0.015	0.945
pooled	0.006	0.000	0.013	0.016	0.967
dind	−0.006	0.000	0.013	0.014	0.952
SNP1, β_4 = −0.032					
ours-cor	−0.012	0.007	0.081	0.093	0.975
ours-mis1	−0.019	0.009	0.095	0.067	0.839
ours-mis2	−0.012	0.007	0.081	0.093	0.975
ours-mis3	−0.019	0.009	0.094	0.068	0.842
ipw	−0.020	0.011	0.101	0.096	0.954
pooled	−0.529	0.286	0.073	0.035	0.000
dind	−0.474	0.228	0.063	0.032	0.000
SNP2, β_5 = −0.040					
ours-cor	0.001	0.001	0.036	0.035	0.931
ours-mis1	0.001	0.001	0.038	0.036	0.935
ours-mis2	0.001	0.001	0.036	0.035	0.928
ours-mis3	0.001	0.001	0.038	0.036	0.935
ipw	0.000	0.001	0.038	0.036	0.935
pooled	0.005	0.001	0.038	0.037	0.948
dind	−0.004	0.001	0.035	0.033	0.944
Active×SNP1, β_6 = −0.021					
ours-cor	0.000	0.006	0.077	0.141	0.999
ours-mis1	−0.007	0.018	0.133	0.106	0.885
ours-mis2	0.000	0.006	0.077	0.141	0.999
ours-mis3	−0.007	0.018	0.133	0.106	0.889
ipw	−0.009	0.025	0.156	0.147	0.944
pooled	−0.008	0.016	0.127	0.057	0.625
dind	−0.004	0.012	0.108	0.051	0.659

Table 13: Simulation results, averaged over 1000 simulations, for estimating the effect of covariates on a the simulated log(BMI) outcomes. The SNPs used had **high** MAF, the effect of SNP1 on the disease distribution was **high**, and its effect on the selection bias function was **high**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
ours-cor	0.000	0.004	0.065	0.084	0.989
ours-mis1	0.001	0.005	0.068	0.088	0.988
ours-mis2	0.000	0.004	0.065	0.084	0.989
ours-mis3	0.001	0.005	0.068	0.088	0.988
ipw	−0.001	0.008	0.092	0.094	0.951
pooled	0.077	0.037	0.176	0.186	0.941
dind	0.351	0.130	0.083	0.083	0.019
Age, $\beta_1 = 0.002$					
ours-cor	0.000	0.000	0.001	0.002	0.996
ours-mis1	0.000	0.000	0.001	0.002	0.994
ours-mis2	0.000	0.000	0.001	0.002	0.996
ours-mis3	0.000	0.000	0.001	0.002	0.995
ipw	0.000	0.000	0.002	0.002	0.954
pooled	0.000	0.000	0.004	0.004	0.952
dind	0.000	0.000	0.002	0.002	0.960
Smoker, $\beta_2 = -0.012$					
ours-cor	0.000	0.001	0.029	0.028	0.937
ours-mis1	0.000	0.001	0.031	0.029	0.930
ours-mis2	0.000	0.001	0.029	0.029	0.938
ours-mis3	0.000	0.001	0.030	0.030	0.940
ipw	−0.001	0.001	0.034	0.032	0.930
pooled	−0.057	0.006	0.054	0.053	0.802
dind	0.117	0.014	0.025	0.024	0.003
Physically active, $\beta_3 = -0.032$					
ours-cor	−0.001	0.001	0.028	0.033	0.974
ours-mis1	−0.001	0.001	0.029	0.038	0.987
ours-mis2	−0.001	0.001	0.028	0.033	0.974
ours-mis3	−0.001	0.001	0.029	0.038	0.987
ipw	−0.002	0.001	0.034	0.035	0.948
pooled	0.017	0.004	0.065	0.106	0.997
dind	−0.040	0.005	0.059	0.047	0.818
SNP1, $\beta_4 = -0.032$					
ours-cor	−0.002	0.001	0.023	0.027	0.978
ours-mis1	−0.002	0.001	0.023	0.021	0.937
ours-mis2	−0.002	0.001	0.023	0.027	0.978
ours-mis3	−0.002	0.001	0.023	0.021	0.937
ipw	−0.003	0.001	0.027	0.028	0.955
pooled	−0.596	0.356	0.038	0.045	0.000
dind	−0.168	0.029	0.027	0.021	0.000
SNP2, $\beta_5 = -0.040$					
ours-cor	0.001	0.000	0.022	0.022	0.970
ours-mis1	0.001	0.000	0.022	0.023	0.968
ours-mis2	0.001	0.000	0.022	0.022	0.970
ours-mis3	0.001	0.000	0.022	0.023	0.969
ipw	0.001	0.001	0.025	0.025	0.959
pooled	0.025	0.002	0.043	0.045	0.927
dind	−0.033	0.002	0.021	0.020	0.614
Active×SNP1, $\beta_6 = -0.021$					
ours-cor	0.001	0.001	0.031	0.040	0.991
ours-mis1	0.001	0.001	0.031	0.032	0.953
ours-mis2	0.001	0.001	0.031	0.040	0.991
ours-mis3	0.001	0.001	0.031	0.032	0.953
ipw	0.002	0.002	0.041	0.042	0.954
pooled	0.012	0.004	0.064	0.072	0.967
dind	0.001	0.002	0.039	0.032	0.892