0.4pt=0pt

CONTROL FUNCTION ASSISTED IPW ESTIMATION WITH A SECONDARY OUTCOME IN CASE-CONTROL STUDIES

Tamar Sofer, Marilyn C. Cornelis, Peter Kraft and Eric J. Tchetgen Tchetgen

University of Washington and Harvard School of Public Health

Supplementary Material

This note provides the mathematical derivations, and extensive simulation studies.

S1 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.

S2 Mathematical derivations.

S2.1 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:

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

Recall the identity

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

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:

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

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.

S2.2 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

1. *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.

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

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

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

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

   We show that

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

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

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

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

Proof of the theorem. Recall

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

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:

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

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

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

Consider the second argument.

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

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

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

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

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

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 $\Pi\big(v\big|\mathbf{U}\big)$. We can decompose

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

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

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

It follows that

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

To complete, it suffices to note that

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

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

$$\frac{Sh_1(\mathbf{X})}{\pi(D)}\bigg[Y-g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\bigg]-\frac{S}{\pi(D)}h_2(\mathbf{X},D)+\Pi\bigg(\frac{S}{\pi(D)}h_2(\mathbf{X},D)\bigg|\Lambda_{D,sub}\bigg). \qquad \square$$

S2.3 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.

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

so that the influence function has to be

$$\frac{Sh_1(\mathbf{X})}{\pi(D)}\bigg[Y-g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\bigg],$$

the IPW influence function. $\square$

S2.4 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:

$$\begin{aligned}\psi(\boldsymbol{\beta}) &= \frac{Sh_1(\mathbf{X})}{\pi(D)}\Big[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\Big] - \Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big|\Lambda_{D,sub}^{\perp}\cap\Lambda_{D,npar}\Big)\\ &= \Pi\Big(\frac{Sh_1(\mathbf{X})}{\pi(D)}\Big[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\Big]\Big|\Lambda_{D,sub}^{\perp}\cap\Lambda_{D,npar}\Big) - \Pi\Big(\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big|\Lambda_{D,sub}^{\perp}\cap\Lambda_{D,npar}\Big)\\ &\quad +\Pi\Big(\frac{Sh_1(\mathbf{X})}{\pi(D)}\Big[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}\Big]\Big|\Lambda_{D,sub}\Big),\end{aligned}$$

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:

$$\begin{aligned}0 &= \mathbb{E}\Big\{\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big(\frac{Sh_1(\mathbf{X})}{\pi(D)}[Y - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - \frac{S}{\pi(D)}h_2^{opt}(\mathbf{X},D)\Big)\Big\}\\ &= \mathbb{E}\Big\{\frac{S}{\pi(D)}h_2(\mathbf{X},D)\Big(\frac{Sh_1(\mathbf{X})}{\pi(D)}[\mathbb{E}(Y|\mathbf{X},D) - g^{-1}\{\mu(\mathbf{X};\boldsymbol{\beta})\}] - \frac{S}{\pi(D)}h_2^{opt}(\mathbf{X},D)\Big)\Big\}.\end{aligned}$$

This equality is satisfied by

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

as required. □

S2.5 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 $h_1(\mathbf{X})$:

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

Then:

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

This equation is satisfied by:

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

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

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

as required. □

S2.6 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:

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

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

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

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

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

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

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

Notice that

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

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

S2.7 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 (**?**). □

S3 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

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

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:

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

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.

S3.1 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).

S3.2 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

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

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

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

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

S3.3 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

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

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

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

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

S4 Identity link simulations - additional information

S4.1 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 **?**, 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

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

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 1 provides comprehensive simulation results (i.e. all summary statistics for all estimators under investigations), while Figure 1 provides graphical results, comparing the bias, MSE and coverage of the unbiased estimators cont-mis, cont-cor, IPW and TT-cor.

S4.2 Table summarizing the identity link simulations provided in Section 4.1 in the manuscript

The following Table 2 provide comprehensive simulation results for the simulation study described in Section 4.1 in the paper.

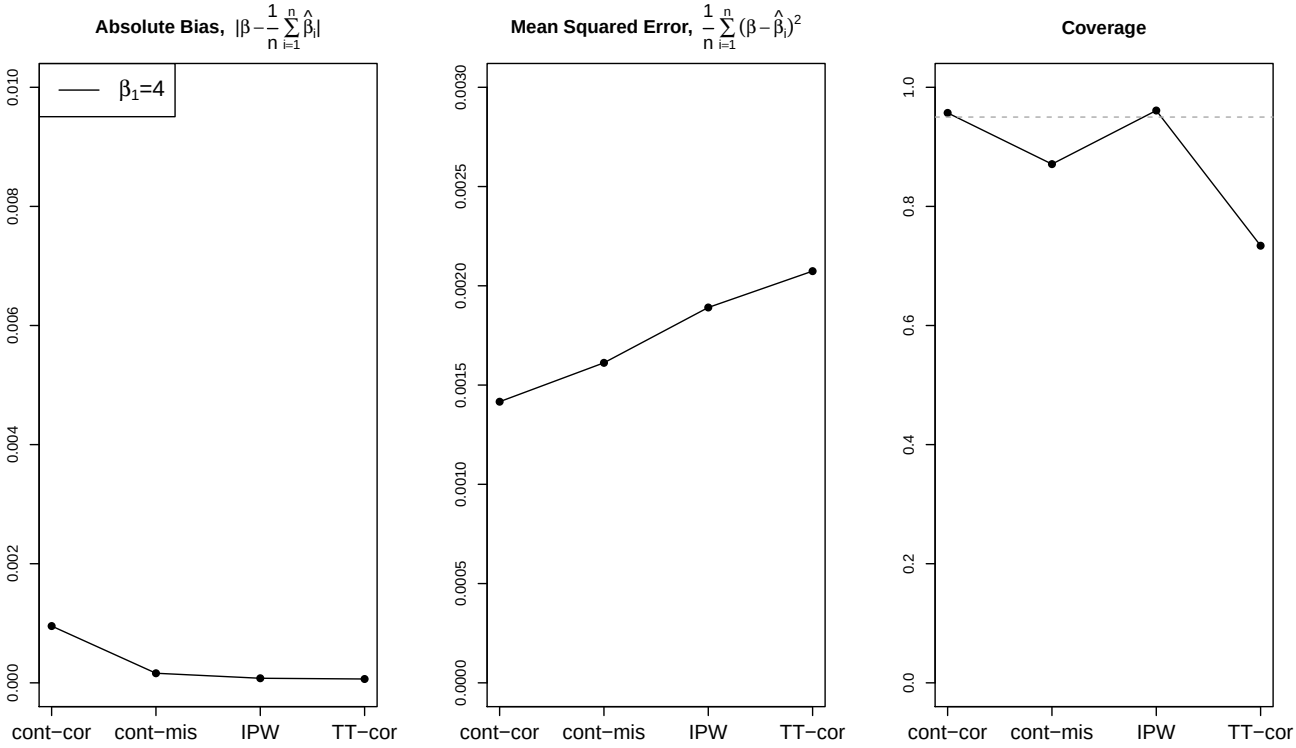


Figure 1: Results from Identity link simulations in the simple settings with a single covariate. Estimated bias, MSE, and coverage probability of the control function under correct and misspecification of the selection bias function (cont-cor, cont-mis, respectively), IPW and TT (correctly specified) estimators, in estimating the population effect of X_1.

Table 1: 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 **?** (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 2: 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

S5 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

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

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

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

so that the population mean is $\exp\left\{\mu(\mathbf{X}, \boldsymbol{\beta})\right\} = \exp(\mathbf{X}^T\boldsymbol{\beta})$ 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 2, provides the bias, MSE and coverage probabilities of the IPW and the control function estimators, calculated over the 1000 simulations. Table 8 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.

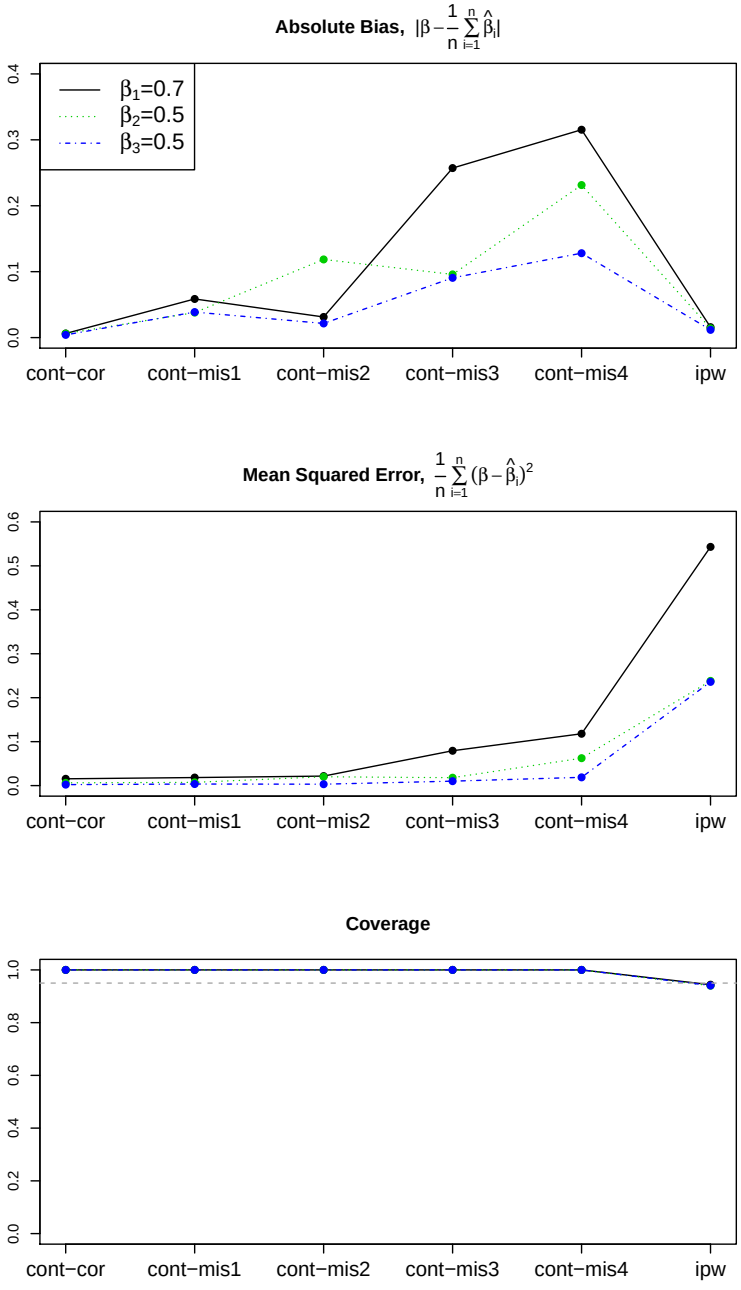


Figure 2: Results from log link simulations. Estimated bias, MSE, and coverage probability of the control function under correct and misspecification of the selection bias function (cont-cor, cont-mis1, ..., cont-mis4), and IPW, in estimating population means.

Table 3: 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 $\nu(\mathbf{X}, D)$ 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, $\beta_0 = 3$		
cont-cor	−0.009	0.023	0.151	0.645	1.000
cont-mis1	0.057	0.025	0.148	0.642	1.000
cont-mis2	−0.181	0.059	0.163	0.637	1.000
cont-mis3	−0.272	0.101	0.165	0.638	1.000
cont-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, $\beta_1 = 0.7$		
cont-cor	0.006	0.015	0.124	0.641	1.000
cont-mis1	−0.059	0.018	0.122	0.639	1.000
cont-mis2	0.031	0.022	0.144	0.635	1.000
cont-mis3	0.257	0.079	0.115	0.627	1.000
cont-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, $\beta_2 = 0.5$		
cont-cor	0.006	0.006	0.079	0.424	1.000
cont-mis1	−0.038	0.007	0.077	0.423	1.000
cont-mis2	0.119	0.020	0.078	0.416	1.000
cont-mis3	0.096	0.018	0.094	0.421	1.000
cont-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, $\beta_3 = 0.5$		
cont-cor	−0.004	0.002	0.047	0.422	1.000
cont-mis1	0.039	0.004	0.046	0.420	1.000
cont-mis2	−0.021	0.003	0.052	0.415	1.000
cont-mis3	−0.091	0.010	0.044	0.413	1.000
cont-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

S6 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.

S6.1 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.

S6.2 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).

S6.3 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).

S6.4 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).

S6.5 Figures and tables summarizing the results

Figure 3: 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.

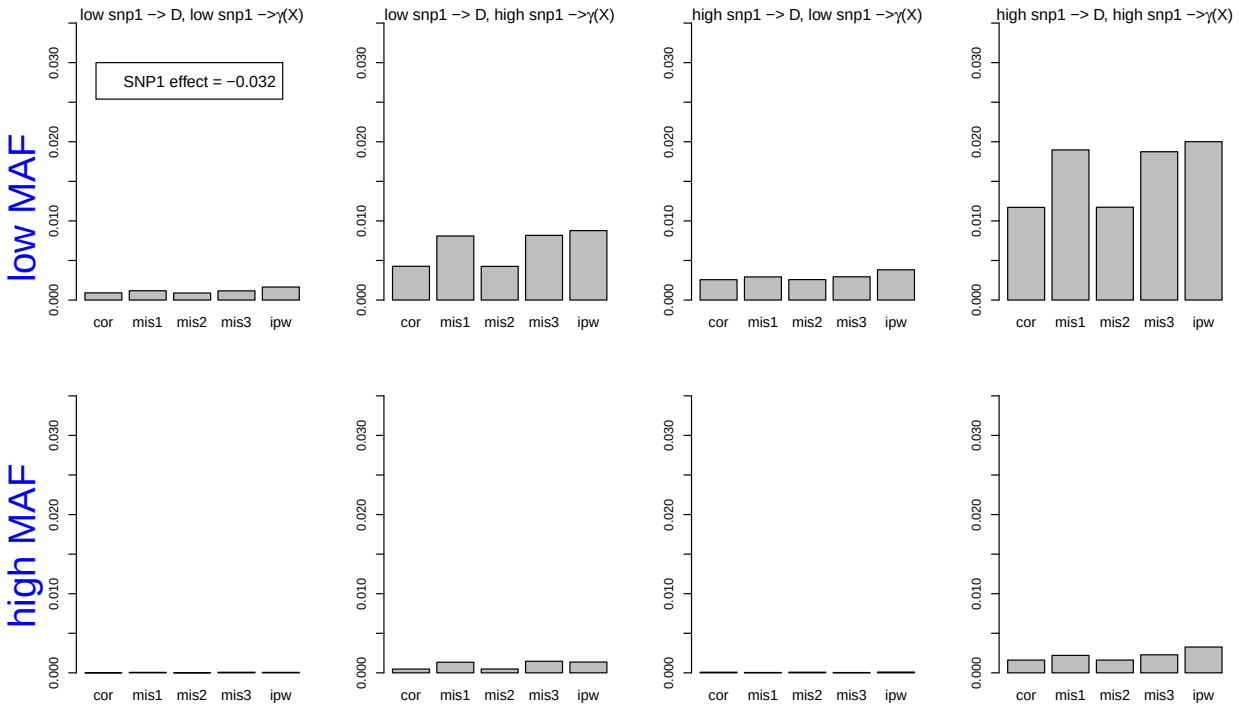


Figure 4: Comparison between the Mean Square Error (MSE) of SNP1 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations 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.

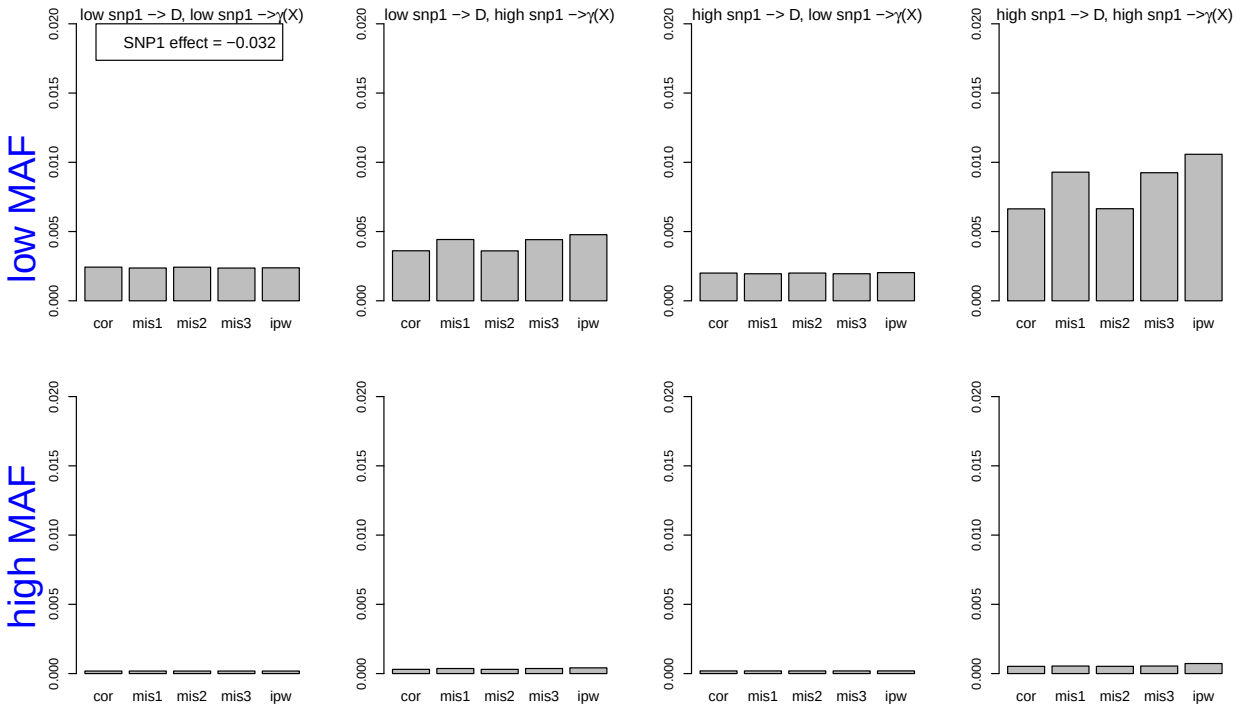


Figure 5: Comparison between the estimated bias of the effect of the interaction Active×SNP1 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. 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.

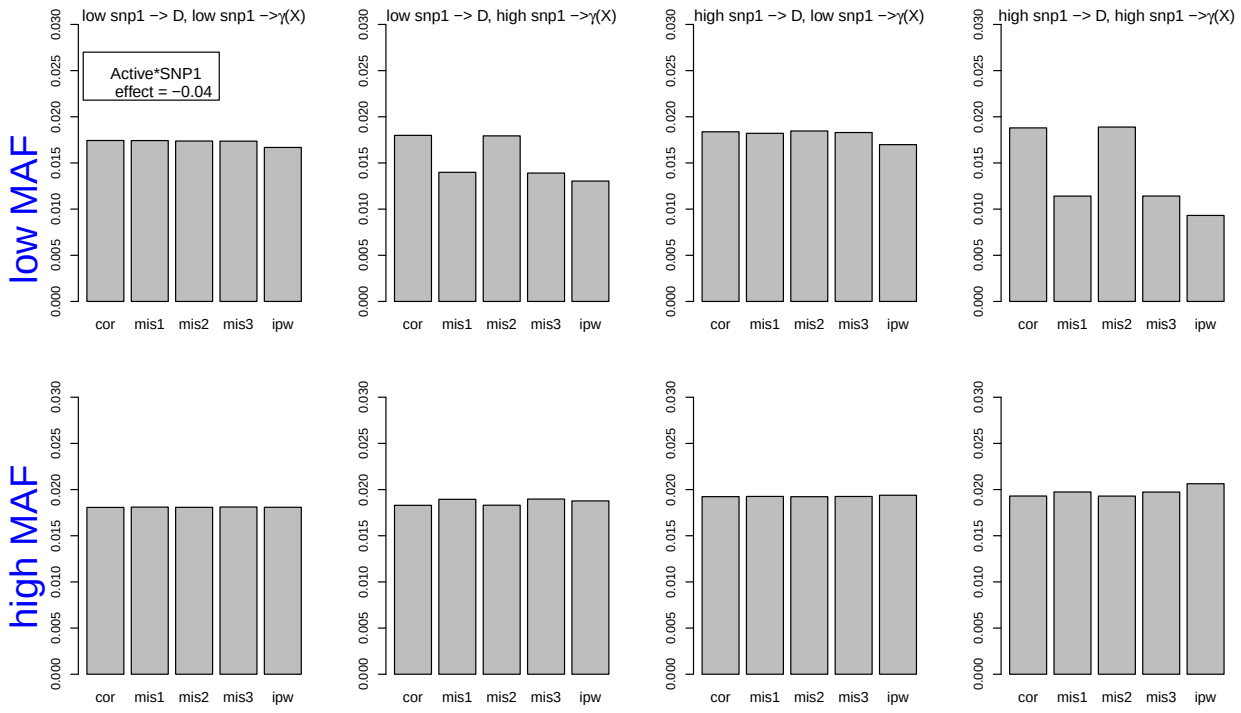


Figure 6: Comparison between the Mean Square Error (MSE) of the interaction Active×SNP1 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare 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.

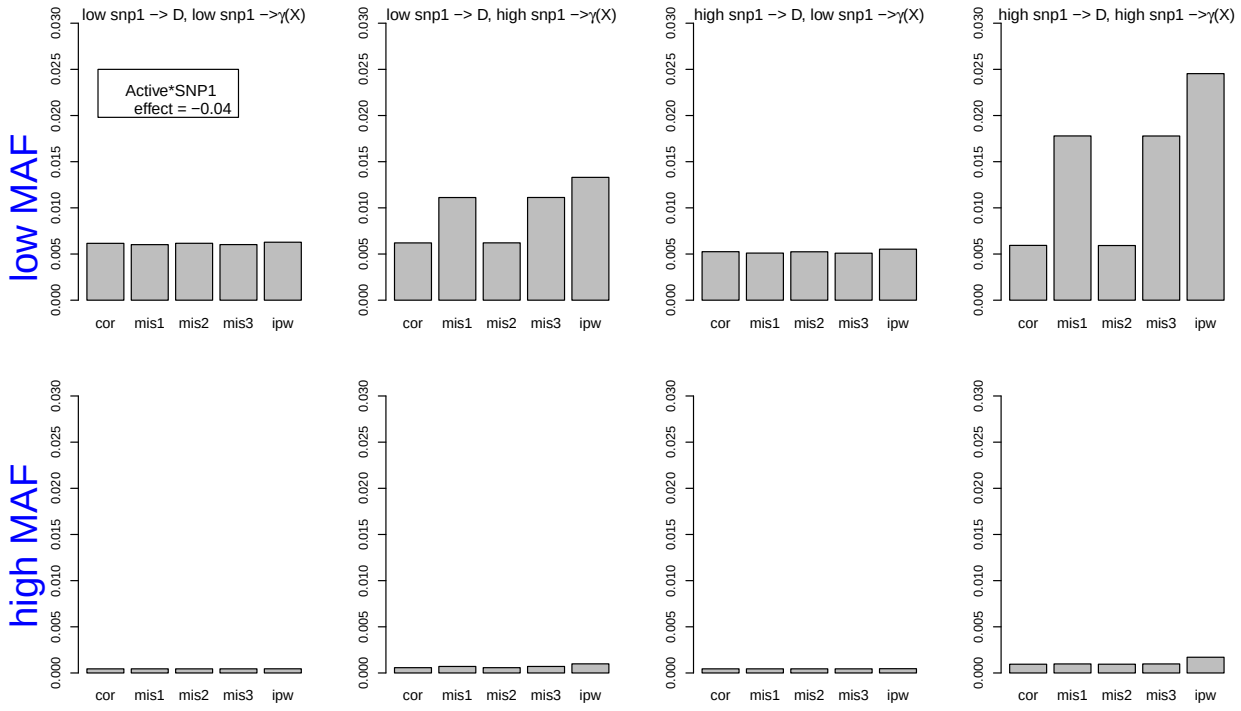


Figure 7: Comparison between the estimated bias of SNP2 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations 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.

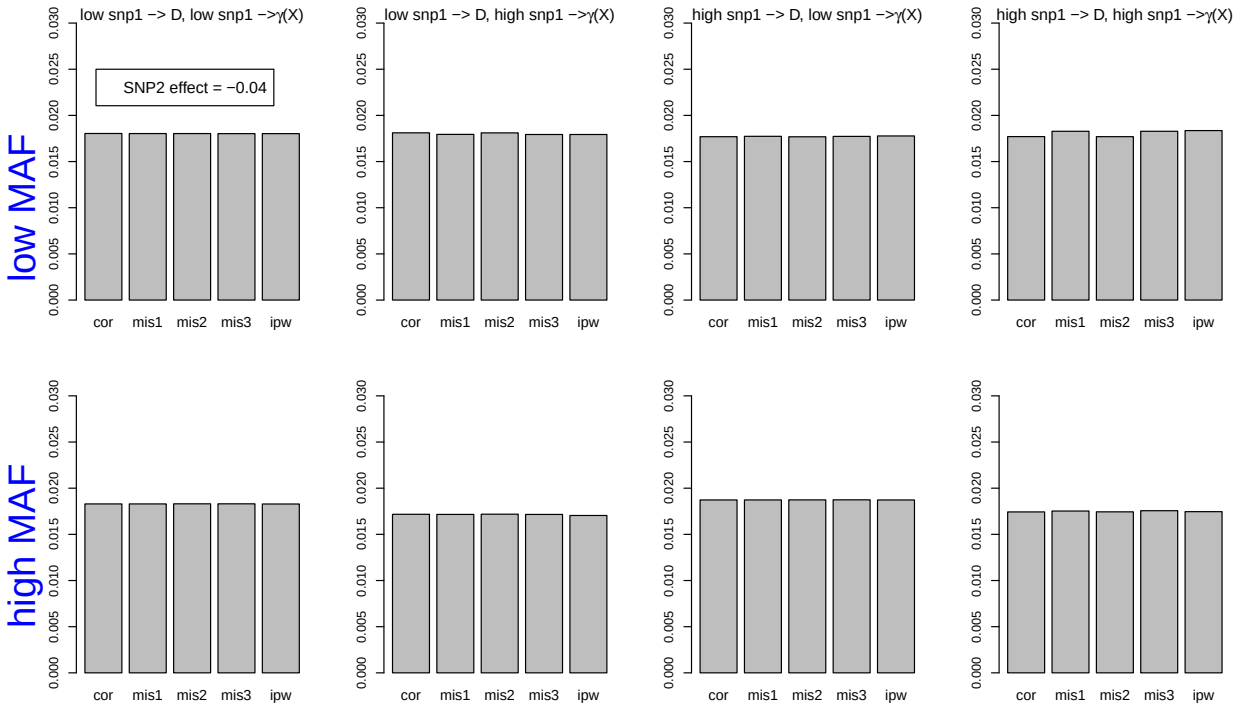
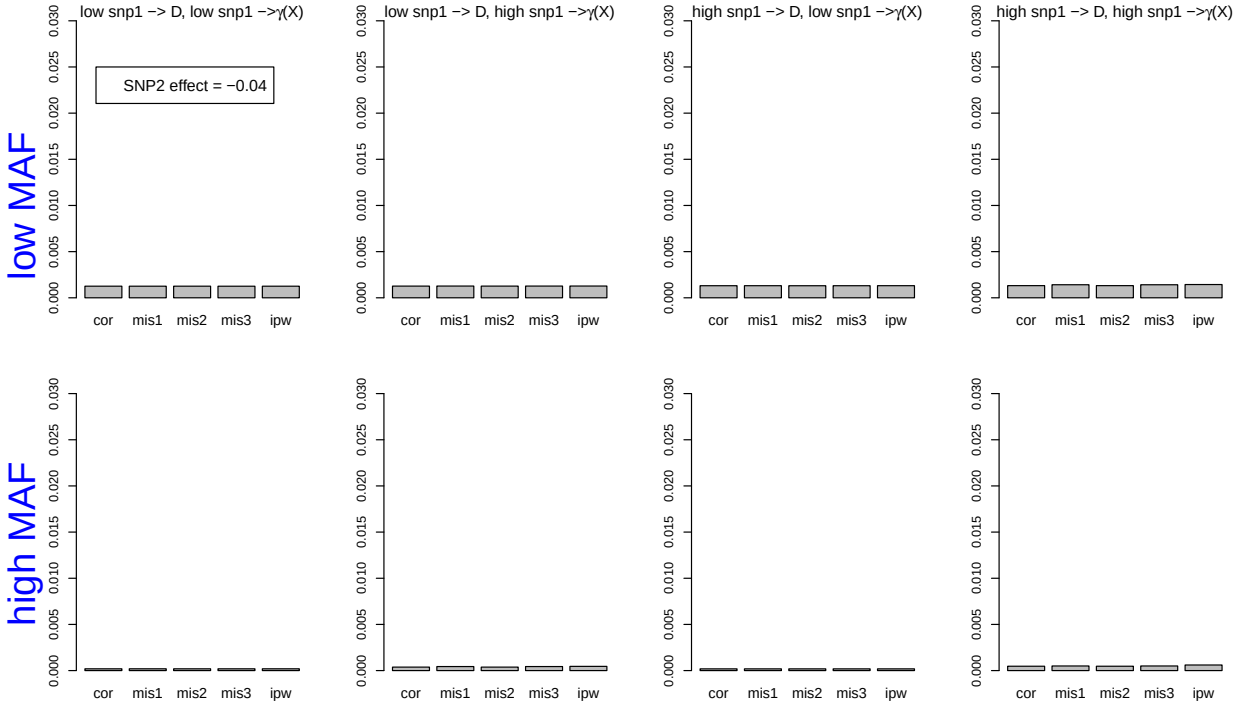


Figure 8: Comparison between the Mean Square Error (MSE) of SNP2 effect, over 1000 simulations, of the control-function estimator under various forms of mispecification (mis1, mis2, mis3) and under correct specification (cor) of the selection bias function, and of the IPW. We compare between all combinations 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.



Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
cont-cor	0.001	0.002	0.047	0.048	0.952
cont-mis1	0.001	0.002	0.046	0.048	0.951
cont-mis2	0.001	0.002	0.046	0.048	0.951
cont-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, $\beta_1 = 0.002$					
cont-cor	0.000	0.000	0.001	0.001	0.943
cont-mis1	0.000	0.000	0.001	0.001	0.944
cont-mis2	0.000	0.000	0.001	0.001	0.943
cont-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, $\beta_2 = -0.012$					
cont-cor	0.001	0.000	0.017	0.017	0.933
cont-mis1	0.001	0.000	0.017	0.017	0.932
cont-mis2	0.001	0.000	0.017	0.017	0.933
cont-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, $\beta_3 = -0.032$					
cont-cor	0.000	0.000	0.015	0.015	0.942
cont-mis1	0.000	0.000	0.015	0.015	0.942
cont-mis2	0.000	0.000	0.015	0.015	0.942
cont-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, $\beta_4 = -0.032$					
cont-cor	−0.001	0.002	0.049	0.045	0.916
cont-mis1	−0.001	0.002	0.049	0.045	0.910
cont-mis2	−0.001	0.002	0.049	0.045	0.915
cont-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, $\beta_5 = -0.040$					
cont-cor	0.001	0.001	0.035	0.034	0.931
cont-mis1	0.001	0.001	0.035	0.034	0.930
cont-mis2	0.001	0.001	0.035	0.034	0.930
cont-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, $\beta_6 = -0.021$					
cont-cor	−0.001	0.006	0.078	0.072	0.926
cont-mis1	−0.001	0.006	0.078	0.071	0.926
cont-mis2	−0.001	0.006	0.079	0.072	0.926
cont-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 4: 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, $\beta_0 = 3.077$					
cont-cor	−0.001	0.003	0.056	0.055	0.943
cont-mis1	−0.001	0.003	0.056	0.055	0.942
cont-mis2	−0.001	0.003	0.056	0.055	0.943
cont-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, $\beta_1 = 0.002$					
cont-cor	0.000	0.000	0.001	0.001	0.944
cont-mis1	0.000	0.000	0.001	0.001	0.944
cont-mis2	0.000	0.000	0.001	0.001	0.944
cont-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, $\beta_2 = -0.012$					
cont-cor	0.000	0.000	0.017	0.017	0.950
cont-mis1	0.000	0.000	0.017	0.017	0.953
cont-mis2	0.000	0.000	0.017	0.017	0.953
cont-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, $\beta_3 = -0.032$					
cont-cor	0.001	0.001	0.027	0.027	0.953
cont-mis1	0.001	0.001	0.027	0.027	0.952
cont-mis2	0.001	0.001	0.027	0.027	0.952
cont-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, $\beta_4 = -0.032$					
cont-cor	0.000	0.000	0.014	0.013	0.944
cont-mis1	0.000	0.000	0.014	0.013	0.944
cont-mis2	0.000	0.000	0.014	0.013	0.944
cont-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, $\beta_5 = -0.040$					
cont-cor	0.000	0.000	0.014	0.014	0.943
cont-mis1	0.000	0.000	0.014	0.014	0.943
cont-mis2	0.000	0.000	0.014	0.014	0.943
cont-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, $\beta_6 = -0.021$					
cont-cor	−0.001	0.000	0.021	0.021	0.949
cont-mis1	−0.001	0.000	0.021	0.021	0.949
cont-mis2	−0.001	0.000	0.021	0.021	0.949
cont-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 5: 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, $\beta_0 = 3.077$					
cont-cor	−0.003	0.002	0.049	0.048	0.947
cont-mis1	−0.003	0.002	0.049	0.048	0.948
cont-mis2	−0.003	0.002	0.049	0.048	0.947
cont-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, $\beta_1 = 0.002$					
cont-cor	0.000	0.000	0.001	0.001	0.944
cont-mis1	0.000	0.000	0.001	0.001	0.945
cont-mis2	0.000	0.000	0.001	0.001	0.945
cont-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, $\beta_2 = -0.012$					
cont-cor	0.000	0.000	0.017	0.017	0.947
cont-mis1	0.000	0.000	0.017	0.017	0.947
cont-mis2	0.000	0.000	0.017	0.017	0.948
cont-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, $\beta_3 = -0.032$					
cont-cor	0.000	0.000	0.015	0.015	0.945
cont-mis1	0.000	0.000	0.015	0.015	0.945
cont-mis2	0.000	0.000	0.015	0.015	0.945
cont-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, $\beta_4 = -0.032$					
cont-cor	−0.003	0.002	0.045	0.043	0.924
cont-mis1	−0.003	0.002	0.044	0.041	0.921
cont-mis2	−0.003	0.002	0.045	0.043	0.926
cont-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, $\beta_5 = -0.040$					
cont-cor	0.001	0.001	0.036	0.035	0.927
cont-mis1	0.001	0.001	0.036	0.035	0.927
cont-mis2	0.001	0.001	0.036	0.035	0.927
cont-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, $\beta_6 = -0.021$					
cont-cor	0.000	0.005	0.073	0.067	0.908
cont-mis1	−0.001	0.005	0.071	0.064	0.907
cont-mis2	0.000	0.005	0.072	0.067	0.906
cont-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 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 **high**, and its effect on the selection bias function was **low**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
cont-cor	0.001	0.003	0.056	0.056	0.956
cont-mis1	0.001	0.003	0.056	0.056	0.956
cont-mis2	0.001	0.003	0.056	0.056	0.956
cont-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, $\beta_1 = 0.002$					
cont-cor	0.000	0.000	0.001	0.001	0.953
cont-mis1	0.000	0.000	0.001	0.001	0.954
cont-mis2	0.000	0.000	0.001	0.001	0.953
cont-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, $\beta_2 = -0.012$					
cont-cor	0.000	0.000	0.017	0.017	0.945
cont-mis1	0.000	0.000	0.017	0.017	0.946
cont-mis2	0.000	0.000	0.017	0.017	0.945
cont-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, $\beta_3 = -0.032$					
cont-cor	−0.001	0.001	0.027	0.028	0.948
cont-mis1	−0.001	0.001	0.027	0.028	0.949
cont-mis2	−0.001	0.001	0.027	0.028	0.947
cont-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, $\beta_4 = -0.032$					
cont-cor	0.000	0.000	0.014	0.014	0.938
cont-mis1	0.000	0.000	0.014	0.013	0.939
cont-mis2	0.000	0.000	0.014	0.014	0.939
cont-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, $\beta_5 = -0.040$					
cont-cor	0.000	0.000	0.014	0.014	0.951
cont-mis1	0.000	0.000	0.014	0.014	0.953
cont-mis2	0.000	0.000	0.014	0.014	0.954
cont-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, $\beta_6 = -0.021$					
cont-cor	0.000	0.000	0.021	0.021	0.948
cont-mis1	0.000	0.000	0.021	0.021	0.947
cont-mis2	0.000	0.000	0.021	0.021	0.946
cont-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 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 **high**, and its effect on the selection bias function was **low**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
cont-cor	0.001	0.002	0.047	0.048	0.952
cont-mis1	0.001	0.002	0.047	0.049	0.948
cont-mis2	0.001	0.002	0.047	0.048	0.952
cont-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, $\beta_1 = 0.002$					
cont-cor	0.000	0.000	0.001	0.001	0.944
cont-mis1	0.000	0.000	0.001	0.001	0.949
cont-mis2	0.000	0.000	0.001	0.001	0.944
cont-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, $\beta_2 = -0.012$					
cont-cor	0.001	0.000	0.017	0.017	0.934
cont-mis1	0.001	0.000	0.017	0.017	0.942
cont-mis2	0.001	0.000	0.017	0.017	0.934
cont-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, $\beta_3 = -0.032$					
cont-cor	0.000	0.000	0.015	0.015	0.942
cont-mis1	0.000	0.000	0.015	0.015	0.947
cont-mis2	0.000	0.000	0.015	0.015	0.942
cont-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, $\beta_4 = -0.032$					
cont-cor	−0.004	0.004	0.060	0.064	0.966
cont-mis1	−0.008	0.004	0.066	0.055	0.900
cont-mis2	−0.004	0.004	0.060	0.064	0.965
cont-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, $\beta_5 = -0.040$					
cont-cor	0.001	0.001	0.036	0.035	0.932
cont-mis1	0.001	0.001	0.036	0.035	0.935
cont-mis2	0.001	0.001	0.036	0.035	0.933
cont-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, $\beta_6 = -0.021$					
cont-cor	−0.001	0.006	0.079	0.098	0.975
cont-mis1	−0.005	0.011	0.105	0.087	0.918
cont-mis2	−0.001	0.006	0.079	0.098	0.978
cont-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 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 **low**, and its effect on the selection bias function was **high**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
cont-cor	−0.003	0.004	0.061	0.073	0.984
cont-mis1	−0.003	0.005	0.068	0.081	0.981
cont-mis2	−0.003	0.004	0.061	0.073	0.984
cont-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, $\beta_1 = 0.002$					
cont-cor	0.000	0.000	0.001	0.001	0.985
cont-mis1	0.000	0.000	0.001	0.002	0.980
cont-mis2	0.000	0.000	0.001	0.001	0.985
cont-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, $\beta_2 = -0.012$					
cont-cor	−0.001	0.001	0.023	0.024	0.964
cont-mis1	−0.001	0.001	0.025	0.026	0.959
cont-mis2	−0.001	0.001	0.023	0.025	0.968
cont-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, $\beta_3 = -0.032$					
cont-cor	0.001	0.001	0.027	0.029	0.967
cont-mis1	0.000	0.001	0.031	0.046	0.996
cont-mis2	0.001	0.001	0.027	0.029	0.967
cont-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, $\beta_4 = -0.032$					
cont-cor	0.000	0.000	0.017	0.019	0.968
cont-mis1	−0.001	0.000	0.019	0.021	0.967
cont-mis2	0.000	0.000	0.017	0.019	0.968
cont-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, $\beta_5 = -0.040$					
cont-cor	0.002	0.000	0.019	0.019	0.940
cont-mis1	0.002	0.000	0.021	0.020	0.945
cont-mis2	0.002	0.000	0.019	0.019	0.940
cont-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, $\beta_6 = -0.021$					
cont-cor	0.000	0.001	0.024	0.029	0.984
cont-mis1	0.000	0.001	0.027	0.031	0.983
cont-mis2	0.000	0.001	0.024	0.029	0.984
cont-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 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 **low**, and its effect on the selection bias function was **high**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
cont-cor	−0.003	0.002	0.050	0.048	0.947
cont-mis1	−0.003	0.003	0.051	0.050	0.941
cont-mis2	−0.003	0.002	0.050	0.048	0.947
cont-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, $\beta_1 = 0.002$					
cont-cor	0.000	0.000	0.001	0.001	0.945
cont-mis1	0.000	0.000	0.001	0.001	0.944
cont-mis2	0.000	0.000	0.001	0.001	0.945
cont-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, $\beta_2 = -0.012$					
cont-cor	0.000	0.000	0.017	0.017	0.946
cont-mis1	0.000	0.000	0.018	0.018	0.948
cont-mis2	0.000	0.000	0.017	0.017	0.948
cont-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, $\beta_3 = -0.032$					
cont-cor	0.000	0.000	0.015	0.015	0.945
cont-mis1	0.000	0.000	0.015	0.015	0.946
cont-mis2	0.000	0.000	0.015	0.015	0.945
cont-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, $\beta_4 = -0.032$					
cont-cor	−0.012	0.007	0.081	0.093	0.975
cont-mis1	−0.019	0.009	0.095	0.067	0.839
cont-mis2	−0.012	0.007	0.081	0.093	0.975
cont-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, $\beta_5 = -0.040$					
cont-cor	0.001	0.001	0.036	0.035	0.931
cont-mis1	0.001	0.001	0.038	0.036	0.935
cont-mis2	0.001	0.001	0.036	0.035	0.928
cont-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, $\beta_6 = -0.021$					
cont-cor	0.000	0.006	0.077	0.141	0.999
cont-mis1	−0.007	0.018	0.133	0.106	0.885
cont-mis2	0.000	0.006	0.077	0.141	0.999
cont-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 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 **high**, and its effect on the selection bias function was **high**.

Estimator	Bias	MSE	emp sd	est sd	coverage
Intercept, $\beta_0 = 3.077$					
cont-cor	0.000	0.004	0.065	0.084	0.989
cont-mis1	0.001	0.005	0.068	0.088	0.988
cont-mis2	0.000	0.004	0.065	0.084	0.989
cont-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, $\beta_1 = 0.002$					
cont-cor	0.000	0.000	0.001	0.002	0.996
cont-mis1	0.000	0.000	0.001	0.002	0.994
cont-mis2	0.000	0.000	0.001	0.002	0.996
cont-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, $\beta_2 = -0.012$					
cont-cor	0.000	0.001	0.029	0.028	0.937
cont-mis1	0.000	0.001	0.031	0.029	0.930
cont-mis2	0.000	0.001	0.029	0.029	0.938
cont-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, $\beta_3 = -0.032$					
cont-cor	−0.001	0.001	0.028	0.033	0.974
cont-mis1	−0.001	0.001	0.029	0.038	0.987
cont-mis2	−0.001	0.001	0.028	0.033	0.974
cont-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, $\beta_4 = -0.032$					
cont-cor	−0.002	0.001	0.023	0.027	0.978
cont-mis1	−0.002	0.001	0.023	0.021	0.937
cont-mis2	−0.002	0.001	0.023	0.027	0.978
cont-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, $\beta_5 = -0.040$					
cont-cor	0.001	0.000	0.022	0.022	0.970
cont-mis1	0.001	0.000	0.022	0.023	0.968
cont-mis2	0.001	0.000	0.022	0.022	0.970
cont-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, $\beta_6 = -0.021$					
cont-cor	0.001	0.001	0.031	0.040	0.991
cont-mis1	0.001	0.001	0.031	0.032	0.953
cont-mis2	0.001	0.001	0.031	0.040	0.991
cont-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

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 **high**, and its effect on the selection bias function was **high**.