Supporting Information for "Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification" by

Lauren J. Beesley*[1] and Bhramar Mukherjee[1]
[1]University of Michigan, Department of Biostatistics
*Corresponding Author: lbeesley@umich.edu

Contents

A Analytical Results

In this section, we provide proofs and derivations for the results in the main paper.

A.1 Proof of *Eq. 3* and *Eq. 4*

Our goal is to relate the overall analysis model, $P(D^* = 1|Z, S = 1)$, to parameters in the conceptual model in *Eq. 1*. We have the following

$$P(D^* = 1|Z, S = 1) = \frac{P(D^* = 1, S = 1|Z)}{P(S = 1|Z)} = \frac{\sum_d P(D^* = 1, S = 1, D = d|Z)}{\sum_d P(S = 1, D = d|Z)}$$
$$= \frac{\sum_d P(D^* = 1|S = 1, D = d, Z)P(S = 1|D = d, Z)P(D = d|Z)}{\sum_d P(S = 1|D = d, Z)P(D = d|Z)}$$

Now, under our model assumptions and notation in *Eq. 2*, $P(D^* = 1|S = 1, D = 1, Z) = c(Z)$ and $P(D^* = 1|S = 1, D = 0, Z) = 1 - b(Z)$ where $b(Z) = \int P(D^* = 0|S = 1, D = 0, Y)f(Y^\dagger|Z, D = 0, S = 1)dY^\dagger$. We have

$$P(D^* = 1|Z, S = 1) = \frac{c(Z)P(S = 1|D = 1, Z)P(D = 1|Z) + \{1 - b(Z)\}P(S = 1|D = 0, Z)P(D = 0|Z)}{P(S = 1|D = 1, Z)P(D = 1|Z) + P(S = 1|D = 0, Z)P(D = 0|Z)}$$

We also note that $r(Z) = \frac{P(S=1|D=1,Z)}{P(S=1|D=0,Z)}$, so we can simplify the above expression to

$$P(D^* = 1|Z, S = 1) = \frac{c(Z)r(Z)P(D = 1|Z) + \{1 - b(Z)\}P(D = 0|Z)}{r(Z)P(D = 1|Z) + P(D = 0|Z)}$$
$$= \frac{1 - b(Z) + [c(Z)r(Z) - \{1 - b(Z)\}]\, P(D = 1|Z)}{1 + [r(Z) - 1]\, P(D = 1|Z)}$$

or equivalently,

$$P(D = 1|Z) = \frac{P(D^* = 1|Z, S = 1) - \{1 - b(Z)\}}{c(Z)r(Z) - \{1 - b(Z)\} - P(D^* = 1|Z, S = 1)\{r(Z) - 1\}}$$

Therefore, we can directly express the analysis model in terms of different contributions to the conceptual model. This gives us the expression in *Eq. 3*. $c(Z)$ reflects contributions of misclassification in terms of sensitivity, $b(Z)$ represents contributions of misclassification in terms of specificity, and $r(Z)$ reflects contributions of the sampling mechanism. Notably, if we set $r(Z) = 1$, we have

$$P(D^* = 1|Z, S = 1) = 1 - b(Z) + [c(Z) - \{1 - b(Z)\}]\, P(D = 1|Z)$$

and $P(D^* = 1|Z, S = 1) = P(D = 1|Z)$ if $c(Z)$ and $b(Z)$ are also equal to 1.

Now, suppose we model $D|Z$ using a logistic regression as in *Eq. 1*. In this case, we have that

$$\text{logit}\left[\frac{P(D^* = 1|Z, S = 1) - \{1 - b(Z)\}}{c(Z)r(Z) - \{1 - b(Z)\} - P(D^* = 1|Z, S = 1)\,[r(Z) - 1]}\right] = \text{logit}\,[P(D = 1|Z)] = \theta_0 + \theta_Z Z$$
$$\implies \log\left[\frac{P(D^* = 1|Z, S = 1) - \{1 - b(Z)\}}{c(Z)r(Z) - r(Z)P(D^* = 1|Z, S = 1)}\right] = \theta_0 + \theta_Z Z$$

So we have that

$$\log\left[\frac{P(D^* = 1|Z, S = 1) - \{1 - b(Z)\}}{c(Z) - P(D^* = 1|Z, S = 1)}\right] = \theta_0 + \theta_Z Z + \log\,[r(Z)]$$

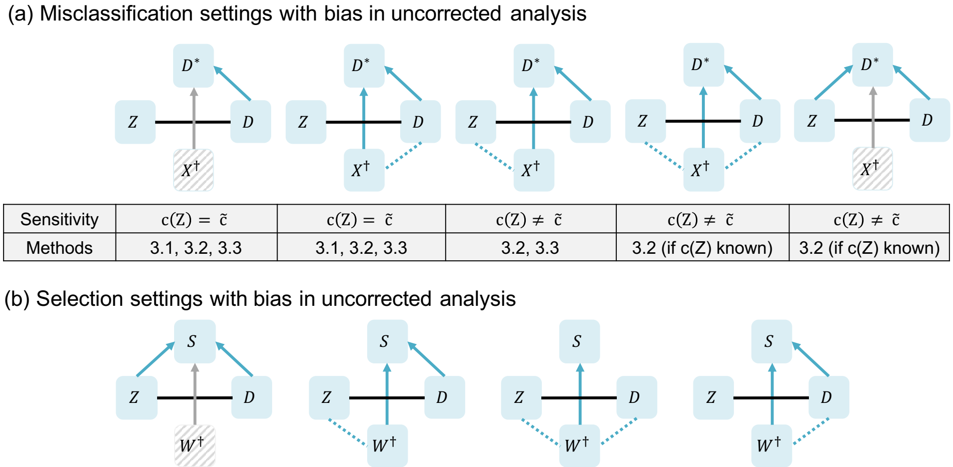
This produces the expression in *Eq. 4*.

A.2 Bias under naive (uncorrected) analysis

The relationship in *Eq. 4* provides insight into settings in which we do and do not expect bias in estimating θ by fitting standard logistic regression model for $D^*|S=1,Z$.

Suppose first that $c(Z)=b(Z)=1$, so we have no misclassification of observed D. In this case, we have the following: $\text{logit}\left[P(D=1|Z,S=1)\right]=\theta_0+\theta_Z Z+\log\left[r(Z)\right]$. Suppose further that we attempt to estimate θ by fitting a logistic regression model for $D|Z$ on the sampled patients using only main effects contributions of Z and *ignoring* the potential contribution of $r(Z)$. We expect bias in estimating θ_Z in this setting if $r(Z)$ depends on Z. This may happen if selection depends directly on Z or if sampling depends on $W^\dagger$ that is associated with Z given D as shown in **Figure A.1b**. If selection depends on W that is *independent* of Z given D, there is still some possibility of small bias in estimating θ_Z if W is independently related to D (Neuhaus and Jewell, 1993).

Figure A.1: Settings resulting in bias in estimating θ_Z from logistic regression under imperfect sensitivity (a) or selection bias (b). Bias will also occur when we have imperfect specificity. *


(a) Misclassification settings with bias in uncorrected analysis

Sensitivity	c(Z) = c̃	c(Z) = c̃	c(Z) ≠ c̃	c(Z) ≠ c̃	c(Z) ≠ c̃
Methods	3.1, 3.2, 3.3	3.1, 3.2, 3.3	3.2, 3.3	3.2 (if c(Z) known)	3.2 (if c(Z) known)

(b) Selection settings with bias in uncorrected analysis

* Solid lines indicate associations, and arrows indicate drivers of patient selection or misclassification. Diagonally-shaded boxes correspond to sets of predictors that may or may not be empty (equal to $\emptyset$). Below each setting in (a), we show implications for $c(Z)$ and list the section or sections of the main paper that can be applied (**Sections 3.1, 3.2, and/or 3.3**). We note that the method in **Section 3.3** can only be applied in cases where $b(Z)=\widetilde{b}=1$. The methods in **Sections 3.1 and 3.2** can only be applied when $b(Z)=\widetilde{b}$. The final setting in (b) will generally only result in small or negligible bias. See Neuhaus and Jewell (1993) for details.

Suppose instead that selection is ignorable ($r(Z)=1$) and that we model potentially misclassified $D^*|Z$ using a standard logistic regression model. In this case, the true relationship is $\log\left[\frac{P(D^*=1|Z,S=1)-\{1-b(Z)\}}{c(Z)-P(D^*=1|Z,S=1)}\right]=\theta_0+\theta_Z Z$. Fitting a standard logistic regression will result in some bias in estimating θ for *any* $c(Z)\neq 1$ or $b(Z)\neq 1$. **Figure A.1a** provides a roadmap for which methods in the main paper can be applied to correct this bias based on the underlying relationships between X, Z, D, and D^*. If both (1) $c(Z)\neq 1$ or $b(Z)\neq 1$ and (2) $r(Z)\neq 1$, there is even greater potential for bias.

A.3 Proof of *Eq. 5* and its extension to non-ignorable sampling

A.3.1 Ignorable sampling or constant sampling ratio

In Beesley et al. (2020), we used Taylor series approximations to express the uncorrected parameter associated with Z from the model for $D^*|Z, S = 1$, denoted θ_Z^{uc}, in terms of the true θ, unknown sensitivity and specificity $\widetilde{c}$ and $\widetilde{b}$, and sampling ratio, $\widetilde{r}$. In that paper, we made additional restricting assumptions on X, Y, and W that, ultimately, boil down to the following: (1) $r(Z) = \widetilde{r}$, (2) $c(Z) = \widetilde{c}$, and (3) $b(Z) = \widetilde{b}$. In this particular setting, we showed that we can approximate θ_Z^{uc} as

$$\theta_Z^{uc} \approx \left[\frac{e^{\theta_0+\theta_Z\bar{Z}}\widetilde{c}\widetilde{r}}{e^{\theta_0+\theta_Z\bar{Z}}\widetilde{c}\widetilde{r}+1-\widetilde{b}} - \frac{e^{\theta_0+\theta_Z\bar{Z}}\{1-\widetilde{c}\}\widetilde{r}}{e^{\theta_0+\theta_Z\bar{Z}}\{1-\widetilde{c}\}\widetilde{r}+\widetilde{b}}\right]\theta_Z$$

where $\bar{Z}$ is the mean of Z. Now, suppose that we replace $\frac{e^{\theta_0+\theta_Z\bar{Z}}}{1+e^{\theta_0+\theta_Z\bar{Z}}} = P(D = 1|\bar{Z})$ with population prevalence $P(D = 1)$. We also note that $p^* = P(D^* = 1|S = 1) = \sum_{d=0,1} P(D^* = 1|D = d|S = 1)P(D = d|S = 1) = \widetilde{c}P(D = 1|S = 1) + [1 - \widetilde{b}]P(D = 0|S = 1)$. We rewrite the above equation as

$$\begin{aligned} p^* &= \widetilde{c}\frac{\widetilde{r}P(D=1)}{\widetilde{r}P(D=1)+P(D=0)} + [1-\widetilde{b}]\frac{P(D=0)}{\widetilde{r}P(D=1)+P(D=0)} \\ \implies P(D=1) &= \frac{p^*-[1-\widetilde{b}]}{p^*+[\widetilde{c}-p^*]\widetilde{r}-[1-\widetilde{b}]} \end{aligned}$$

putting these together, we have

$$\begin{aligned} \theta_Z^{uc} &\approx \left[\frac{\frac{e^{\theta_0+\theta_Z\bar{Z}}}{e^{\theta_0+\theta_Z\bar{Z}}+1}\widetilde{c}\widetilde{r}}{\frac{e^{\theta_0+\theta_Z\bar{Z}}}{e^{\theta_0+\theta_Z\bar{Z}}+1}\widetilde{c}\widetilde{r}+\frac{1-\widetilde{b}}{e^{\theta_0+\theta_Z\bar{Z}}+1}} - \frac{\frac{e^{\theta_0+\theta_Z\bar{Z}}}{e^{\theta_0+\theta_Z\bar{Z}}+1}\{1-\widetilde{c}\}\widetilde{r}}{\frac{e^{\theta_0+\theta_Z\bar{Z}}}{e^{\theta_0+\theta_Z\bar{Z}}+1}\{1-\widetilde{c}\}\widetilde{r}+\frac{\widetilde{b}}{e^{\theta_0+\theta_Z\bar{Z}}+1}}\right]\theta_Z \\ &\approx \left[\frac{P(D=1)\widetilde{c}\widetilde{r}}{P(D=1)\widetilde{c}\widetilde{r}+P(D=0)\{1-\widetilde{b}\}} - \frac{P(D=1)\{1-\widetilde{c}\}\widetilde{r}}{P(D=1)\{1-\widetilde{c}\}\widetilde{r}+P(D=0)\widetilde{b}}\right]\theta_Z \\ &= \left[p^*-\{1-\widetilde{b}\}\right]\left[\frac{\widetilde{c}\widetilde{r}}{\left\{p^*-(1-\widetilde{b})\right\}\widetilde{c}\widetilde{r}+\{\widetilde{c}-p^*\}\widetilde{r}\{1-\widetilde{b}\}} - \frac{\{1-\widetilde{c}\}\widetilde{r}}{\left\{p^*-(1-\widetilde{b})\right\}\{1-\widetilde{c}\}\widetilde{r}+\{\widetilde{c}-p^*\}\widetilde{r}\widetilde{b}}\right]\theta_Z \\ &= \left[p^*-(1-\widetilde{b})\right]\left[\frac{\widetilde{c}}{p^*\left\{\widetilde{c}-(1-\widetilde{b})\right\}} - \frac{\{1-\widetilde{c}\}}{[1-p^*]\left\{\widetilde{c}-(1-\widetilde{b})\right\}}\right]\theta_Z \\ \implies \theta_Z &\approx \theta_Z^{uc}\frac{\left[\widetilde{c}-\{1-\widetilde{b}\}\right]p^*[1-p^*]}{\left[p^*-\{1-\widetilde{b}\}\right][\widetilde{c}-p^*]} \end{aligned}$$

This gives us *Eq. 5*. Now, if we also have that $\widetilde{b} = 1$, this expression further reduces to

$$\theta_Z \approx \theta_Z^{uc}\frac{\widetilde{c}[1-p^*]}{\widetilde{c}-p^*}$$

This is the exact same structure as the estimator in Duffy et al. (2004). Our derivations show that this estimator is justified for Z that is non-binary and for $\widetilde{r} \neq 1$ as well. One notable feature of the above estimator is that it does not depend on $\widetilde{r}$. Under the restrictive assumptions on $r(Z)$, $c(Z)$, and $b(Z)$ above, we can adjust for both misclassification and selection using the above estimator. Intuitively, this is because p^* will be impacted by both the misclassification and sampling mechanisms.

Treating $\widetilde{c}$ and $\widetilde{b}$ as fixed and replacing θ_Z^{uc} with an estimate, we can express

$$Var(\hat{\theta}_Z) = Var(\hat{\theta}_Z^{uc}) \left[\frac{\left\{ \widetilde{c} - (1-\widetilde{b}) \right\} p^* \left\{ 1 - p^* \right\}}{\left\{ p^* - (1-\widetilde{b}) \right\} \left\{ \widetilde{c} - p^* \right\}} \right]^2$$

or the following under $\widetilde{b} = 1$:

$$Var(\hat{\theta}_Z) = Var(\hat{\theta}_Z^{uc}) \left[\frac{\widetilde{c} \left\{ 1 - p^* \right\}}{\widetilde{c} - p^*} \right]^2$$

In reality, $\widetilde{c}$ unknown. However, we can obtain an estimate of $\widetilde{c}$ and incorporate our uncertainty about this value. We will still treat p^* as fixed due to the large sample we will be applying these methods to. Additionally, we will consider the case where $\widetilde{b} = 1$ for simplicity. We have

$$\begin{aligned}
\text{Var}(\hat{\theta}_Z) &= Var\left[E(\hat{\theta}_Z|c) \right] + E\left[Var(\hat{\theta}_Z|c) \right] \\
&\approx Var\left[E\left(\hat{\theta}_Z^{uc} \frac{\widetilde{c}\{1-p^*\}}{\widetilde{c}-p^*} |c \right) \right] + E\left[Var\left(\hat{\theta}_Z^{uc} \frac{\widetilde{c}\{1-p^*\}}{\widetilde{c}-p^*} |c \right) \right] \\
&= Var\left[\frac{\widetilde{c}\{1-p^*\}}{\widetilde{c}-p^*} E\left(\hat{\theta}_Z^{uc}|c \right) \right] + E\left[\left\{ \frac{\widetilde{c}(1-p^*)}{\widetilde{c}-p^*} \right\}^2 Var\left(\hat{\theta}_Z^{uc}|c \right) \right] \\
&\approx E\left(\hat{\theta}_Z^{uc} \right)^2 [1-p^*]^2 Var\left(\frac{\widetilde{c}}{\widetilde{c}-p^*} \right) + Var\left(\hat{\theta}_Z^{uc} \right) [1-p^*]^2 E\left(\left[\frac{\widetilde{c}}{\widetilde{c}-p^*} \right]^2 \right)
\end{aligned}$$

Using Taylor series and other approximations, we have

$$\text{Var}(\hat{\theta}_Z) \approx \hat{\theta}_Z^{uc\,2} [1-p^*]^2 \frac{[p^*]^2}{[E(\widetilde{c}) - p^*]^2} Var(\widetilde{c}) + \hat{Var}\left(\hat{\theta}_Z^{uc} \right) [1-p^*]^2 \left[\frac{E(\widetilde{c})}{E(\widetilde{c}) - p^*} \right]^2 \qquad \textit{(Eq. S1)}$$

This is now a function of known values along with $E(\widetilde{c})$ and $Var(\widetilde{c})$. We can insert our prior uncertainty about $\widetilde{c}$ or its estimate into this expression to get the resulting variance.

A.3.2 Sampling ratio related to Z

We now consider the setting where the sampling ratio $r(Z)$ is *not* assumed to be equal to a constant. This is a more plausible setting for EHR data. We first take another look at the estimator from *Eq. 5*. Under ignorable selection ($r(Z) = \widetilde{r}$), we get expression

$$\theta_Z \approx \theta_Z^{uc} \frac{\left[\widetilde{c} - \{1 - \widetilde{b}\} \right] p^* \left[1 - p^* \right]}{\left[p^* - \{1 - \widetilde{b}\} \right] \left[\widetilde{c} - p^* \right]}$$

where $p^* = P(D^* = 1|S = 1) = P(D^* = 1)$ and θ_Z^{uc} is from $f(D^*|Z, S = 1) = f(D^*|Z)$.

In order to apply this estimator in the more general setting, we estimate $p^* = P(D^* = 1)$ and θ_Z^{uc} from $f(D^*|Z)$ directly. Given the observed data on the sampled patients and IPW or calibration weights ω, we can estimate

$$p^* = P(D^* = 1) = \frac{\sum_{i \text{ in sample}} \omega_i D_i^*}{\sum_{i \text{ in sample}} \omega_i}$$

We can estimate θ_Z^{uc} by fitting a model for $D^*|Z$ on the sampled data *weighted* by ω. The resulting estimator takes a similar form to the setting with ignorable missingness, but the estimation of θ_Z^{uc} and p^* incorporates sampling weights.

A.4 Replacing $c(Z)$ with $c_{true}(X)$

In **Section 3.2** of the main paper, we discuss replacing $c(Z)$ with $c_{true}(X)$ for estimation of θ. In this section, we make the assumption that specificity is a constant in Z, i.e. $b(Z) = \widetilde{b}$. We provide two conditions under which the replacement of $c(Z)$ with $c_{true}(X)$ is appropriate. Here, we provide some support for these assertions. First, we note that

$$c_{true}(X)P(D = 1|Z, X^{\dagger}) + (1 - \widetilde{b})P(D = 0|Z, X^{\dagger}) = P(D^{*} = 1|Z, X^{\dagger})$$

Under a logistic regression model, this relationship implies

$$\log\left[\frac{P(D^{*} = 1|Z, X^{\dagger}) - (1 - \widetilde{b})}{c_{true}(X) - P(D^{*} = 1|Z, X^{\dagger})}\right] = \text{logit}\left[P(D = 1|Z, X^{\dagger})\right]$$

Suppose first that $D \perp X^{\dagger}|Z$. In this case, the above expression reduces to

$$\log\left[\frac{P(D^{*} = 1|Z, X^{\dagger}) - (1 - \widetilde{b})}{c_{true}(X) - P(D^{*} = 1|Z, X^{\dagger})}\right] = \theta_0 + \theta_Z Z$$

which is the expression we want to apply to estimate θ_Z after replacing $c(Z)$ with $c_{true}(X)$.

In practice, it may not be reasonable to assume that D is independent of factors in $X^{\dagger}$ such as length of follow-up or number of doctor's visits. Therefore, we want to explore alternative assumptions that will allow for this substitution. First, we note that

$$\begin{aligned} P(D = 1|Z, X^{\dagger}) &= \frac{f(X^{\dagger}|D = 1, Z)P(D = 1|Z)}{f(X^{\dagger}|Z)} \\ &= \frac{f(X^{\dagger}|D = 1, Z)P(D = 1|Z)}{f(X^{\dagger}|D = 1, Z)P(D = 1|Z) + f(X^{\dagger}|D = 0, Z)P(D = 0|Z)} \end{aligned}$$

Replacing this expression into the logistic regression above, we have that

$$\log\left[\frac{P(D^{*} = 1|Z, X^{\dagger}) - (1 - \widetilde{b})}{c_{true}(X) - P(D^{*} = 1|Z, X^{\dagger})}\right] = \theta_0 + \theta_Z Z - \log\left[\frac{f(X^{\dagger}|D = 0, Z)}{f(X^{\dagger}|D = 1, Z)}\right]$$

Again, this last term is zero if $f(X^{\dagger}|D = 0, Z) = f(X^{\dagger}|D = 1, Z)$, so if $D \perp X^{\dagger}|Z$. Alternatively, suppose that $Z \perp X^{\dagger}|D$. In this case, the above expression reduces to

$$\log\left[\frac{P(D^{*} = 1|Z, X^{\dagger}) - (1 - \widetilde{b})}{c_{true}(X) - P(D^{*} = 1|Z, X^{\dagger})}\right] = \theta_0 + \theta_Z Z - \log\left[\frac{f(X^{\dagger}|D = 0)}{f(X^{\dagger}|D = 1)}\right]$$

The final term will be a function of $X^{\dagger}$ or possibly a constant. In either case, we do not expect failure to include this offset term will result in much bias in estimating θ_Z (Neuhaus and Jewell, 1993). However, θ_0 may be impacted by a failure to include this term. Usually, however, we are primarily interested in estimating θ_Z, and inference about θ_Z obtained by replacing $c(Z)$ with $c_{true}(X)$ and ignoring the offset term will have little residual bias.

A.5 Proof of *Eq. 6* and *Eq. 9*

In this section, we explore how to estimate $c_{true}(X)$. We can apply this strategy for estimating $c_{true}(X)$ whether or not we have perfect specificity under assumptions that specificity $b(Z) = \widetilde{b} = P(D^* = 0|Z, S = 1, D = 0)$ is a known constant. We will also assume selection is ignorable $(\widetilde{r} = 1)$. We observe that

$$c_{true}(X) = \frac{P(D^* = 1|X) - \{1 - \widetilde{b}\} + \{1 - \widetilde{b}\}P(D = 1|X)}{P(D = 1|X)} = \text{expit}\left(\beta_0 + \beta_X X\right)$$

If we assume a logistic regression model structure for sensitivity as in *Eq. 1*, we have

$$\text{logit}\left[\frac{P(D^* = 1|X) - \{1 - \widetilde{b}\} + \{1 - \widetilde{b}\}P(D = 1|X)}{P(D = 1|X)}\right] = \beta_0 + \beta_X X$$

So we have that

$$\log\left[\frac{P(D^* = 1|X) - \{1 - \widetilde{b}\}P(D = 0|X)}{P(D = 1|X) + \{1 - \widetilde{b}\}P(D = 0|X) - P(D^* = 1|X)}\right] = \beta_0 + \beta_X X \qquad (\textit{Eq. S2})$$

In the special case where $\widetilde{b} = 1$, we have that

$$\log\left[\frac{P(D^* = 1|X)}{P(D = 1|X) - P(D^* = 1|X)}\right] = \beta_0 + \beta_X X$$

These expressions allow us to estimate β if $P(D = 1|X)$ is known, but in reality we will not know this term. For example, X may contain information such as the length of follow-up in the EHR, and we will likely not know how this is related to true disease status. However, we can incorporate some prior beliefs about $P(D = 1|X)$ to estimate β using the above expression.

Specifying $P(D = 1|X)$ in practice

Suppose first that D is independent of X, so $P(D = 1|X) = P(D = 1)$. In this case, we can replace $P(D = 1|X)$ with $P(D = 1)$, the population disease prevalence. For EHR data, it may be that known risk factors such as age and gender are indicators for enhanced disease screening and, therefore, may be incorporated into X and related to D. In this case, we may know the relationship $P(D = 1|X_{sub})$ for some subset X_{sub} of X from population summary statistics. If we assume D is *independent* of the elements of X *not included* in X_{sub}, then we have $P(D = 1|X) = P(D = 1|X_{sub})$, which can be replaced with population summary statistics. This will allow us to estimate β.

Suppose, instead, that there are elements of X that are related to D and that the relationship between those elements and D is unknown. In that case, $P(D = 1|X)$ is unknown. In this case, we propose *approximating* $P(D = 1|X)$ with what information is available, i.e. $P(D = 1)$ or $P(D = 1|X_{sub})$. The extent to which estimates of β and downstream estimates of θ are impacted is considered in simulations in **Section B.3**.

Importantly, *Eq. S2* may not always have a solution for a given *estimate* $P(D = 1|X_{sub})$, and it could produce inaccurate sensitivity estimates when $P(D = 1|X_{sub})$ is poorly specified (see simulations for details). An alternative strategy for estimating $c_{true}(X)$ is to fit a standard regression model for $P(D^* = 1|X)$ and estimate $c_{true}(X)$ using the ratio

$$\min\left(\frac{P(D^* = 1|X) - \{1 - \widetilde{b}\} + \{1 - \widetilde{b}\}P(D = 1|X)}{P(D = 1|X)}, 1\right)$$

using estimates for both the numerator and denominator. This "ratio" estimator will provide estimates of $c_{true}(X)$ in settings where there is no solution to *Eq. S2*.

Now, we consider the setting where we have potential selection bias. We first observe that

$$P(D=1|S=1,X) = \frac{P(S=1|X,D=1)P(D=1|X)}{\sum_d P(S=1|X,D=d)P(D=d|X)}$$

Suppose we approximate $\frac{P(S=1|X,D=1)}{P(S=1|X,D=0)}$ with $\widetilde{r} = \frac{P(S=1|D=1)}{P(S=1|D=0)}$. This then gives that $P(D=1|S=1,X) \approx \frac{\widetilde{r}P(D=1|X)}{\widetilde{r}P(D=1|X)+P(D=0|X)}$. We may expect broadly different factors to be driving selection and misclassification given D, so it may be reasonable to assume X is independent of W given D, which gives $P(S=1|X,D) = P(S=1|D)$. This may not always be the case if, for example, age is a driver of both selection and misclassification given D. Future work can explore the sensitivity of estimated $c_{true}(X)$ and resulting θ_Z estimates to the implicit assumption that $P(S=1|X,D) = P(S=1|D)$.

Suppose that we can make this approximation. Using logic as before, we have that

$$\log\left[\frac{P(D^*=1|X,S=1) - \{1-\widetilde{b}\}P(D=0|X,S=1)}{P(D=1|X,S=1) + \{1-\widetilde{b}\}P(D=0|X,S=1) - P(D^*=1|X,S=1)}\right] = \beta_0 + \beta_X X$$

Substituting the approximation for $P(D=1|S=1,X)$, we have

$$\log\left[\frac{P(D^*=1|X,S=1) - \{1-\widetilde{b}\}\frac{P(D=0|X)}{\widetilde{r}P(D=1|X)+P(D=0|X)}}{\frac{\widetilde{r}P(D=1|X)+\{1-\widetilde{b}\}P(D=0|X)}{\widetilde{r}P(D=1|X)+P(D=0|X)} - P(D^*=1|X,S=1)}\right] \approx \beta_0 + \beta_X X$$

Assuming $\widetilde{b} = 1$, this gives

$$\log\left[\frac{P(D^*=1|X,S=1)}{\frac{\widetilde{r}P(D=1|X)}{\widetilde{r}P(D=1|X)+P(D=0|X)} - P(D^*=1|X,S=1)}\right] \approx \beta_0 + \beta_X X$$

As before, this expression may not always have a solution in β for a given specification of $\widetilde{r}$ or $P(D=1|X)$ incompatible with the data. In this case, we could also estimate $c_{true}(X)$ using the ratio

$$\min\left(\frac{P(D^*=1|X,S=1) - \{1-\widetilde{b}\} + \{1-\widetilde{b}\}\frac{\widetilde{r}P(D=1|X)}{\widetilde{r}P(D=1|X)+P(D=0|X)}}{\frac{\widetilde{r}P(D=1|X)}{\widetilde{r}P(D=1|X)+P(D=0|X)}}, 1\right) \qquad (\textit{Eq. S3})$$

A.6 Jointly estimating θ and β

In this section, we describe how we can jointly estimate θ and β to deal with misclassification. In this section, we are assuming that $b(Z) = \widetilde{b} = 1$, so we have perfect specificity.

A.6.1 Some assumptions

First, we notice that $P(D^* = 1|Z, X^\dagger) = c_{true}(X)P(D = 1|Z, X^\dagger)$. As shown in **Supporting Section A.4**, we have that
(a) $P(D^* = 1|Z, X^\dagger) = \text{expit}(\beta_0 + \beta_X X)\text{expit}(\theta_0 + \theta_Z Z)$ if $D \perp X^\dagger|Z$ or that
(b) $P(D^* = 1|Z, X^\dagger) = \text{expit}(\beta_0 + \beta_X X)\text{expit}\left[\theta_0 + \theta_Z Z - \log\left(\frac{f(X^\dagger|D=0)}{f(X^\dagger|D=1)}\right)\right]$ if $Z \perp X^\dagger|D$.
Fixing β, we would expect little bias in estimating θ_Z in the latter case if we were to drop the offset term involving $X^\dagger$ from the equation (Neuhaus and Jewell, 1993). Therefore, we will define the observed data log-likelihood using model structure

$$P(D^* = 1|Z, X^\dagger) = \text{expit}(\beta_0 + \beta_X X)\text{expit}(\theta_0 + \theta_Z Z)$$

with an understanding that either (a) $D \perp X^\dagger|Z$ or (b) $Z \perp X^\dagger|D$ must hold and resulting inference about θ_0 may be subject to residual bias under (b) and not (a). Diop et al. (2011) shows that this model is identifiable if we have a continuous covariate that is included in X but not Z or vice-versa. For EHR data, we expect factors such as length of follow-up in the EHR to be included in X but not Z, so we will often have identifiability, and we can improve identifiability by fixing β_0 as discussed later on.

A.6.2 Direct maximization of observed data log-likelihood

Under these assumptions, we define the *observed* data log-likelihood as follows:

$$l_{obs}(\theta, \beta) = \sum_i D_i^* \log\left[\frac{e^{\beta_0+\beta_X X_i}}{1 + e^{\beta_0+\beta_X X_i}} \frac{e^{\theta_0+\theta_Z Z_i}}{1 + e^{\theta_0+\theta_Z Z_i}}\right] + (1 - D_i^*)\log\left[1 - \frac{e^{\beta_0+\beta_X X_i}}{1 + e^{\beta_0+\beta_X X_i}} \frac{e^{\theta_0+\theta_Z Z_i}}{1 + e^{\theta_0+\theta_Z Z_i}}\right]$$
$$= \sum_i D_i^* \log\left[K_i(\theta, \beta)\right] + (1 - D_i^*)\log\left[1 - K_i(\theta, \beta)\right]$$

We can estimate θ and β by directly maximizing this likelihood through a Newton-Raphson algorithm or numerical optimization method. We have the following score and expected information matrices.

$$U^u_{obs}(\theta, \beta) = \sum_i \frac{D_i^* - K_i(\theta, \beta)}{K_i(\theta, \beta)[1 - K_i(\theta, \beta)]} \frac{\partial K_i(\theta, \beta)}{\partial u}$$
$$I^{uv}_{obs}(\theta, \beta) = \sum_i \frac{1}{K_i(\theta, \beta)[1 - K_i(\theta, \beta)]} \frac{\partial K_i(\theta, \beta)}{\partial u} \frac{\partial K_i(\theta, \beta)}{\partial v^T}$$
$$\frac{\partial K_i(\theta, \beta)}{\partial u} = \begin{bmatrix} u = \theta : & \frac{K_i(\theta,\beta)}{1+e^{\theta_0+\theta_Z Z_i}}(1, Z_i^T) \\ u = \beta : & \frac{K_i(\theta,\beta)}{1+e^{\beta_0+\beta_X X_i}}(1, X_i^T) \end{bmatrix}$$

These expressions can be easily calculated given the observed data.

The task of jointly maximizing θ and β, however, can be numerically challenging. In particular, the likelihood surface can be difficult to maximize when both intercepts θ_0 and β_0 are left unspecified. Therefore, we perform parameter estimation using a profile likelihood strategy across β_0, where we specify discrete values of β_0, perform maximization to estimate other parameters given that value of β_0, and ultimately choose the value of β_0 that results in the largest log-likelihood values. In simulation, we have found that this strategy tends to have improved performance over joint maximization of all model parameters. Additionally, one can specify a single fixed value for β_0 a priori. One strategy is to set β_0 to the logit of an estimate of $\widetilde{c}$ as in

Section 3.1 for mean-centered X. This may be a useful strategy for improving our ability to estimate other model parameters and tends to perform well in simulation.

A.6.3 Maximization using an EM algorithm

Direct numerical maximization of the observed data log-likelihood can sometimes be cumbersome for large datasets. In this setting, it can be faster to perform parameter estimation using the following expectation-maximization (EM) algorithm. Firstly, we can write the *complete* data log-likelihood as follows:

$$l_{com}(\theta,\beta) = \sum_i D_i \text{log}\left[\frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}\right] + (1-D_i)\text{log}\left[\frac{1}{1+e^{\theta_0+\theta_Z Z_i}}\right]$$
$$+ D_i^* D_i \text{log}\left[\frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}}\right] + (1-D_i^*)D_i\text{log}\left[\frac{1}{1+e^{\beta_0+\beta_X X_i}}\right]$$

This expression is linear in D_i. Given the observed data and our modeling assumptions, we can replace D_i in the E-step of the EM-algorithm with

$$\begin{aligned} p = P(D=1|D^*,X,Z) &= D^* + (1-D^*)\frac{P(D^*=0|X,D=1)P(D=1|Z)}{\sum_d P(D^*=0|X,D=d)P(D=d|Z)} \\ &= D^* + (1-D^*)\frac{P(D^*=0|X,D=1)P(D=1|Z)}{P(D=0|Z)+P(D^*=0|X,D=1)P(D=1|Z)} \\ &= D^* + (1-D^*)\frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}+e^{\beta_0+\beta_X X_i}} \end{aligned}$$

In the M-step, we maximize the following expected log-likelihood with respect to θ and β:

$$Q = \sum_i p_i \text{log}\left[\frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}\right] + (1-p_i)\text{log}\left[\frac{1}{1+e^{\theta_0+\theta_Z Z_i}}\right]$$
$$+ D_i^* p_i \text{log}\left[\frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}}\right] + (1-D_i^*)p_i\text{log}\left[\frac{1}{1+e^{\beta_0+\beta_X X_i}}\right]$$

In practice, this can be accomplished by (1) fitting a logistic regression with p_i as the outcome and Z_i as covariates and (2) fitting a logistic regression with D_i^* given X_i weighted by p_i.

A.6.4 Incorporating weights into the algorithms

We can address selection bias and misclassification simultaneously by maximizing a weighted version of the observed data log-likelihood, called a pseudo log-likelihood, as follows:

$$\sum_i \omega_i D_i^* \text{log}\left[\frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}}\frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}\right] + \omega_i(1-D_i^*)\text{log}\left[1-\frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}}\frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}\right]$$

We can similarly estimate θ using a weighted version of the above EM algorithm. In particular, let ω_i be our weights. In the E-step, we replace D_i as before. In the M-step, we maximize the following expected pseudo log-likelihood

$$Q = \sum_i \omega_i p_i \text{log}\left[\frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}\right] + \omega_i(1-p_i)\text{log}\left[\frac{1}{1+e^{\theta_0+\theta_Z Z_i}}\right]$$
$$+ \omega_i D_i^* p_i \text{log}\left[\frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}}\right] + \omega_i(1-D_i^*)p_i\text{log}\left[\frac{1}{1+e^{\beta_0+\beta_X X_i}}\right]$$

Similar to before, we can obtain estimates of θ and β in the M-step by (1) fitting a logistic regression for p_i given Z_i weighted by ω_i and (2) fitting a logistic regression for D_i^* given X_i weighted by $p_i \times \omega_i$.

Justification for the usual EM algorithm is based on properties of likelihoods. In the weighted example, however, we no longer are working with a valid likelihood. Therefore, convergence properties are not immediately clear. However, this strategy can be justified under literature exploring a variant of the EM algorithm called the expectation-solution (ES) algorithm. In this variant, we transform the problem from maximizing a log-likelihood to solving corresponding score equations. Theoretical properties of the ES algorithm are explored in Elashoff (2004) and Rosen (2000).

A more challenging concern is estimation of the covariance matrix. Since we are no longer maximizing a valid observed data log-likelihood, we can no longer rely on the observed data information matrix directly. Instead, we apply the following commonly-used sandwich estimation strategy (e.g. as implemented by the R package *sandwich*). First, we define the "bread" of the sandwich matrix as follows

$$B(\theta,\beta) = \left[\sum_i \omega_i \frac{1}{K_i(\theta,\beta)[1-K_i(\theta,\beta)]} \frac{\partial K_i(\theta,\beta)}{\partial[\theta,\beta]}^{\otimes 2}\right]^{-1}$$

This is the inverse of a weighted version of the information matrix for the observed data log-likelihood of interest. For the "meat" of the sandwich estimator, we express the weighted variance of the observed data score matrix as follows:

$$M(\theta,\beta) = \sum_i \left[\omega_i \frac{D_i^* - K_i(\theta,\beta)}{K_i(\theta,\beta)[1-K_i(\theta,\beta)]} \frac{\partial K_i(\theta,\beta)}{\partial[\theta,\beta]}\right]^{\otimes 2}$$

Using these components, we express

$$Var([\hat{\theta},\hat{\beta}]) = B(\hat{\theta},\hat{\beta})M(\hat{\theta},\hat{\beta})B(\hat{\theta},\hat{\beta})$$

Suppose we perform this estimation fixing β_0. We then obtain corresponding standard errors for the other parameters by calculating B and M excluding the column and row corresponding to β_0. In the case of B, we exclude this column and row prior to inverting the weighted matrix. In simulations, this estimator resulted in nominal coverage.

A.7 Proof of *Eq. 7* and *Eq. 10*

A.7.1 Assuming no phenotype misclassification

In this section, obtain expressions for estimating $P(S = 1|D, W)$ that can be used to obtain IPW weights as discussed in **Section 4.1**. First, we review some notation. Suppose we have an external dataset with corresponding selection indicator S_{ext}. For this dataset, we assume we know the selection mechanism $P(S_{ext} = 1|D, W)$ or have corresponding selection weights from which to estimate the selection mechanism. For now, we will assume that the population used to define $P(S_{ext} = 1|D, W)$ is the same as our target population. Define S_{all} to take the value 1 if $S = 1$ or if $S_{ext} = 1$. When individual subjects are in both datasets, all three selection indicators equal 1 ($S = S_{ext} = S_{all} = 1$). For now, we allow such overlap in our non-probability sample (internal data) and probability sample (external data).

We first note that

$$P(S = 1|D, W) = \frac{P(D, W|S = 1)P(S = 1)}{P(D, W)}$$
$$P(S_{ext} = 1|D, W) = \frac{P(D, W|S_{ext} = 1)P(S_{ext} = 1)}{P(D, W)}$$

Putting those pieces together, we have

$$P(S = 1|D, W) = P(S_{ext} = 1|D, W)\frac{P(D, W|S = 1)P(S = 1)}{P(D, W|S_{ext} = 1)P(S_{ext} = 1)} \qquad (\textit{Eq. S4})$$

In *Eq. S4*, we relate the selection mechanism of interest ($P(S = 1|D, W)$) to the known selection model for the probability sample and to joint distributions of D and W in the internal and external samples, which can be estimated. We can use this expression directly to obtain estimates of $P(S = 1|D, W)$. We note some parallels between this expression and calibration weighting. Firstly, we **do not require individual-level data** on D and W from the probability sample to use this expression as long as the joint distribution in the probability sample is known along with the corresponding selection model for the probability sample. Secondly, *Eq. S4* involves the ratio of joint distributions for the internal data and some external data. Unlike usual calibration weighting, this joint distribution comes from a probability sample rather than the population of interest, and the resulting estimator is modified by $P(S_{ext} = 1|D, W)$ to account for this. Under simple random sampling for the external dataset, this expression produces weights proportional to those obtained through poststratification as discussed in **Section 4**.

Eq. S4 can be used to estimate IPW weights, but it may be unappealing to model or estimate the joint distributions of D and W when W is high-dimensional. The following expressions may be very useful in this case. We first recall that patients in the combined internal and external datasets may fall into one of three groups: (1) $S = 1$ and $S_{ext} = 1$, (2) $S = 0$ and $S_{ext} = 1$, and (3) $S = 1$ and $S_{ext} = 0$. We have that

$$P(S = j, S_{ext} = k|D, W, S_{all} = 1) = \frac{P(S = j, S_{ext} = k, D, W|S_{all} = 1)}{P(D, W|S_{all} = 1)}$$
$$= \frac{P(D, W|S = j, S_{ext} = k, S_{all} = 1)P(S = j, S_{ext} = k|S_{all} = 1)}{\sum_{ab\in(10,01,11)} P(D, W|S = a, S_{ext} = b, S_{all} = 1)P(S = a, S_{ext} = b|S_{all} = 1)}$$
$$= \frac{P(D, W|S = j, S_{ext} = k)P(S = j, S_{ext} = k)}{\sum_{ab\in(10,01,11)} P(D, W|S = a, S_{ext} = b)P(S = a, S_{ext} = b)}$$

for $jk \in (10, 01, 11)$. Defining

$$\mu_{jk} = P(D, W|S = j, S_{ext} = k)$$
$$\alpha_{jk} = P(S = j, S_{ext} = k)$$

$$p_{jk} = P(S = j, S_{ext} = k|D, W, S_{all} = 1)$$

we have

$$p_{jk} = \frac{\mu_{jk}\alpha_{jk}}{\mu_{11}\alpha_{11} + \mu_{01}\alpha_{01} + \mu_{10}\alpha_{10}} \implies \frac{\mu_{11}\alpha_{11}}{p_{11}} = \frac{\mu_{10}\alpha_{10}}{p_{10}} = \frac{\mu_{01}\alpha_{01}}{p_{01}}$$

We also note that

$$\begin{aligned} P(D, W|S = 1) &= \sum_b P(D, W|S = 1, S_{ext} = b)P(S_{ext} = b|S = 1) \\ &= \frac{\mu_{11}\alpha_{11} + \mu_{10}\alpha_{10}}{P(S = 1)} = \frac{\mu_{11}\alpha_{11} + \mu_{11}\alpha_{11}\frac{p_{10}}{p_{11}}}{P(S = 1)} \\ P(D, W|S_{ext} = 1) &= \sum_a P(D, W|S = a, S_{ext} = 1)P(S = a|S_{ext} = 1) \\ &= \frac{\mu_{11}\alpha_{11} + \mu_{01}\alpha_{01}}{P(S_{ext} = 1)} = \frac{\mu_{11}\alpha_{11} + \mu_{11}\alpha_{11}\frac{p_{01}}{p_{11}}}{P(S_{ext} = 1)} \end{aligned}$$

Replacing these terms in *Eq. S4*, we have that

$$P(S = 1|D, W) = P(S_{ext} = 1|D, W)\frac{p_{11} + p_{10}}{p_{11} + p_{01}}$$

Replacing our notation, we have

$$\begin{aligned} &P(S = 1|D, W) \qquad (\textit{Eq. S5}) \\ &= P(S_{ext} = 1|D, W)\frac{P(S = 1, S_{ext} = 1|D, W, S_{all} = 1) + P(S = 1, S_{ext} = 0|D, W, S_{all} = 1)}{P(S = 1, S_{ext} = 1|D, W, S_{all} = 1) + P(S = 0, S_{ext} = 1|D, W, S_{all} = 1)} \end{aligned}$$

This expression depends on terms that we can easily estimate given the external and internal data. In particular, we can fit a **multinomial regression** for whether a person is in the internal dataset only, external dataset only, or both given D and W.

Now, suppose further that our population is so large and sampling fractions are small enough such that no people are included in both the internal and external samples. In this case, we can approximate $P(S = 1|D, W)$ as follows:

$$P(S = 1|D, W) \approx P(S_{ext} = 1|D, W)\frac{P(S = 1|D, W, S_{all} = 1)}{1 - P(S = 1|D, W, S_{all} = 1)} \qquad (\textit{Eq. S6})$$

Eq. S6 is used in Elliot (2009) to account for non-probability selection using an external probability sample. Here, we extend this expression to allow for overlap between the internal and external datasets through *Eq. S5* and accompanying multinomial regression modeling of selection.

Of course, W may not be available in practice for either the internal or external datasets, and a subset, W_{sub}, must be used in its place. We would effectively be approximating $P(S = 1|D, W)$ using available $P(S = 1|D, W_{sub})$. In this case, we may hope to weight our analysis to *reduce* the bias rather than expecting it to *remove* bias.

A.7.2 Relationship between selection model and calibration weights

In the main paper, we describe how we can use summary statistics on D and W (or possibly a subset W_{sub}) to define poststratification weights as follows:

$$\omega \propto \frac{f(D,W)}{f(D,W|S=1)}$$

To help clarify the link between poststratification weights and inverse probability of selection weights, we note the following:

$$P(S=1|D,W) = \frac{f(D,W,S=1)}{f(D,W)} = \frac{f(D,W|S=1)P(S=1)}{f(D,W)}$$

If we were to define inverse probability of selection weights using the above expression, we would define

$$\omega \propto \frac{1}{P(S=1|D,W)} \propto \frac{f(D,W)}{f(D,W|S=1)}$$

These weights take the exact same form as the poststratification weights, so we can view poststratification weights as a similar type of weight as inverse probability of selection weights but using different types of information (individual patient data vs. summary statistics) to estimate $P(S=1|D,W)$.

A.7.3 Assuming phenotype misclassification

Now, we suppose that we have phenotype misclassification, so D is not observed. We will further assume that we can write specificity $b(Z) = \widetilde{b}$, where $\widetilde{b}$ is a known constant. We further assume that specificity is a constant in W, so $P(D^*=0|D=0,W,S=1) = P(D^*=0|D=0,S=1) = \widetilde{b}$. Ideally, we would obtain selection bias adjustment weights using the inverse of $P(S=1|D,W)$, but this is difficult to estimate in this setting. Instead, the best we can do is weight by the inverse of $P(S=1|D^*,W)$. We observe the following

$$P(S=1|D^*,W) = \frac{f(D^*|S=1,W)}{f(D^*|W)}P(S=1|W)$$

The first term, $f(D^*|S=1,W)$, can be estimated by modeling D^* directly using the observed data. $P(S=1|W)$ can be estimated using the methods in **Section A.7.1** but only conditioning on W rather than W and D. If D^* is measured on the external probability sample, then $f(D^*|W)$ can also be estimated directly. Usually, however, our external dataset may have D measured. In this case, we can estimate $f(D^*|W)$ using that

$$P(D^*=1|W) = \sum_{d=0,1} P(D^*=1|W,D=d)P(D=d|W)$$

If we further assume that D^* is independent of S given D and W (as we do in *Eq. 1* by specifying that selection independently depends on W and D but not D^*), we have that

$$P(D^*=1|W) = P(D^*=1|W,D=1,S=1)P(D=1|W) + (1-\widetilde{b})P(D=0|W)$$

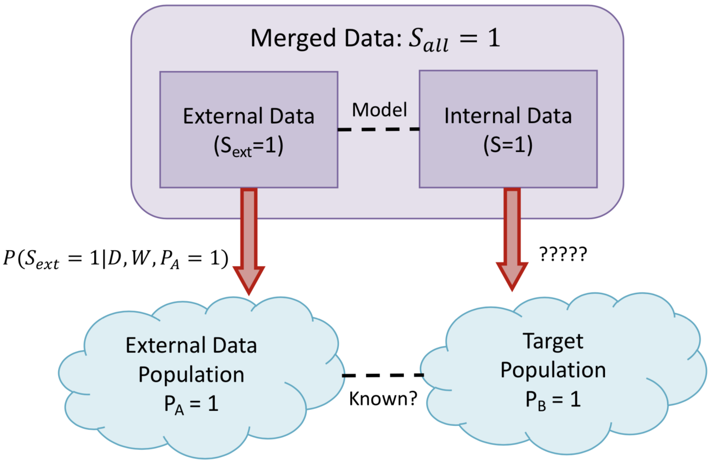
In practice, we will approximate $P(D^*=1|W,D=1,S=1)$ with either the (1) marginal sensitivity estimate $\widetilde{c}$ or (2) sensitivity estimate $c_{true}(X)$. The former substitution can be applied when either sensitivity is a constant or X is independent of W given $D=1$ and $S=1$. The reasonableness of this assumption has been discussed previously. The impact of this substitution when X and W are related is briefly demonstrated in simulations later on (**Figure B.14**). Briefly, we should use $c_{true}(X)$ instead of $\tilde{c}$ when Z is related to X and W given D, which induces a relationship between X and W given D. As before, we might not always have W measured in the internal and external datasets in practice, and we might approximate the above distributions using available predictors, W_{sub}.

A.7.4 External data with different target population

It may often be the case that the target population used for the external data is different than our desired target population. A natural question, then, is to what extent we can still apply the expressions in **Section A.7.1** under different target populations. This is important, because it impacts how we incorporate the selection probabilities or weights provided for the probability sample.

Suppose we define two populations, population A and population B. These populations may overlap or may not. Let P_A and P_B be indicators corresponding to whether a random person in some shared base population is included in population A or population B respectively. We define our goal selection probability as $P(S = 1|D, W, P_B = 1)$ and suppose $P(S_{ext} = 1|D, W, P_A = 1)$ is known or estimable using the probability sample. **Figure A.2** provides a visualization of our notation and modeling setting.

Figure A.2: Visualizing an external probability sample with different target population



The main idea is that we want to relate our observed data to our target population using an external dataset. In doing so, we will need to consider the relationship between the two populations associated with the external data and our target analysis. We have that

$$P(S = 1|D, W, P_B = 1) = \frac{P(D, W|S = 1, P_B = 1)P(S = 1|P_B = 1)}{P(D, W|P_B = 1)}$$
$$P(S_{ext} = 1|D, W, P_A = 1) = \frac{P(D, W|S_{ext} = 1, P_A = 1)P(S_{ext} = 1|P_A = 1)}{P(D, W|P_A = 1)}$$

Suppose we assume that $P(D, W|P_A = 1) = P(D, W|P_B = 1)$, so we set the denominators on the righthand side of the two equations above to be equal. Replacing the second equation into the denominator of the first equation and ignoring terms that do not depend on D and/or W, we have that

$$P(S = 1|D, W, P_B = 1) \propto P(S_{ext} = 1|D, W, P_A = 1)\frac{P(D, W|S = 1, P_B = 1)}{P(D, W|S_{ext} = 1, P_A = 1)}$$

The expression follows the same form as *Eq. S4*, and we can obtain expressions similar to *Eq. S5* and *Eq. S6* as well. In other words, we can apply the expressions in **Section A.7.1 ignoring the different populations** assuming that $P(D, W|P_A = 1) = P(D, W|P_B = 1)$. This will occur when the two populations differ with respect to factors that are independent of D and W.

Suppose we cannot assume $P(D, W|P_A = 1) = P(D, W|P_B = 1)$. This is the more common case. Suppose instead that **population A contains population B**, so the target population is a subset of the larger population associated with the probability sample. We have that

$$P(S=1|D,W,P_B=1) = \frac{P(D,W|S=1,P_B=1)P(S=1|P_B=1)P(P_B=1)}{P(P_B=1|D,W)P(D,W)}$$

$$P(S_{ext}=1|D,W,P_A=1) = \frac{P(D,W|S_{ext}=1,P_A=1)P(S_{ext}=1|P_A=1)P(P_A=1)}{P(P_A=1|D,W)P(D,W)}$$

Putting these expressions together, we have that

$$\begin{aligned} &P(S=1|D,W,P_B=1) \\ &\propto P(S_{ext}=1|D,W,P_A=1)\frac{P(P_A=1|D,W)}{P(P_B=1|D,W)}\frac{P(D,W|S=1,P_B=1)}{P(D,W|S_{ext}=1,P_A=1)} \\ &\propto P(S_{ext}=1|D,W,P_A=1)\frac{P(D,W|P_A=1)}{P(D,W|P_B=1)}\frac{P(D,W|S=1,P_B=1)}{P(D,W|S_{ext}=1,P_A=1)} \qquad (\textit{Eq. S7}) \end{aligned}$$

If either (1) $P(P_A=1|D,W)$ and $P(P_B=1|D,W)$ (i.e. the probabilities that a person in some shared base population in included in each sub-population) or (2) $P(D,W|P_A=1)$ and $P(D,W|P_B=1)$ are known, we can apply *Eq. S7* to estimate $P(S=1|D,W,P_B=1)$. In other words, we could theoretically handle the problem of different populations for the target analysis and the external probability sample if we understand how the two **populations** differ in terms of D and W. Intuitively, this is akin to incorporating calibration weighting accounting for differences between the populations. We note, however, that the second form of *Eq. S7* relies on $P(D,W|P_B=1)$. If this were known, we could just apply calibration weighting relating our internal data to population B and ignore the external dataset entirely.

Since population B is assumed to be a subset of population A, we could instead define the larger shared base population as population A, so $P(P_A=1|D,W)=1$. In this case, we have that $P(P_B=1|D,W)=P(P_B=1|D,W,P_A=1)$. We may have some sense of how population B relates to larger population A. For example, population A may represent all adults over 50 in the US, and population B may represent all adults age 50-65 in the US. In this case, we have that

$$\begin{aligned} &P(S=1|D,W,P_B=1) \qquad (\textit{Eq. S8}) \\ &\propto \frac{1}{P(P_B=1|D,W,P_A=1)}P(S_{ext}=1|D,W,P_A=1)\frac{P(D,W|S=1,P_B=1)}{P(D,W|S_{ext}=1,P_A=1)} \end{aligned}$$

This expression does not require us to know the joint distribution of D and W in either population and involves one more term (i.e., $P(P_B=1|D,W,P_A=1)$) compared to expressions assuming the populations are the same. We can then derive expressions similar to *Eq. S5* and *Eq. S6* in this setting, avoiding the need to model the joint distributions of D and W entirely.

A.8 Proof of *Eq. 8*

In this section, we develop an expression to relate $\widetilde{r}$ to $\widetilde{c}$ and $\widetilde{b}$. We have that

$$\widetilde{r} = \frac{P(S=1|D=1)}{P(S=1|D=0)} = \frac{P(D=1|S=1)}{P(D=0|S=1)}\frac{P(D=0)}{P(D=1)}$$

Now, we also have that

$$\begin{aligned}
&P(D^*=1|S=1) = \sum_d P(D^*=1|S=1,D=d)P(D=d|S=1)\\
&= \widetilde{c}P(D=1|S=1) + (1-\widetilde{b})P(D=0|S=1)\\
&\implies P(D=1|S=1) = \frac{P(D^*=1|S=1)-(1-\widetilde{b})}{\widetilde{c}-(1-\widetilde{b})}\\
&\implies P(D=0|S=1) = \frac{\widetilde{c}-P(D^*=1|S=1)}{\widetilde{c}-(1-\widetilde{b})}
\end{aligned}$$

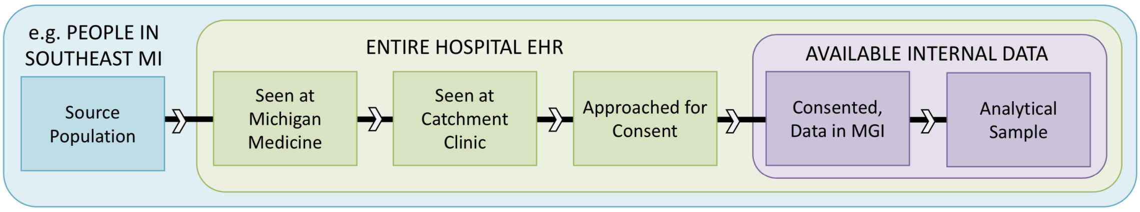
Putting these pieces together, we have

$$\widetilde{r} = \frac{P(D^*=1|S=1)-(1-\widetilde{b})}{\widetilde{c}-P(D^*=1|S=1)}\frac{P(D=0)}{P(D=1)}$$

A.9 Combining multiple complicated selection mechanisms

As discussed in Haneuse and Daniels (2016), the mechanism governing patient selection in our EHR analytical dataset may be complicated and composed of many different sub-mechanisms. **Figure A.3** provides a visualization of the various selection stages generating patient inclusion in MGI.

Figure A.3: Stages of selection from source population to analytical sample in MGI

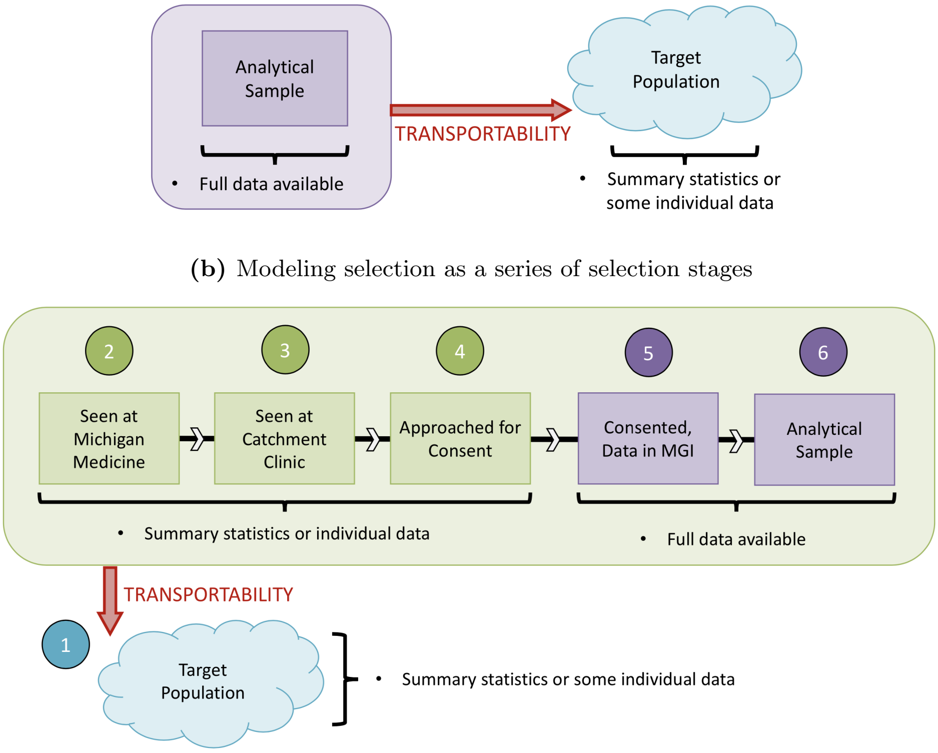


We consider two over-arching conceptual strategies for addressing selection. In the first, we view the selection probability as an aggregate across all these various selection stages and relate our analytical sample to our target population using a single selection model as in **Figure A.4a**. Of course, this has corresponding transportability assumptions involved, since the source population (the population from which we sample) and the target population (the population we want to make inference about) may not be the same. A second approach (as in Haneuse and Daniels (2016)) is to model individual selection stages separately. For example, we may model individual selection stages patients pass through to be selected into the analytical dataset from the broader hospital EHR. Then, we relate the hospital EHR patients to our target population. This approach is visualized in **Figure A.4b**. An advantage of this latter approach is the ability to incorporate prior knowledge and additional data available for individual selection stages.

Figure A.4: Stages of selection from source population to analytical sample in MGI

(a) Viewing selection as a single mechanism

(b) Modeling selection as a series of selection stages



We can define a set of intermediate selection indicators corresponding to different stages of selection. In the MGI example, let S_1 indicate whether a patient was seen at Michigan Medicine, let S_2 indicate whether the patient visited a clinic involved in MGI recruitment, let S_3 indicate whether the patient was approached for consent, let S_4 indicate whether the patient was included in MGI, and let S_5 indicate whether the patient was included in our analytical sample. We define these indicators such that $S_k = 1$ only if $S_{k-1} = 1$. Define $S_0 = 1$ for all people in the target population. The target population could vary based on our analysis, and we may often suppose it could be the Michigan or US adult population. This target population could even be the set of people seen at Michigan Medicine, but for generality we will assume a broader target population. In this updated notation, the overall sampling indicator S corresponds to S_5 and can be written as

$$P(S = 1|D, W) = \prod_{k=1}^{5} P(S_k = 1|D, W, S_{k-1} = 1) \qquad (\textit{Eq. S9})$$

First, we make a distinction between the selection model $P(S_1 = 1|D, W, S_0 = 1) = P(S_1 = 1|D, W)$, which corresponds to the relationship between the EHR and the target population, and the other models. For the selection models conditional on inclusion in the EHR, we may have available individual-level data from which to estimate each $P(S_k = 1|D, W, S_{k-1} = 1)$ as is done in Haneuse and Daniels (2016). In contrast, we will not have individual-level data on everyone in the target population for modeling $P(S_1 = 1|D, W, S_0 = 1)$. A particular challenge, therefore, is linking the EHR population to the target population through $P(S_1 = 1|D, W)$.

In **Section 4** of the main paper, we describe how we can estimate $P(S = 1|D, W)$ either using a probability sample from the target population or using summary statistics from the target population. Here, we can apply the same approach to estimate $P(S_1 = 1|D, W)$ or corresponding calibration weights, viewing the $S_1 = 1$ sample as the "internal" data. We can then work to estimate each stage of selection within the EHR $S_1 = 1$ sample using the available data.

In an ideal world, we could model each selection step within the EHR separately using data on D and W from every individual in the population. In practice, we may not have individual-level data on all patients in the EHR $S_1 = 1$ sample. Even if we do, we may not have as much detailed data available for patients in the larger EHR compared to patients in the analytical sample. In MGI, for example, we have a much wider spectrum of patient information available for patients in MGI than in the entire Michigan Medicine sample. Since we do not expect W and D to be necessarily available for all patients in the $S_1 = 1$ sample (or indeed, the final analytical sample), we can take a patchwork approach to estimating each component selection model using the information that is available for every stage of the selection. Our ultimate goal is to generate selection weights that may help *reduce* selection bias.

Often, we may only have individual-level data for some stages of selection and summary statistics for others. When we have individual-level data on all $S_{k-1} = 1$ patients, we can directly model $P(S_k|W_k, S_{k-1} = 1)$ using the set of covariates W_k available in the S_{k-1} and S_k samples. When only summary statistics are available for the S_{k-1} or S_k samples, we can take a calibration weighting approach and estimate $P(S_k = 1|D, W, S_{k-1} = 1)$ proportional to $\frac{P(D, W_k|S_k=1)}{P(D, W_k|S_{k-1}=1)}$. In this way, we can estimate each component piece in *Eq. S9* using the most information possible at each stage. By combining traditional modeling of multi-stage selection with strategies for relating non-probability samples to the target population, we can bridge the gap between our analytical sample and the target population in our data analyses.

A.10 Estimating standard errors

In the main paper, we develop statistical methods for obtaining bias-corrected point estimates for θ, but we do not directly address estimation of corresponding standard errors. Here, we describe how this can be done, appealing to existing results in the maximum likelihood estimation and survey sampling literature. We focus on the setting where we have perfect specificity, and the setting with imperfect specificity is similar. **Table A.1** provides details about the proposed variance estimators for each of the bias-correction methods proposed in this paper. We provide estimators for each one of the bias-correction methods treating estimated sensitivity and/or IPW/calibration weights ω as fixed. In the footnote, we describe how we can account for additional uncertainty due to estimating sensitivity and/or ω using bootstrap methods. Derivations motivating these variance estimators can be found elsewhere in the text (e.g. **Web Appendices A.3 and A.6** and below).

Table A.1: Strategies for estimating standard errors for $\hat{\theta}$ *

Bias	Method
Misclass.	Approximating $D^* \vert Z$ distribution (Section 3.1) • $\text{Var}(\hat{\theta}_Z) \approx \text{Var}(\hat{\theta}_Z^{uc}) \left[\frac{\widetilde{c}(1-P(D^*=1))}{\widetilde{c}-P(D^*=1)}\right]^2$ where $\hat{\theta}_Z^{uc}$ is the uncorrected log-odds ratio. • $\text{Var}(\hat{\theta}_Z) \approx \hat{\theta}_Z^{uc\,2} \left[\frac{P(D^*=1)[1-P(D^*=1)]}{E(\widetilde{c})-P(D^*=1)}\right]^2 \text{Var}(\widetilde{c}) + \text{Var}\left(\hat{\theta}_Z^{uc}\right) \left[\frac{E(\widetilde{c})[1-P(D^*=1)]}{E(\widetilde{c})-P(D^*=1)}\right]^2$
Misclass.	Non-logistic link function (Section 3.2) • $\text{Var}(\hat{\theta}) = \left[\sum_i \frac{c(Z)}{1+[1-c(Z)]e^{\theta_0+\theta_Z Z}} \frac{e^{\theta_0+\theta_Z Z}}{(1+e^{\theta_0+\theta_Z Z})^2}(1,Z)^{\otimes 2}\right]^{-1}$ where we replace $c(Z)$ with an estimate.
Misclass.	Obs. data log-likelihood (Section 3.3) • Using the expected obs. data information matrix, we have $\text{Var}(\hat{\theta}) = \left[\sum_i \frac{1}{K_i(\theta,\beta)[1-K_i(\theta,\beta)]} \frac{\partial K_i(\theta,\beta)}{\partial(\theta,\beta)} \frac{\partial K_i(\theta,\beta)}{\partial(\theta,\beta)^T}\right]^{-1}$ where $K_i(\theta,\beta) = \frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}} \frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}$
Selection	Weighting by ω (Section 4) • Apply Huber-White sandwich estimator with survey weights as implemented in R package *survey* (Freedman, 2006).
Both	Approximating $D^* \vert Z$ distribution + weighting (Section 5.1) • We can use the same general variance structure as in the unweighted case except we estimate θ_Z^{uc} using a weighted regression model fit with Huber-White standard errors. We also replace $P(D^*=1)$ with $p^* = \frac{\sum_i \omega_i D_i^*}{\sum_i \omega_i}$
Both	Non-logistic link function + weighting (Section 5.2) • We can again apply the Huber-White sandwich estimator with survey weights as implemented in R package *survey* (Freedman, 2006), except this time we specify a non-logistic link function for the estimation and define the meat and bread matrices corresponding to the modified link function given $c_{true}(X)$.
Both	Obs. data log-likelihood + weighting (Section 5.3) • We no longer have a valid likelihood, and we apply the following sandwich estimator We have $B(\theta,\beta) = \left[\sum_i \omega_i \frac{1}{K_i(\theta,\beta)[1-K_i(\theta,\beta)]} \frac{\partial K_i(\theta,\beta)}{\partial[\theta,\beta]} \frac{\partial K_i(\theta,\beta)}{\partial[\theta,\beta]^T}\right]^{-1}$ $M(\theta,\beta) = \sum_i \left[\omega_i \frac{D_i^*-K_i(\theta,\beta)}{K_i(\theta,\beta)[1-K_i(\theta,\beta)]} \frac{\partial K_i(\theta,\beta)}{\partial[\theta,\beta]}\right]^{\otimes 2}$ and $Var([\hat{\theta},\hat{\beta}]) = B(\hat{\theta},\hat{\beta})M(\hat{\theta},\hat{\beta})B(\hat{\theta},\hat{\beta})$

* Many of the above estimators treat sensitivity and/or IPW/calibration weights ω as fixed and do not take into account the uncertainty in estimating sensitivity or ω. One could account for this uncertainty through bootstrap methods, where sensitivity, ω, and θ are estimated for each of many bootstrap samples of the data. The resulting distribution of $\hat{\theta}$ can then be used to obtain standard errors.

A.10.1 Comparison between naive and misclassification-corrected standard errors

In this section, we focus on the setting where we have misclassification and where selection is ignorable. We want to compare the magnitude of the standard errors obtained using the various bias-correction strategies amongst each other. We also will compare these bias-correction strategies to naive analysis.

Naive: We suppose we fit a logistic regression model to the observed data and treat the resulting parameters as if they were θ. The structure of the resulting expected information matrix is as follows:

$$I_{uc}(\theta) = \sum_i \frac{e^{\theta_0+\theta_Z Z}}{(1+e^{\theta_0+\theta_Z Z})^2}(1,Z)^{\otimes 2}$$

Approximation of $D^*|Z$ method: The variance estimation equation for $\hat{\theta}_Z$ from approximating the $D^*|Z$ distribution is $\text{Var}(\hat{\theta}_Z) \approx \text{Var}(\hat{\theta}_Z^{uc})\left[\frac{\widetilde{c}(1-P(D^*=1))}{\widetilde{c}-P(D^*=1)}\right]^2$. Since $\widetilde{c}$ and $P(D^*=1)$ are both strictly less than 1 under imperfect sensitivity, we have that $\text{Var}(\hat{\theta}_Z) > \text{Var}(\hat{\theta}_Z^{uc})$. Additionally, we can write the expected information matrix implied by this model as a function of θ as follows:

$$I_{approx}(\theta) = \left[\frac{\widetilde{c}(1-P(D^*=1))}{\widetilde{c}-P(D^*=1)}\right]^{-2} \sum_i \frac{e^{\theta_0+\theta_Z Z}}{(1+e^{\theta_0+\theta_Z Z})^2}(1,Z)^{\otimes 2}$$

Non-logistic link function method: Consider the likelihood function corresponding to the distribution of $D^*|Z$ and its relationship to θ and $c(Z)$ as follows:

$$L = \prod_i \left[c(Z)\frac{e^{\theta_0+\theta_Z Z}}{1+e^{\theta_0+\theta_Z Z}}\right]^{D^*}\left[1-c(Z)\frac{e^{\theta_0+\theta_Z Z}}{1+e^{\theta_0+\theta_Z Z}}\right]^{1-D^*}$$

$$\log(L) = \sum_i D^*(\theta_0+\theta_Z Z) - \log\left[1+e^{\theta_0+\theta_Z Z}\right] + (1-D^*)\log\left[1+(1-c(Z))e^{\theta_0+\theta_Z Z}\right] + \text{constant}$$

with score function

$$U(\theta) = \sum_i \left\{D^* - \frac{e^{\theta_0+\theta_Z Z}}{1+e^{\theta_0+\theta_Z Z}} + (1-D^*)\frac{(1-c(Z))e^{\theta_0+\theta_Z Z}}{1+(1-c(Z))e^{\theta_0+\theta_Z Z}}\right\}(1,Z)$$

and information matrix

$$J(\theta) = \sum_i \left\{\frac{e^{\theta_0+\theta_Z Z}}{(1+e^{\theta_0+\theta_Z Z})^2} - (1-D^*)\frac{(1-c(Z))e^{\theta_0+\theta_Z Z}}{(1+(1-c(Z))e^{\theta_0+\theta_Z Z})^2}\right\}(1,Z)^{\otimes 2}$$

This information matrix is strictly less than the information matrix for naïve logistic regression when $c(Z) < 1$. Therefore, $c(Z)$ less than 1 will result in an increase in corresponding standard errors when we correctly account for the misclassification.

We might also be interested in the expected information matrix, where we replace D^* with its expectation, $c(Z)\text{expit}(\theta_0+\theta_Z Z)$. Replacing D^* in the above equation and re-writing, we have that

$$I_{link}(\theta) = \sum_i \frac{c(Z)}{1+[1-c(Z)]e^{\theta_0+\theta_Z Z}}\frac{e^{\theta_0+\theta_Z Z}}{(1+e^{\theta_0+\theta_Z Z})^2}(1,Z)^{\otimes 2}$$

Again, this will be strictly less than the information matrix from naive analysis.

Suppose we estimate θ replacing $c(Z)$ with $c_{true}(X)$, which is a function of β. We can write the expected information matrix as a function of β as follows:

$$I_{link}(\theta) = \sum_i \frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}+e^{\theta_0+\theta_Z Z}}\frac{e^{\theta_0+\theta_Z Z}}{(1+e^{\theta_0+\theta_Z Z})^2}(1,Z)^{\otimes 2} \qquad (\textit{Eq. S10})$$

We will use this quantity later on.

Observed data log-likelihood maximization method: When we jointly estimate θ and β using the observed data log-likelihood, we have corresponding expected observed data information matrix as follows:

$$
\begin{aligned}
I_{obs}(\theta, \beta) &= \sum_i \frac{1}{K_i(\theta,\beta)[1 - K_i(\theta,\beta)]} \frac{\partial K_i(\theta,\beta)}{\partial(\theta,\beta)} \frac{\partial K_i(\theta,\beta)}{\partial(\theta,\beta)^T} \\
&= \sum_i \frac{\frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}} \frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}}{[1 - \frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}} \frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}]} \left[\frac{1}{1+e^{\theta_0+\theta_Z Z_i}}(1, Z_i), \frac{1}{1+e^{\beta_0+\beta_X X_i}}(1, X_i) \right]^{\otimes 2} \\
&= \sum_i \frac{e^{\beta_0+\beta_X X_i} e^{\theta_0+\theta_Z Z_i}}{1+e^{\beta_0+\beta_X X_i} + e^{\theta_0+\theta_Z Z_i}} \left[\frac{1}{1+e^{\theta_0+\theta_Z Z_i}}(1, Z_i), \frac{1}{1+e^{\beta_0+\beta_X X_i}}(1, X_i) \right]^{\otimes 2} \\
&= \left[\begin{array}{lll} \theta; & \sum_i \frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}+e^{\theta_0+\theta_Z Z_i}} \frac{e^{\theta_0+\theta_Z Z_i}}{(1+e^{\theta_0+\theta_Z Z_i})^2}(1, Z_i)^{\otimes 2} & \sum_i \frac{e^{\beta_0+\beta_X X_i} e^{\theta_0+\theta_Z Z_i}}{1+e^{\beta_0+\beta_X X_i}+e^{\theta_0+\theta_Z Z_i}} \frac{(1,Z_i)(1,X_i)^T}{(1+e^{\beta_0+\beta_X X_i})(1+e^{\theta_0+\theta_Z Z_i})} \\ \beta; & \sum_i \frac{e^{\beta_0+\beta_X X_i} e^{\theta_0+\theta_Z Z_i}}{1+e^{\beta_0+\beta_X X_i}+e^{\theta_0+\theta_Z Z_i}} \frac{(1,Z_i)^T(1,X_i)}{(1+e^{\beta_0+\beta_X X_i})(1+e^{\theta_0+\theta_Z Z_i})} & \sum_i \frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\beta_0+\beta_X X_i}+e^{\theta_0+\theta_Z Z_i}} \frac{e^{\beta_0+\beta_X X_i}}{(1+e^{\beta_0+\beta_X X_i})^2}(1, X_i)^{\otimes 2} \end{array} \right]
\end{aligned}
$$

Now, we appeal to results in the linear algebra literature to relate the corresponding covariance matrix with the covariance matrix we would obtain if we fit the naive, uncorrected model. Denote the terms in I_{obs} as

$$
I_{obs}(\theta, \beta) = \left[\begin{array}{cc} A & B \\ B^T & D \end{array} \right]
$$

Assuming D is invertible, we have that

$$
[I_{obs}(\theta, \beta)]^{-1} = \left[\begin{array}{ll} (A - BD^{-1}B^T)^{-1} & -(A - BD^{-1}B^T)^{-1}BD^{-1} \\ -D^{-1}B^T(A - BD^{-1}B^T)^{-1} & D^{-1} + D^{-1}B^TA - BD^{-1}B^T)^{-1}BD^{-1} \end{array} \right]
$$

following Lu and Shiou (2002). Now, let's take a closer look at the element corresponding to the covariance matrix of $\hat{\theta}$, $(A - BD^{-1}B^T)^{-1}$. Using properties of the inverse of sums of matrices, we have that

$$
(A - BD^{-1}B^T)^{-1} = A^{-1} + \frac{1}{1 - \text{trace}(BD^{-1}B^TA^{-1})} A^{-1}BD^{-1}B^TA^{-1}
$$

Assuming D is invertible (which it is) and has non-negative diagonal elements (which it does), we have that $BD^{-1}B$ will also have non-negative diagonal elements. Assuming A is also invertible (which it is), $A^{-1}BD^{-1}B^TA^{-1}$ will also have non-negative diagonal elements. Now, we need to determine the sign of $\frac{1}{1-\text{trace}(BD^{-1}B^TA^{-1})}$. We have already concluded that $BD^{-1}B^T$ has non-negative diagonal elements. Additionally, A^{-1} is invertible and will have non-negative diagonal elements. Therefore, trace$(BD^{-1}B^TA^{-1})$ will be positive. The question remains whether it will be greater than or less than 1. We generally expect trace$(BD^{-1}B^TA^{-1})$ will be less than 1 for sufficient sample size, since A^{-1} will have small entries in this setting. We make this assertion noting that A^{-1} is equal to the inverse of $I_{link}(\theta)$ when $c(Z)$ is replaced by $c_{true}(X)$ as in *Eq. S10*. Therefore, A^{-1} is the variance of $\hat{\theta}$ when sensitivity is fixed to be equal to $c_{true}(X)$.

For sufficient sample size, we have that

$$
diag([I_{obs}(\theta; \beta)]^{-1}_{\theta,\theta}) = diag(A^{-1} + \frac{1}{1 - \text{trace}(BD^{-1}B^TA^{-1})} A^{-1}BD^{-1}B^TA^{-1}) > diag(A^{-1})
$$

where 'diag' represents the diagonal elements of the matrix.

Noting that $A = I_{link}(\theta)$ with $c(Z)$ replaced by $c_{true}(X)$, we showed previously $A^{-1} > I_{uc}(\theta)^{-1}$. Putting things together, we have that the *diagonal elements* covariance matrix associated with $\hat{\theta}$ from the observed data log-likelihood maximization follows

$$
diag([I_{obs}(\theta, \beta)]^{-1}_{\theta,\theta}) > diag(A^{-1}) > diag([I_{uc}(\theta)]^{-1})
$$

This shows that for a fixed value of θ, the standard errors will be larger under the observed data log-likelihood maximization method than the naive method. For fixed values of the corrected and uncorrected maximum likelihood estimates, however, it is possible for the standard errors to be smaller. In generally, however, we expect larger standard errors under the observed data log-likelihood method.

Overall comparisons: Putting everything together, we have the following for a fixed θ

$$diag(I_{uc}(\theta)^{-1}) < diag(I_{link}(\theta)^{-1}), diag(I_{approx}(\theta)^{-1}) < diag([I_{obs}(\theta,\beta)]^{-1}_{\theta,\theta})$$

noting that $A = I_{link}(\theta)$. This states that the standard errors for all bias correction methods will tend to be larger than the naive method and that the method using the observed data log-likelihood will tend to be the largest. This may not always be the case for a single data analysis, however, because these functions will be evaluated at different estimates for θ. In general, however, we expect the above orderings.

Overall, we expect the methods that use fixed sensitivity to produce smaller estimated standard errors than the observed data log-likelihood method (without fixed β_0). We expect this to be often true even when we account for the estimation of sensitivity for the non-logistic link function and approximation methods, since external information is incorporated into these methods. It is difficult to determine the relative orderings of standard errors for the non-logistic link function method and the method approximating the $D^*|Z$ distribution in general.

A.11 Extension to allow for probabilistic phenotyping

Most methods in the statistical literature for dealing with outcome misclassification assume either (1) sensitivity and specificity are known (e.g. Neuhaus (1999)) or (2) D and D^* are available for some validation subset of patients (e.g. Carroll (2006)). The PIE (Prior knowledge guided Integrated Estimation) method in Huang et al. (2018) provides a strategy for estimation that incorporates our prior beliefs for plausible values of sensitivity and specificity without using either the true values or a validation dataset. Although mentioned in Neuhaus (1999), very little work has been done in the setting with misclassification related to covariates. All of these methods assume the observed outcome, D^*, takes the form of a **binary variable.**

Recently, many researchers have used validation datasets to develop statistical models for the **predicted probability** of having the disease, estimated as $\hat{p}_D$, given a spectrum of electronic health record data (e.g. Castro et al. (2015)). The functional form of $\hat{p}_D$ is obtained using a validation dataset for which EHR variables X and true disease status D are known. This $\hat{p}_D$ can be viewed as the EHR-derived phenotype and can be considered a random variable when we do not condition on EHR variables X used to generate $\hat{p}_D$. Given estimated $\hat{p}_D$, a common strategy is to then apply a threshold on this probability to obtain a binary disease status outcome for analysis. In contrast, Sinnott et al. (2014) describes an approach for analyzing a **transformed** version of $\hat{p}_D$ to produce a non-binary disease status outcome for analysis. In this section, we describe how our proposed approach relates to the methods in Sinnott et al. (2014) and propose new extensions of their work to account for covariate-related misclassification and selection.

Existing method in Sinnott et al. (2014)

First, we summarize the method in Sinnott et al. (2014) using our notation. Sinnott et al. (2014) shows that we can obtain valid inference about θ by solving

$$\sum_{i=1}^{n}(1, Z_i^T)\,[Y_i - g(\theta; Z_i)] = 0 \qquad (\textit{Eq. S11})$$

for θ, where Y_i is some known transformation of $\hat{p}_D$ and $g(\theta; Z_i) = P(D_i = 1|Z_i)$ is the disease model of interest. In our case, function g is the expit function corresponding to logistic regression. This approach will provide an unbiased estimate for θ if we define function $Y(\hat{p}_D)$ such that $E(Y|Z) = E[Y(\hat{p}_D)|Z] = P(D = 1|Z)$. Assuming that $\hat{p}_D \perp Z|D$, Sinnott et al. (2014) defines

$$Y = Y(\hat{p}_D) = \frac{\hat{p}_D - E(\hat{p}_D|D=0)}{E(\hat{p}_D|D=1) - E(\hat{p}_D|D=0)}. \qquad (\textit{Eq. S12})$$

where $E(\hat{p}_D|D = 1)$ and $E(\hat{p}_D|D = 0)$ are calculated ahead of time from the validation data.

Relationship to *Eq. 4*

Suppose we do not assume that $\hat{p}_D \perp Z|D$. We can apply the same approach in Sinnott et al. (2014) to obtain the following transformation:

$$Y = Y(\hat{p}_D) = \frac{\hat{p}_D - E(\hat{p}_D|Z, D=0)}{E(\hat{p}_D|Z, D=1) - E(\hat{p}_D|Z, D=0)}$$

Suppose we define $\hat{p}_D$ such that $\hat{p}_D = D^*$ is binary. We can re-write the above expression as

$$Y = \frac{D^* - P(D^*=1|Z, D=0)}{P(D^*=1|Z, D=1) - P(D^*=1|Z, D=0)} = \frac{D^* - [1 - b(Z)]}{c(Z) - [1 - b(Z)]}$$

Under logistic regression for $g(\theta; Z_i)$, one can show that solving *Eq. S11* for θ using the above transformation Y is equivalent to the non-logistic link function method in **Section 3.2** given fixed $c(Z)$. By "equivalent", we mean that the two approaches result in the **exact same score**

equation, *Eq. S11*. Similarly, we can account for selection bias by solving a modified version of *Eq. S11* incorporating selection weights as follows:

$$\sum_{i=1}^{n} \omega_i (1, Z_i^T) \left[Y_i - g(\theta_0 + \theta_Z Z_i))\right] = 0 \qquad (\textit{Eq. S13})$$

Suppose instead that $r(Z)$ is known and $\hat{p}_D$ is allowed to be non-binary. In this case, we can estimate θ by solving *Eq. S11* using an appropriate transformation Y of $\hat{p}_D$ such that $E(Y|Z, S = 1) = g(\theta; Z)$. It is difficult to obtain an exact transformation of $\hat{p}_D$ such that $E(Y|Z, S = 1) = g(\theta; Z)$. However, we can obtain a transformation of $\hat{p}_D$ with expectation with *zero-th order approximation* $g(\theta; Z)$. Suppose we define

$$\widetilde{Y} = \frac{\hat{p}_D - E(\hat{p}_D|Z, D = 0, S = 1)}{r(Z)E(\hat{p}_D|Z, D = 1, S = 1) - E(\hat{p}_D|Z, D = 0, S = 1) - \hat{p}_D[r(Z) - 1]} \qquad (\textit{Eq. S14})$$

We can show that

$$E(\widetilde{Y}|Z, S = 1) = E\left[\frac{\hat{p}_D - E(\hat{p}_D|Z, D = 0, S = 1)}{r(Z)E(\hat{p}_D|Z, D = 1, S = 1) - E(\hat{p}_D|Z, D = 0, S = 1) - \hat{p}_D[r(Z) - 1]}|Z, S = 1\right]$$
$$\neq \frac{E(\hat{p}_D|Z, S = 1) - E(\hat{p}_D|Z, D = 0, S = 1)}{r(Z)E(\hat{p}_D|Z, D = 1, S = 1) - E(\hat{p}_D|Z, D = 0, S = 1) - E(\hat{p}_D|Z, S = 1)[r(Z) - 1]} = g(\theta; Z)$$

The expectation of a function is not equal to the function of the expectations in this case. However, the expectation of $\widetilde{Y}$ given $Z, S = 1$ is a **zero-th order approximation** to $g(\theta; Z)$. Future work can explore the performance of this approach for estimating θ in the presence of both misclassification and selection bias. In the main text, we consider a particular case where $\hat{p}_D = D^*$. If we solve *Eq. S11* using the transformation in *Eq. S14* in this setting, we can obtain an estimate of θ that is **equivalent** to solving the score equation associated with the non-logistic regression model in *Eq. 4*.

B Simulations

B.1 Simulation study set-up

The simulation study is broken up into three parts: (1) misclassification only, (2) selection bias only, and (3) misclassification and selection bias. In all simulation settings, we first generate 500 datasets with 5000 patients each. This sample of 5000 represents the true population. For each simulated dataset, we started by generating covariates Z, W, and X from a multivariate normal with mean 0, unit variances, and covariances σ_{zw}, σ_{zx}, and σ_{wx}. True disease status D was then generated using the following relation: $\text{logit}\left(P(D=1|Z)\right) = -2 + 0.5Z$. In all simulations, we had $X^{\dagger} = X$ and $W^{\dagger} = W$, so Z was not a direct driver of either misclassification or selection given D. In presenting these simulation results, we often use X and $X^{\dagger}$ interchangeably. Unless otherwise noted, all simulations assume that we have perfect specificity, so $\widetilde{b} = 1$.

Simulation part 1:
We considered several different scenarios for the relationships between X, D, and Z. The independent relationship between X and Z given D was controlled by σ_{zx} above. We allowed for additional correlation between X and D by defining $X_{new} = X_{original} + \sigma_{dx}D$, where σ_{dx} controls the strength of the relationship between D and X. We considered 4 different simulation scenarios for σ_{dx} and σ_{zx} as shown in **Table B.1**. In each scenario, we then generated D^* using the sensitivity relation $\text{logit}\left(P(D^*=1|X=X_{new}, D=1)\right) = \beta_0 + X$ and assuming perfect specificity. We performed a large number of simulations across different combinations of parameters, and we present results for $\beta_0 = -0.4$, which corresponds to marginal sensitivities $\widetilde{c}$ between roughly 0.4 and 0.45 in each simulation scenario.

Simulation part 2:
In *simulation part 2*, we define $D^* = D$. We allowed for the possibility of correlation between W and D by defining $W_{new} = W_{original} + \sigma_{dw}D$, where σ_{dw} controls the strength of the relationship between D and W. We then imposed sub-sampling to obtain our analytical sample using the following relation: $\text{logit}\left(P(S=1|W=W_{new}, D)\right) = \phi_0 + \phi_D D + \phi_W W$. We considered 4 different simulation scenarios as shown in **Table B.1**. The ϕ values were chosen to give roughly a 50% selection probability on average.

Simulation part 3:
In *simulation part 3*, we simulate data as in part 2 but also generate D^* using
$\text{logit}\left(P(D^*=1|X, D=1)\right) = 0.65 + X$ with $\sigma_{zx} = 0.5$ and $\sigma_{dx} = 0$. This corresponds to roughly a 65% marginal sensitivity with X related to Z given D. Many other simulation settings were explored with similar results, but these will not be presented here.

Methods:
For each dataset in *simulation part 1*, we corrected for misclassification bias by applying the various methods discussed in **Section 3**. Unless otherwise specified, these methods were implemented using **estimates** for sensitivity based on the simulated data. $\widetilde{c}$ was estimated as $\frac{P(D^*=1)}{P(D=1)}$. In the main paper, $c_{true}(X)$ was estimated using the method in *Eq. 6* and assuming known $P(D=1|X)$. In this **Supporting Information**, we also estimated $c_{true}(X)$ as the ratio of $P(D^*=1|X)$ and $P(D=1|X)$ instead of using *Eq. 6*. As a sensitivity analysis, we further considered the setting where $c_{true}(X)$ is estimated with unknown $P(D=1|X)$. Unless otherwise stated, implementation of the observed data log-likelihood maximization method assumed fixed intercept $\beta_0 = \text{logit}(\widetilde{c})$

In *simulation part 2*, we corrected selection bias using IPW or calibration weighting. Inverse probability weights were obtained either by fitting a model for selection using the entire population or estimated using a probability sample from that population using equations in **Section A.7**. In the main paper, weights were estimated using *Eq. S4*, and we also consider weights estimated using *Eq. S5* in this **Supporting Information**. Poststratification weights were estimated using the correct population summary statistics for W and D after binning continuous W.

For each dataset in *simulation part 3*, we corrected selection bias and bias due to phenotype misclassification using the methods discussed in **Section 5**. In the main paper, $\widetilde{c}$ and $c_{true}(X)$ were estimated using $\widetilde{r}$ fixed at the simulation truth and using $\widetilde{c} = \frac{P(D^*=1|S=1)}{P(D=1|S=1)} = P(D^* = 1|S = 1)\frac{\widetilde{r}P(D=1)+P(D=0)}{\widetilde{r}P(D=1)}$ or *Eq. 9* respectively. In this **Supporting Information**, we also explore settings where $\widetilde{r}$ is misspecified and where $c_{true}(X)$ is estimated as in *Eq. S3*. In the main paper, we present simulation results using correct IPW weights rather than sample-estimated weights. Results are similar when weights are estimated using correctly-specified $\widetilde{r}$. We explore different sample-estimated weights and their relationship to chosen $\widetilde{r}$ and sensitivity estimation.

For each simulated dataset, we apply the above methods to estimate the log-odds ratio of Z corresponding to the logistic regression for $D|Z$. In all settings, we then estimate the average and median deviation from the truth of 0.5 across the 500 simulated datasets. We also estimate coverage of 95% confidence intervals and corresponding statistical power. For each simulation setting, we also run a paired simulation where true θ_Z is set to 0, allowing us to assess false positive rates. Standard errors were estimated as discussed in **Supporting Section A.10**, treating estimated sensitivities or selection adjustment weights as fixed.

In the main paper, we present a set of three simulation studies exploring the performance of our proposed methods for handling (1) phenotype misclassification, (2) selection bias, and (3) both misclassification and selection bias. In the following sections, we provide additional explorations into these simulation study results and additional evaluation of our proposed estimators for sensitivity and sampling/calibration weights.

Table B.1: Simulation set-up

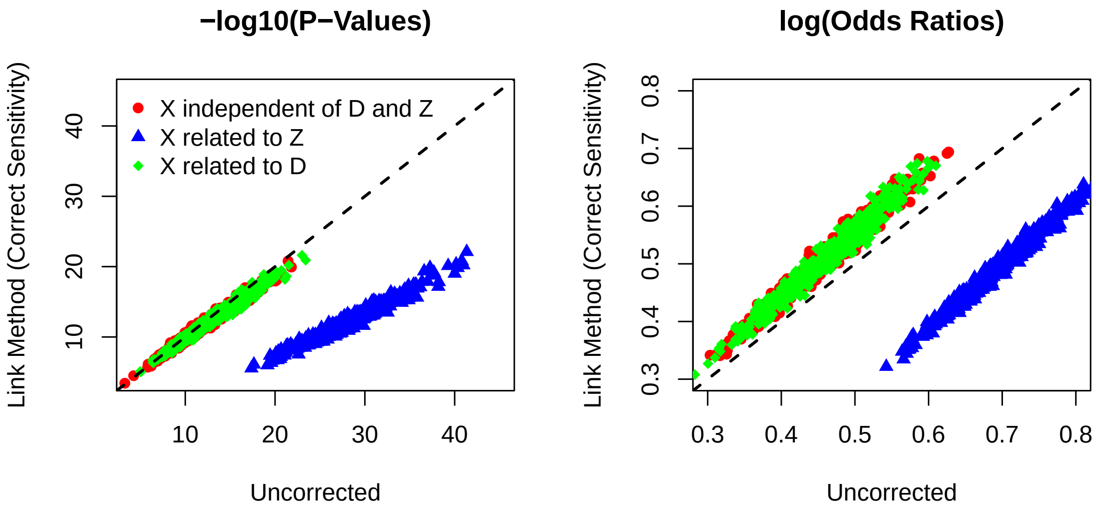
	Part 1			Part 2				
Setting	β_0	σ_{zx}	σ_{dx}	ϕ_0	ϕ_D	ϕ_W	σ_{zw}	σ_{dw}
1	-0.4	0	0	-0.6	2	-1	0.4	0
2	-0.4	0.5	0	-0.6	2	-1	0.4	1
3	-0.4	0	0.2	-0.2	0	-1	0.4	0
4	-0.4	0.5	0.5	-0.1	0	-1	0.4	1
5	0.65	0.5	0	-	-	-	-	-

▨ Part 3

B.2 Simulation part 1: p-values, power, and type I error

In the main paper, we focus on assessing bias in estimating θ, but we may also be interested in studying the impact of misclassification and our methods on the resulting p-values. **Figure B.1** shows the estimated p-values and θ_Z across 500 in Setting 1 (X independent of Z), Setting 2 (X related to Z given D), and Setting 3 (X related to D given Z) from **Table B.1**.

Figure B.1: Estimated p-values and θ_Z across 500 simulations after imposing phenotype misclassification*



* Applying method from **Section 3.2** using correct $c_{true}(X)$. True log-odds ratio is 0.5.

The left panel of **Figure B.1** demonstrates that p-values for the uncorrected and corrected analysis are nearly identical when X and Z are independent given D (Settings 1 and 3). This is consistent with existing literature in the area of outcome misclassification and was shown in the setting of link function misspecification in Li and Duan (1989). Importantly, p-values **differ** when X is related to Z given D. In Settings 1 and 3 we have that $c(Z) = \widetilde{c}$, but this is not true for Setting 2. As shown in the right panel, however, the resulting θ estimates differ between corrected and uncorrected analysis in all three settings. This figure illustrates the property that p-values are not impacted by ignoring misclassification when $c(Z) = \widetilde{c}$, but they are when $c(Z) \neq \widetilde{c}$.

Suppose our interest is in estimating p-values as in a PheWAS study, which compares p-values resulting from regression modeling of many different phenotypes, each of which has different sensitivity properties. These results indicate that there should not be a large impact of the differential misclassification across diseases on the resulting p-value comparison when X and Z are reasonably assumed to be independent given D. When X and Z may be related given D, however, accounting for misclassification across diseases can be important. When θ_Z itself is of primary interest, uncorrected analysis will produce bias in all settings with imperfect sensitivity/specificity.

Now, we take a closer look at the impact of misclassification and our corrections on type I error and power. We simulate data as before but vary the true value of θ_Z. **Figure B.2** shows the results across 500 simulated datasets corresponding to 95% confidence intervals.

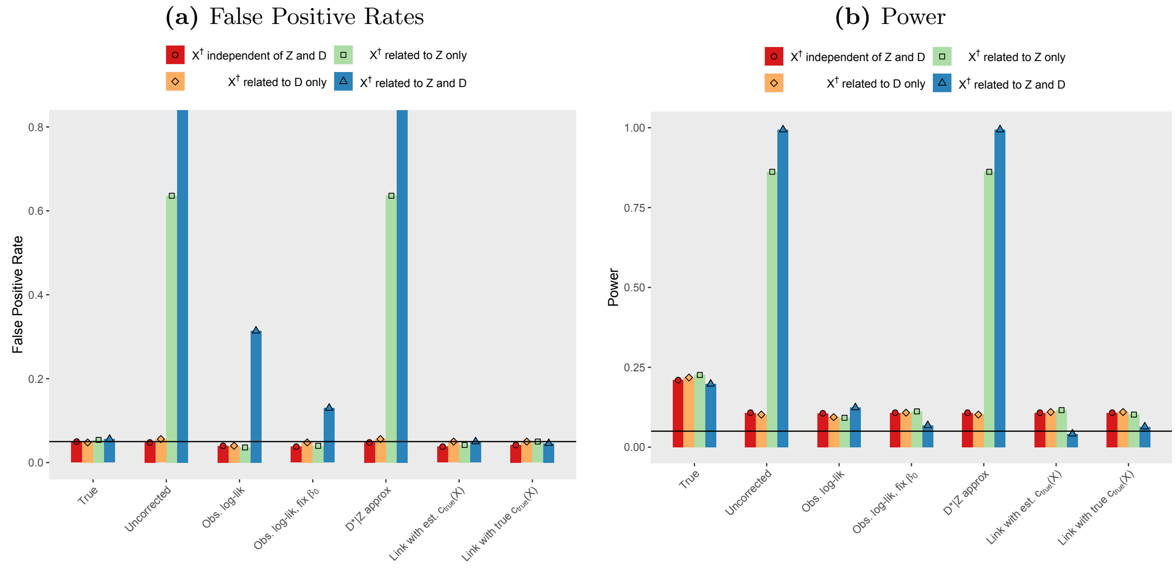
Figure B.2a shows the type I error rates. When X and Z are uncorrelated given D, we see nominal type I error rate across simulation settings considered, where the horizontal line in **Figure B.2a** corresponds to a type I error rate of 0.05. This is consistent with **Figure B.1**, which showed little difference in the resulting p-values. When X is related to Z given D but not related to D given Z, we see nominal type I error rates for analyses that correct for misclassification. That is, all methods except the uncorrected analysis and the method where we approximate the $D^*|Z$ distribution, which assumes constant sensitivity $c(Z) = \widetilde{c}$.

In the setting where X is independently related to both Z and D, the observed data log-likelihood maximization method and the $D^*|Z$ distribution approximation methods perform poorly. This is because we are violating the assumptions required for these methods. For the non-logistic link function method, we only require this independence when we are estimating $c_{true}(X)$ and using it to replace $c(Z)$. In these simulations, we see that use of both estimated and true $c_{true}(X)$ results in nominal type I error rates even when these assumptions are violated.

Figure B.2a also emphasizes that we cannot ignore misclassification related to covariates when X is related to Z given D (so $c(Z)$ is not a constant). This is particularly important because, unlike other types of misclassification, **we can have bias toward or away from the null** when sensitivity depends on covariates independently related to Z.

Figure B.2b shows the power when $\theta_Z = 0.05$. Note that this is a small value for θ_Z. We chose a small value to allow for imperfect power and easier comparison across methods. In all settings where bias is corrected appropriately, power is fairly low but generally still above the 0.05 level as expected.

Figure B.2: Estimated false positive rates and power across 500 simulations after imposing phenotype misclassification



(a) False Positive Rates

(b) Power

B.3 Simulation part 1: sensitivity estimators and misspecification of $P(D = 1|X)$

In the main paper, we propose several methods for estimating either marginal sensitivity or individual-level sensitivity using the observed data and some additional information about the population of interest. In this section, we will refer to these methods as follows:

Method 1: (**Section 3.1**) Estimate "crude" marginal sensitivity as $\widetilde{c} = \frac{P(D^*=1)}{P(D=1)}$
Method 2a: (**Section 3.2**) Estimate $c_{true}(X)$ using non-logistic link function method assuming $P(D = 1|X)$ is known. We fit the following model for $D^*|X$:

$$\log\left[\frac{P(D^* = 1|X)}{P(D = 1|X) - P(D^* = 1|X)}\right] = \beta_0 + \beta_X X = \text{logit}(c_{true}(X))$$

Method 2b: (**Section 3.2**) Estimate $c_{true}(X)$ using the following ratio

$$c_{true}(X) = \min\left(\frac{P(D^* = 1|X)}{P(D = 1|X)}, 1\right)$$

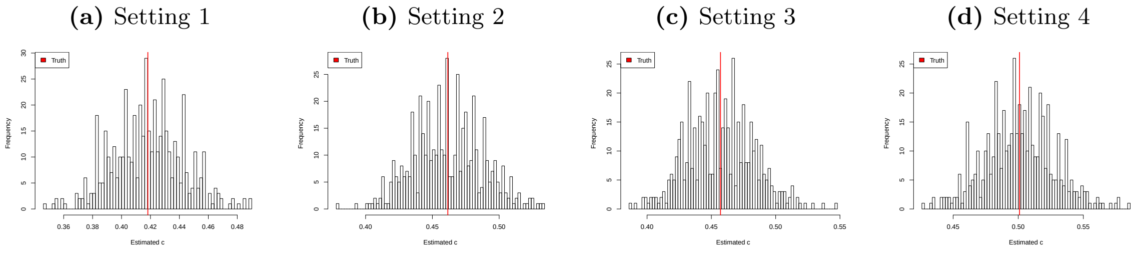
Method 3a: (**Section 3.3**) Estimate $c_{true}(X)$ through joint estimation of β and θ.
Method 3b: (**Section 3.3**) Estimation as in Method 3a but with β_0 fixed at $\text{logit}(\widetilde{c})$.

We will evaluate our ability to estimate $\widetilde{c}$ and $c_{true}(X)$ in several settings. We consider four general scenarios corresponding to different relationships between D, X, and Z. In particular, we consider simulation Settings 1-4 in **Table B.1.** In Settings 1 and 3, we have that Z is independent of X given D, and we have conditional dependence in Settings 2 and 4. Additionally, we have X and D associated given Z in Settings 3 and 4. We note that Settings 1 and 3 correspond to settings where $c(Z) = \widetilde{c}$ even though sensitivity $c_{true}(X)$ depends on covariates. In all simulation settings considered, the average sensitivity is roughly 0.4-0.5.

B.3.1 Estimating Marginal Sensitivity

In **Figure B.3**, we plot a histogram of the estimated $\widetilde{c}$ using Method 1 across 500 simulations. In all settings, these estimates are well-centered around the true marginal sensitivity (vertical line). We emphasize that we can do a good job in estimating the marginal sensitivity $\widetilde{c}$ even when sensitivity $c_{true}(X)$ does depend on covariates.

Figure B.3: Estimated $\widetilde{c}$ across 500 simulations using Method 1 *

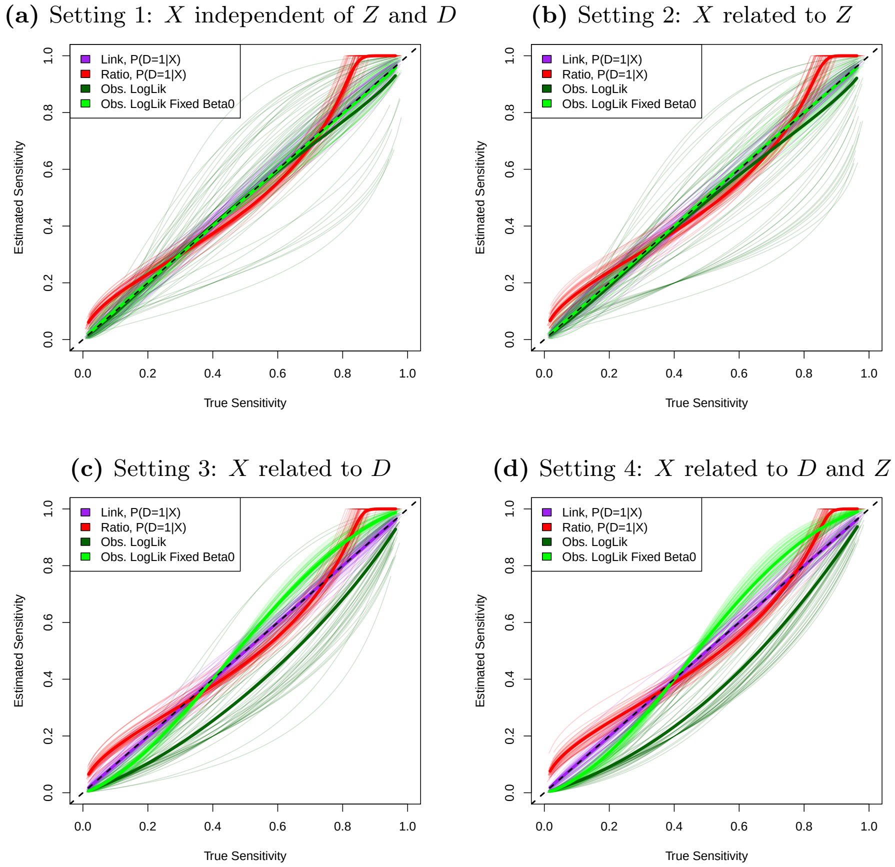

(a) Setting 1 **(b)** Setting 2 **(c)** Setting 3 **(d)** Setting 4

* The vertical line in each figure represents the simulation truth value for $\widetilde{c}$. We assume the population $P(D = 1)$ is known.

B.3.2 Estimating $c_{true}(X)$

Next, we evaluate our ability to estimate $c_{true}(X)$ using Methods 2a, 2b, 3a, and 3b. In **Figure B.4**, we compare $c_{true}(X)$ estimates and the truth for 50 simulated datasets, where $c_{true}(X)$ is estimated using different methods. For Methods 2a and 2b, we assume that $P(D = 1|X)$ is known. For Method 3b, β_0 was fixed at the truth. Method 2a performs very well across all four simulation settings. In all settings, Method 2b does not quite capture the true $c_{true}(X)$ at the upper end, but the general magnitude and trend for $c_{true}(X)$ are close to the truth. Unlike Methods 2a and 2b, Method 3 does not incorporate any outside information about $D|X$. Methods 3a and 3b both perform well when X is not independently related to D given Z. Both methods struggle more to estimate $c_{true}(X)$ when X is independently related to D given Z, and Method 3b (where we fix the intercept in the sensitivity model) has better performance.

Figure B.4: Estimated $c_{true}(X)$ for 50 simulated datasets using Methods 2 and 3*


(a) Setting 1: X independent of Z and D

(b) Setting 2: X related to Z

(c) Setting 3: X related to D

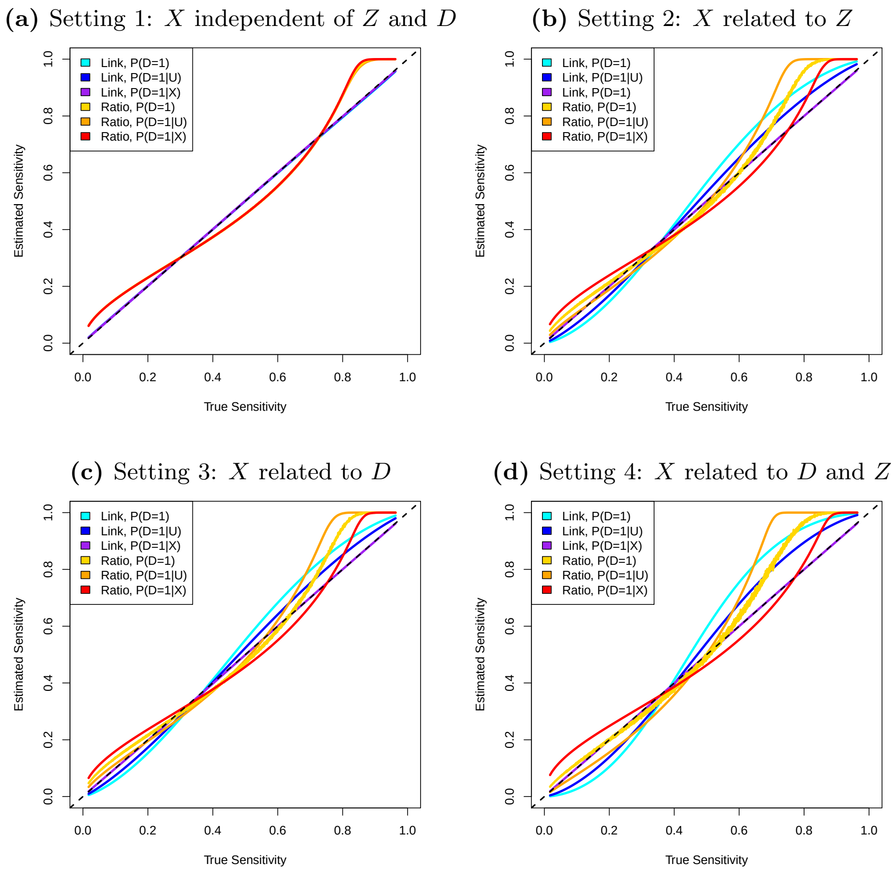
(d) Setting 4: X related to D and Z

* Estimated sensitivities for 50 individual datasets are plotted using thin lines. Bolded lines correspond to the average estimated sensitivity, sorted according to the corresponding average true sensitivity value across 50 simulated datasets.

Method 2 sensitivity estimation relies on known $P(D = 1|X)$, but this may not often be known in practice. For example, suppose X includes the length of follow-up in the EHR. We may rarely know the relationship between length of follow-up in *this* EHR and *true* disease status D. Instead, we may approximate $P(D = 1|X)$ using available $P(D = 1)$ or $P(D = 1|U)$ for some U related to X or a subset of X. In **Figure B.5**, we explore how well we can estimate $c_{true}(X)$ using Method 2 when $P(D = 1)$, $P(D = 1|U)$, or true $P(D = 1|X)$ is known. Here, we generate $U = X + e$ where $e \sim Normal(0, 1)$. When X is independent of Z and D (Setting 1), $P(D = 1|X) = P(D = 1)$, and all approaches perform well. In all other settings, use of

$P(D=1)$ or $P(D=1|U)$ results in some error in estimating $c_{true}(X)$. This is a particular concern when X is related to both D and Z (Setting 4). We note that, while there is still some error in estimating $c_{true}(X)$ using $P(D=1|U)$ in Settings 2-4, we see greater error when we just use $P(D=1)$. This demonstrates a benefit to incorporating what information that is available to best approximate $P(D=1|X)$ if the goal is to estimate $c_{true}(X)$.

Figure B.5: Estimated $c_{true}(X)$ using Method 2 when $P(D=1|X)$ not known*



(a) Setting 1: X independent of Z and D

(b) Setting 2: X related to Z

(c) Setting 3: X related to D

(d) Setting 4: X related to D and Z

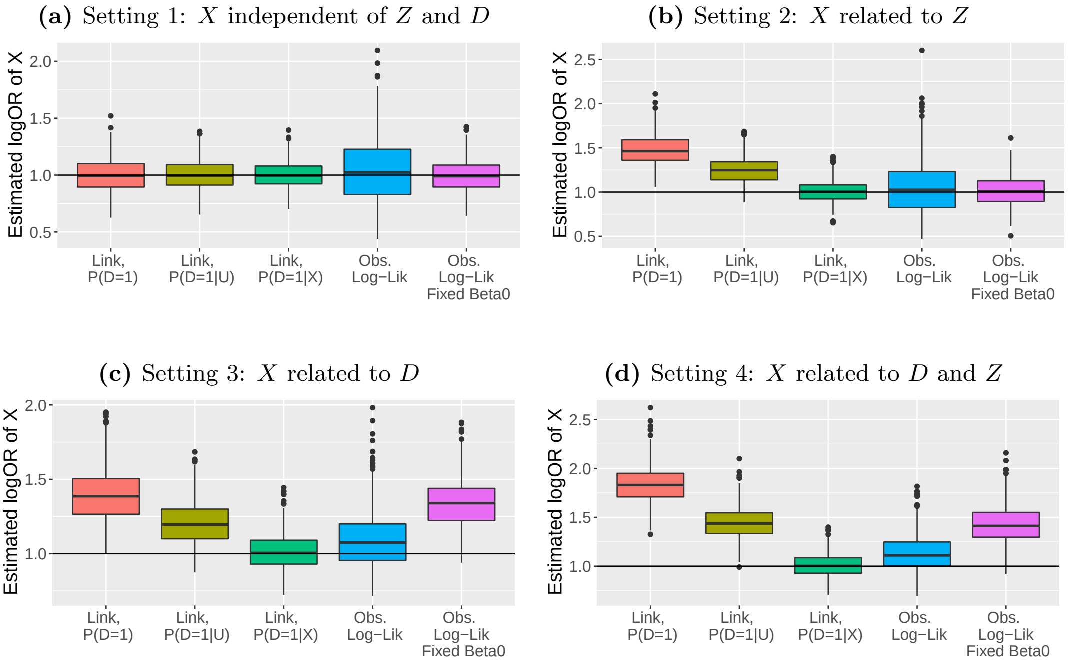
* Lines correspond to the average estimated sensitivity, sorted according to the corresponding average true sensitivity value across 50 simulated datasets.

B.3.3 Estimating β_X

We might also be interested in estimating β_X rather than $c_{true}(X)$. In **Figure B.6**, we show the estimated values for β_X for several methods and across 500 simulations. When X is independent of both Z and D, all methods perform well, with estimates of β_X well-centered around the truth of 1. We notice that the observed data log-likelihood method with no fixed parameters results in greater spread in estimated β_X compared to the other methods. This is due to the more difficult task of jointly estimating β and θ, resulting in less efficient estimates with greater variability.

When X is related to Z given D but is not related to D given Z (Setting 2), the observed log-likelihood methods perform well for estimating β, but some bias can be seen in estimating β with misspecified $P(D = 1|X)$. This bias is smaller when we specify $P(D = 1|U)$ rather than just $P(D = 1)$. When X is related to D given Z (Settings 3 and 4), all methods struggle somewhat in estimating β_X unless true $P(D = 1|X)$ is known. However, it should be noted that all methods produce estimated β_X of generally similar magnitude and correct direction. If the goal is to estimate directions of association in the sensitivity model, we may be less concerned about our specification of $P(D = 1|X)$. Interestingly, we also note that Method 3a out-performs Method 3b (with β_0 fixed at the truth) in terms of estimated θ_Z in Settings 3 and 4. This may indicate greater adaptability of Method 3a (no fixed intercept) compared to 3b (fixed intercept) when $P(D = 1|Z, X) \neq P(D = 1|Z)$ in terms of estimating $c_{true}(X)$.

Figure B.6: Estimated β_X across 500 simulations (true $\beta_X = 1$)



(a) Setting 1: X independent of Z and D

(b) Setting 2: X related to Z

(c) Setting 3: X related to D

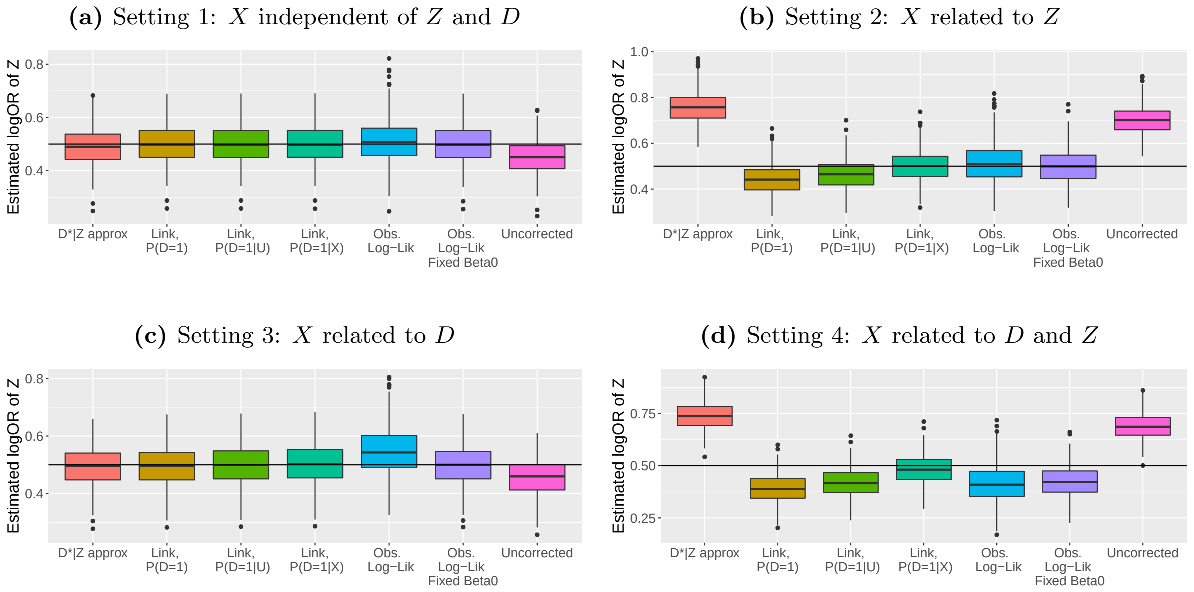
(d) Setting 4: X related to D and Z

B.3.4 Impact of sensitivity estimation on estimated θ

Now, suppose we are interested in θ_Z, and estimation of sensitivity is more of a means to an end. We want to understand to what extent difficulty in estimating sensitivity impacts estimation of θ. In **Figure B.7**, we provide boxplots of the estimated θ_Z across 500 simulated datasets with sensitivity estimated in different ways. Observed log-likelihood maximization without a fixed intercept performs well in estimating θ_Z when X is independent of D given Z but poorly otherwise. When we fix the intercept β_0 at or near the truth, we do a good job estimating θ_Z when X is independently related to Z given D, but neither observed log-likelihood maximization strategy performs well when X is independently related to both D and Z.

In Settings 1 and 3, we have that $c(Z) = \widetilde{c}$. In these settings, the method in **Section ??** in which we approximate the $D^*|Z$ distribution and use estimated $\widetilde{c}$ performs well. This method performs poorly in Settings 2 and 4, where $c(Z)$ is not truly a constant. Similarly, the non-logistic link function method performs well in both Setting 1 and 3. This is notable, since we are replacing constant $c(Z) = \widetilde{c}$ with $c_{true}(X)$ in these settings. This good performance comes from results in **Section A.4**, which indicate we can replace $c(Z)$ with $c_{true}(X)$ for estimation with the non-logistic link function method in Settings 1-3. In Settings 2 and 4, estimation using a non-logistic link function from **Section 3.2**) performs well when $P(D = 1|X)$ is correctly specified. We can see some error in estimating θ_Z when $P(D = 1|X)$ is not correctly specified. Importantly, all methods in **Section 3.2 and 3.3** out-perform estimation assuming constant sensitivity **Section 3.1** when X is independently related to Z (even when $P(D = 1|X)$ is misspecified). This suggests that accounting for covariate relationships with sensitivity using methods in **Section 3.2** may be a good idea even when $P(D = 1|X)$ is not well-known.

Figure B.7: Estimated θ_Z across 500 simulations (true $\theta_Z = 0.5$)



(a) Setting 1: X independent of Z and D

(b) Setting 2: X related to Z

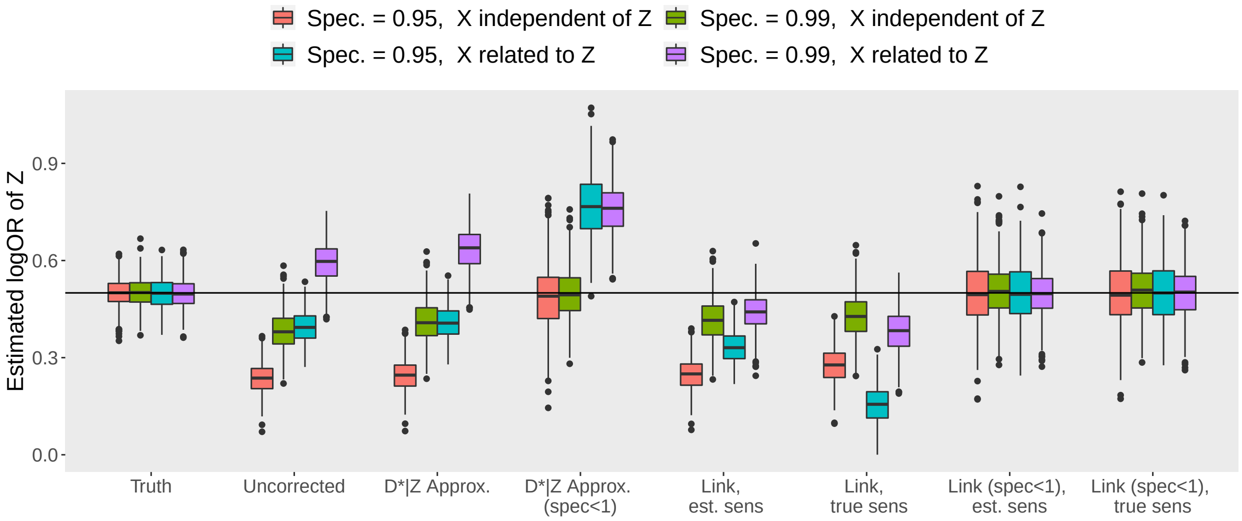
(c) Setting 3: X related to D

(d) Setting 4: X related to D and Z

B.4 Simulation part 1: estimation under imperfect specificity

In other simulations, we assume we have perfect specificity, but this may not always be the case. We consider the setting where specificity is a known constant less than 1. We generate data as in Settings 1 and 2 in **Table B.1** under imperfect specificity with $\widetilde{b}$ equal to 0.99 or 0.90. This results in 4 simulation settings. We then estimate θ_Z using uncorrected analysis, after applying the methods in **Sections 3.1 and 3.2** assuming perfect specificity, and after applying the methods in **Sections 3.1 and 3.2** assuming known $\widetilde{b} < 1$. Estimated θ_Z values across each of 500 simulated datasets for each simulation setting and method combination are shown in **Figure B.8**. We first note that uncorrected analysis results in bias in all simulation settings. Application of methods that correct for imperfect sensitivity but incorrectly assume imperfect specificity do not correct this bias and can sometimes make the bias worse. When we apply our methods that correctly account for imperfect specificity, we can do a good job at estimating θ_Z.

Figure B.8: Estimated θ_Z across 500 simulations under imperfect specificity*



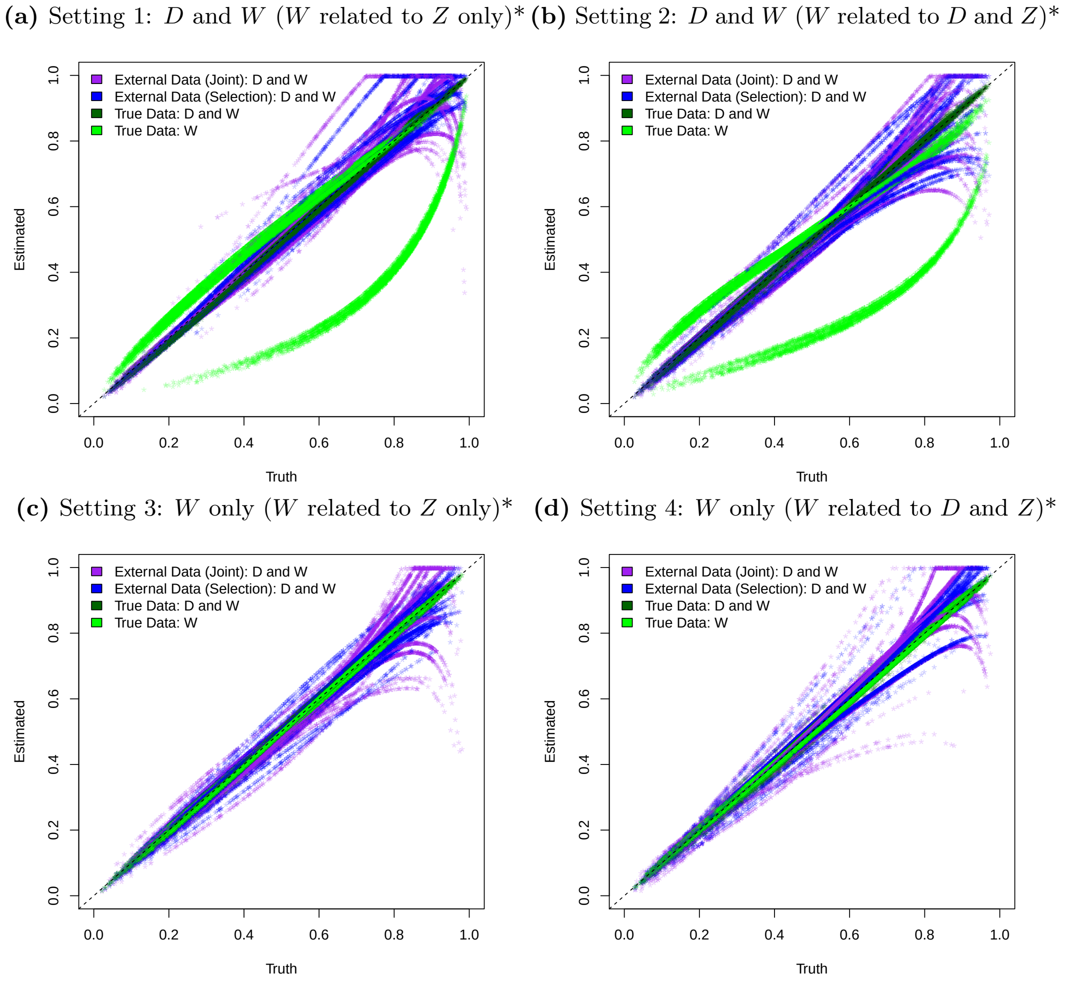
* Spec. = specificity. For the non-logistic link function method, we compare results using estimated ("est.") and true $c_{true}(X)$.

B.5 Simulation part 2: estimating selection probabilities using external data

In each scenario in Settings 1-4 of Simulation part 2 in **Table B.1**, we generate a probability sample from the population with a 50% selection probability. We use this external probability sample combined with the internal data to estimate sampling probabilities using either *Eq. S4* ("joint" method: relying on the joint distribution of D and W) or *Eq. S5* ("selection" method: relying on a multinomial selection model for inclusion in the non-probability sample, the probability sample, or both). We assume both D and W are available for all people.

Figure B.9 compares the estimated and true selection probabilities for 10 simulated datasets. We also plot the estimated selection probabilities obtained using data from the "true" (entire) population and consider modeling based on (1) D and W and (2) W only. In all simulation settings, the estimators in *Eq. S5* and *Eq. S4* do a good job at recovering the true selection probabilities. Unsurprisingly, estimated sampling probabilities obtained using these methods have greater variability across simulated datasets than when sampling probabilities are estimated using the entire population. When selection depends directly on D in addition to W (Settings 1 and 2), failure to include D in the selection model resulted in poor estimation of the selection probabilities. These simulations demonstrate that the methods in *Eq. S4* and *Eq. S5* can do a good job at estimating the selection probabilities using an external probability sample as long as the various selection models are correctly specified.

Figure B.9: Estimated sampling probabilities for 10 simulated datasets



(a) Setting 1: D and W (W related to Z only)* **(b)** Setting 2: D and W (W related to D and Z)*

(c) Setting 3: W only (W related to Z only)* **(d)** Setting 4: W only (W related to D and Z)*

*Variables included in selection model (dependence structure). "Joint" corresponds to estimation using an external probability sample and *Eq. S4*. "Selection" corresponds to estimation using an external probability sample and *Eq. S5*.

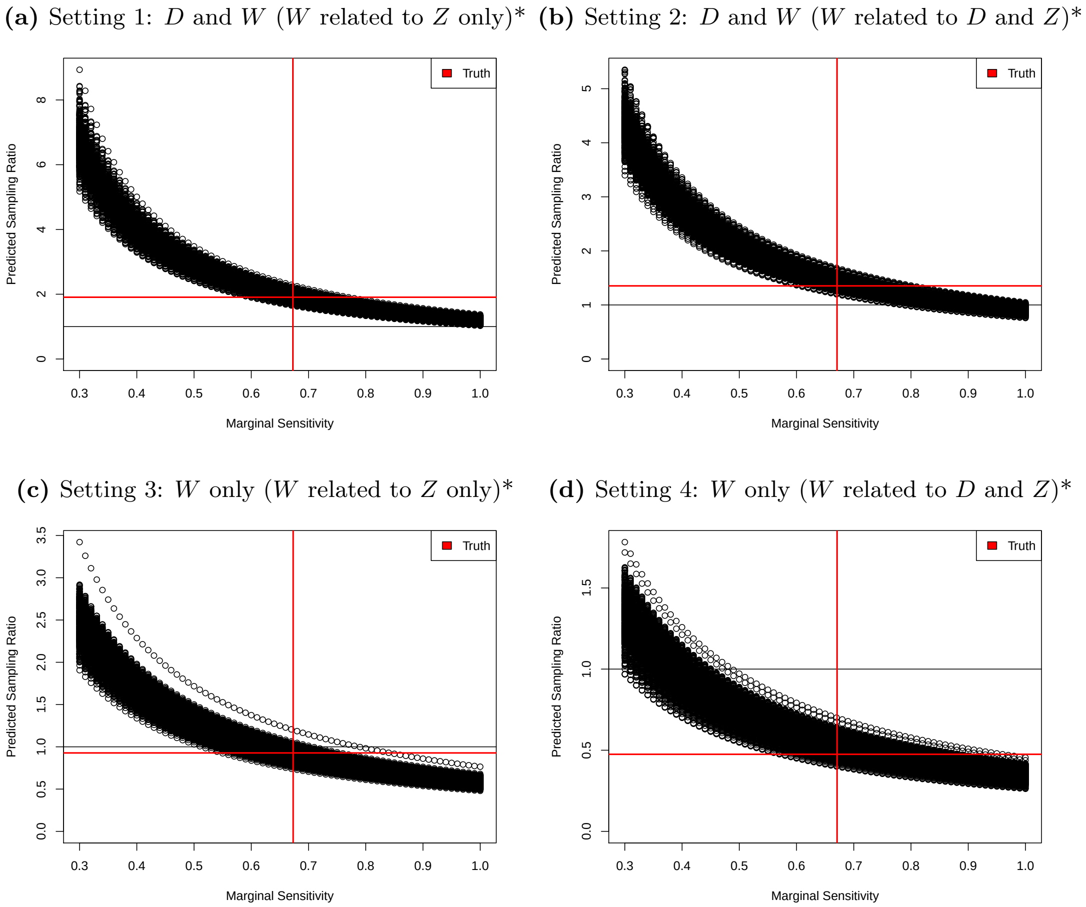
B.6 Simulation part 3: exploring plausible values for sampling ratio

In *Eq. 8*, we express $\widetilde{r}$ as a function of $\widetilde{c}$ and $P(D=1)$ as follows:

$$\widetilde{r} = \frac{P(D^*=1|S=1)}{\widetilde{c}-P(D^*=1|S=1)}\frac{1-P(D=1)}{P(D=1)}$$

We can use this relationship to plot a curve relating $\widetilde{r}$ and $\widetilde{c}$ as shown for 500 simulations under Simulation part 3 Settings 1-4 in **Figure B.10**. True values for $\widetilde{c}$ and $\widetilde{r}$ are plotted as red lines, and we can see that the predicted curves intersect the true values. Using this plot, we can estimate either $\widetilde{r}$ or $\widetilde{c}$ by fixing a value for the other. Alternatively, we can use these plots to inform plausible values of $\widetilde{r}$ based on our beliefs about $\widetilde{c}$ and repeat our analysis for several plausible values of $\widetilde{r}$.

Figure B.10: Estimated relationship between $\widetilde{r}$ and $\widetilde{c}$ for 500 simulated datasets



(a) Setting 1: D and W (W related to Z only)*

(b) Setting 2: D and W (W related to D and Z)*

(c) Setting 3: W only (W related to Z only)*

(d) Setting 4: W only (W related to D and Z)*

*Variables included in selection model (dependence structure).

B.7 Simulation part 3: estimating sensitivity as a function of the sampling ratio

In **Section 5**, we describe several strategies for estimating $c_{true}(X)$. Here, we will explore how these approaches perform for different fixed $\widetilde{r}$ in several simulation settings. Strategies include use of the observed data log-likelihood maximization method, which can be applied with or without fixing $\beta_0 = \text{logit}(\widetilde{c})$. For this exploration, we apply the observed data log-likelihood method to estimate $c_{true}(X)$ with and without IPW weighting to adjust for selection bias. Fixing $\widetilde{r}$ and assuming $P(D = 1|X)$ is known, we can also estimate $c_{true}(X)$ using two strategies. In the first "Link" method, we fit the non-logistic link function model in **Section 5** to estimate β and, therefore, $c_{true}(X)$. In the second "Ratio" method, we estimate $c_{true}(X)$ as the ratio of $P(D^* = 1|X, S = 1)$ and $P(D = 1|X, S = 1)$.

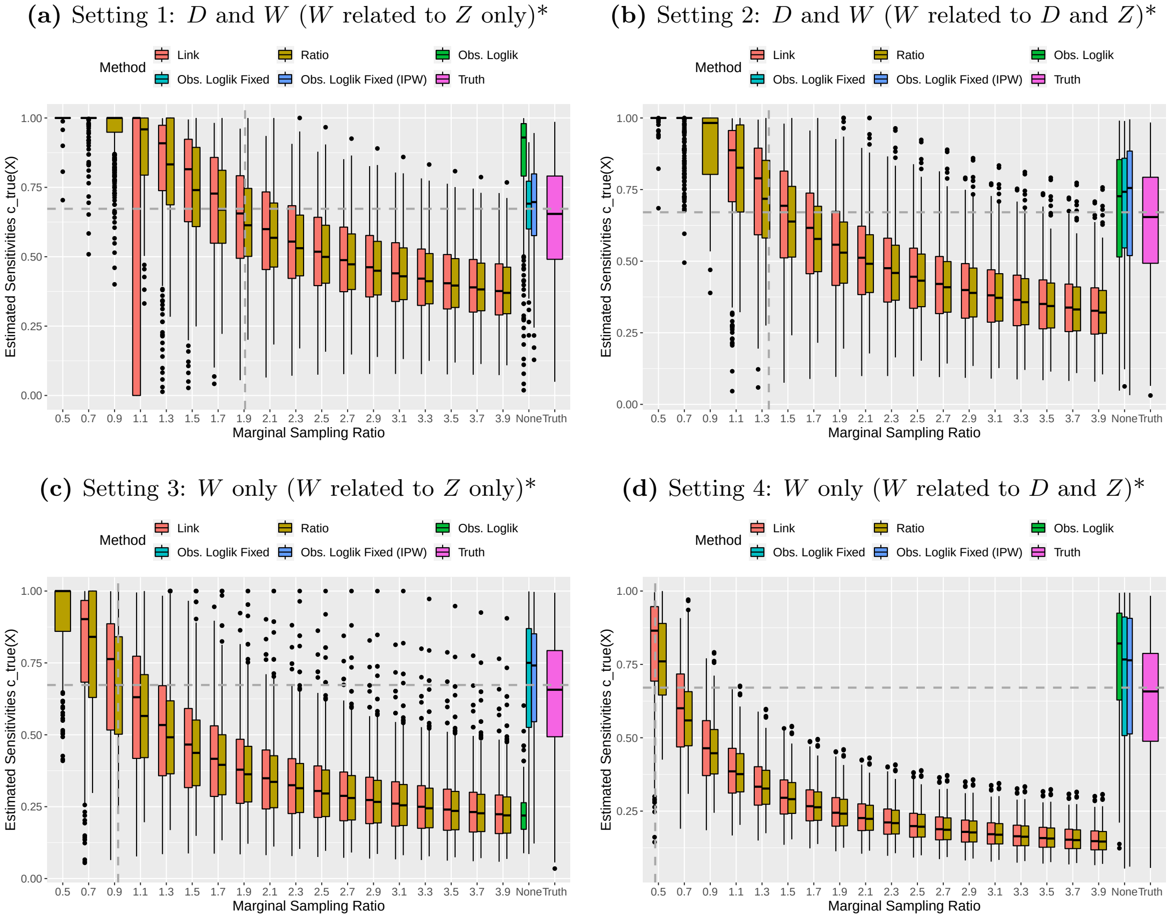
Figure B.11 provides boxplots of the estimated individual-level sensitivities in a single simulated dataset. When we specify the correct marginal sampling ratio and $P(D = 1|X)$, we see that we can do a good job at estimating $c_{true}(X)$. However, estimation of $c_{true}(X)$ using the non-logistic link function or ratio methods may be somewhat sensitive to the choice of $\widetilde{r}$. Alternatively, suppose we specify β_0 instead of $\widetilde{r}$. We can apply the observed data log-likelihood maximization methods. Whether or not we also account for selection, we tend to do a good job at estimating sensitivities when we fix β_0, suggesting that selection bias may play a larger role in estimation of θ than β_X if β_0 is roughly known. In contrast, observed data log-likelihood estimation without a fixed β_0 value performs poorly at estimating $c_{true}(X)$ (i.e. has a hard time estimating θ, β_X, and β_0 at once).

Rather than $c_{true}(X)$, we might also be interested in estimating β_X. **Figure B.12** provides boxplots of the estimated values of β_X across 50 simulated datasets using a variety of estimation strategies. Using the non-logistic link function method, estimated β_X does appear to be somewhat sensitive to the choice of $\widetilde{r}$. However, all β_X estimates are in the correct direction from zero, so we may be less worried about the exact value of $\widetilde{r}$ if we want to get a general sense of the important drivers of sensitivity. When $\widetilde{r}$ is correctly specified, the β_X estimate is near the truth. The observed data log-likelihood method with fixed β_X does a reasonable job at estimating β_0 with or without correcting for selection bias in Settings 1-3.

Key Takeaway: Generally, we need to fix either β_0 or $\widetilde{r}$ in order to do a good job estimating β_X and $c_{true}(X)$.

It appears that $c_{true}(X)$ and β_X are at least moderately impacted by the specification of $\widetilde{r}$ and/or β_X. However, we are more interested in the downstream implications for estimated θ_Z. **Figure B.13** provides boxplots of estimated θ_Z across 50 simulated datasets across values for $\widetilde{r}$ (assuming θ_Z was estimable using the non-logistic link function method in **Section 5**). For example, in Setting 4, many values of $\widetilde{r}$ are not plotted since no solution existed for *Eq. 6* for any of the 50 simulated datasets. We see that θ_Z can be somewhat sensitive to the choice of $\widetilde{r}$, but the bias in mis-specifying $\widetilde{r}$ is often smaller than bias of uncorrected data analysis, particularly when selection is related to D.

Figure B.11: Estimated $c_{true}(X)$ across sampling ratios in single simulated dataset



(a) Setting 1: D and W (W related to Z only)*

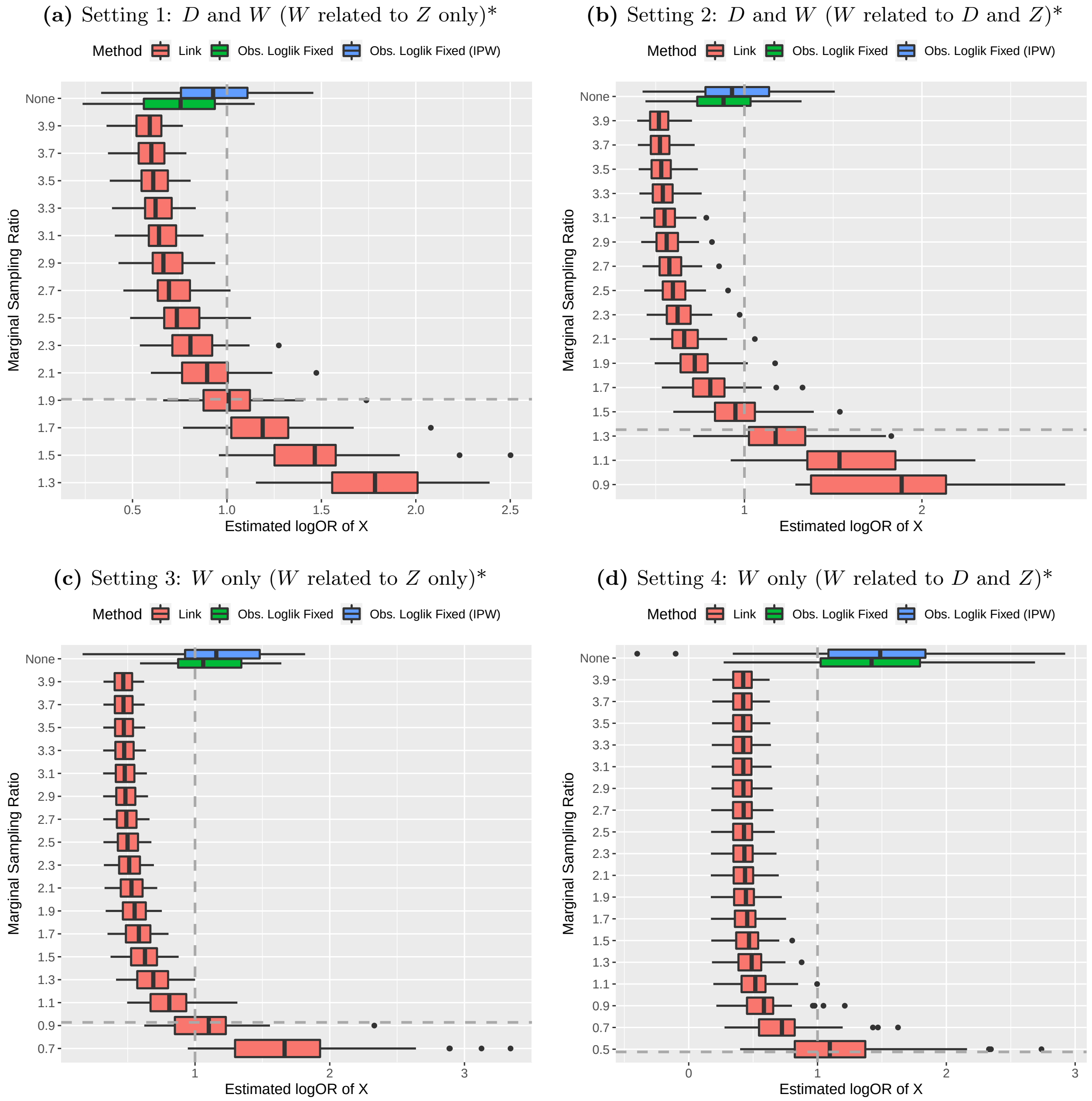
(b) Setting 2: D and W (W related to D and Z)*

(c) Setting 3: W only (W related to Z only)*

(d) Setting 4: W only (W related to D and Z)*

*Variables included in selection model (dependence structure). The horizontal and vertical lines correspond to the true values of $\widetilde{r}$ and $\widetilde{c}$ respectively. Boxplots were not plotted when sensitivity could not be measured using the non-logistic link function method.

Figure B.12: Estimation of β_X across sampling ratios in 50 simulated datasets



(a) Setting 1: D and W (W related to Z only)*

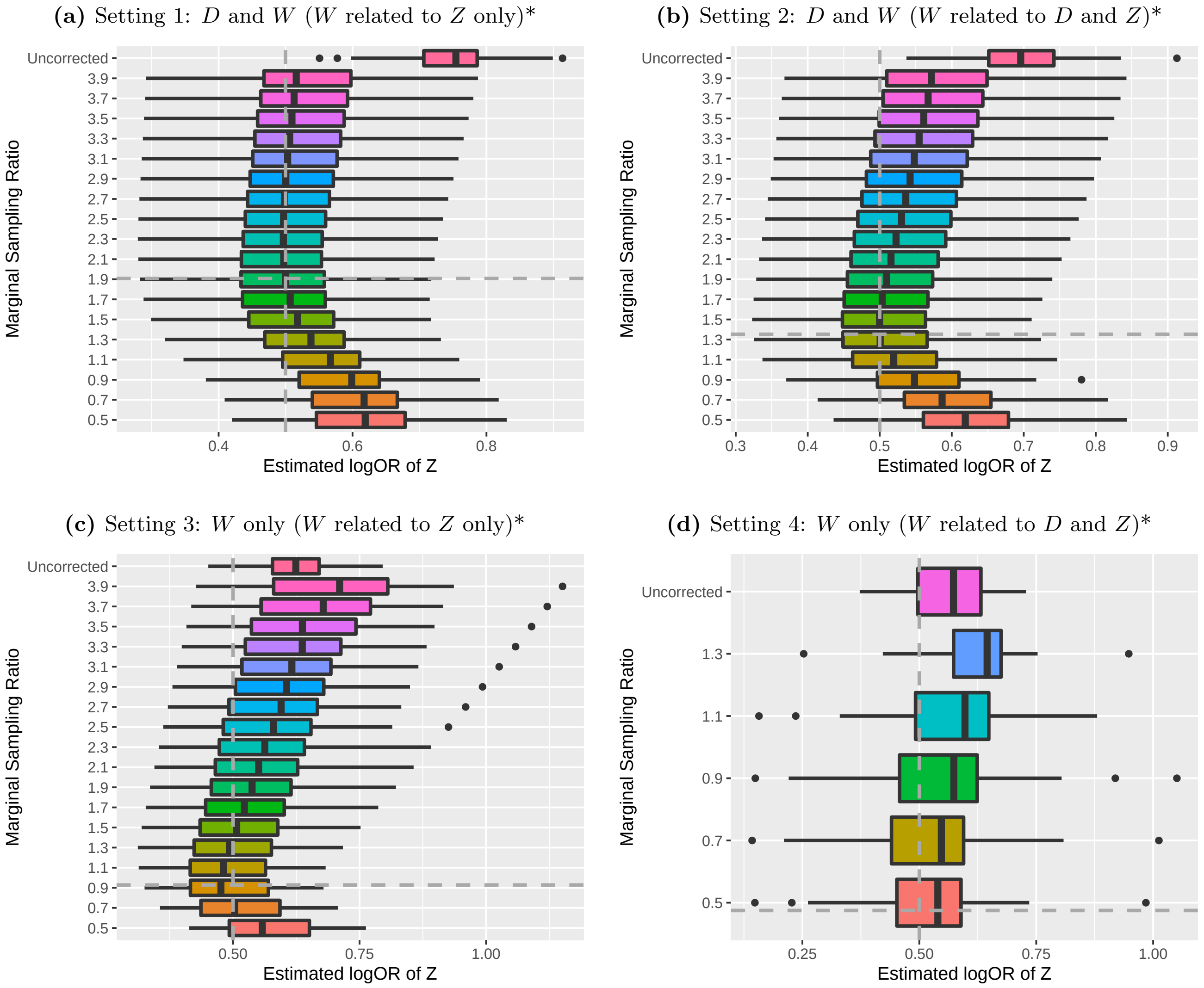
(b) Setting 2: D and W (W related to D and Z)*

(c) Setting 3: W only (W related to Z only)*

(d) Setting 4: W only (W related to D and Z)*

*Variables included in selection model (dependence structure). The horizontal and vertical lines correspond to the true values of β_X and $\widetilde{r}$ respectively. Sensitivities were estimated using the non-logistic link function method fixing $\widetilde{r}$ or by maximizing the observed data log-likelihood with fixed β_0.

Figure B.13: Estimation of θ_Z across sampling ratios in 50 simulated datasets (using correct selection weights)



(a) Setting 1: D and W (W related to Z only)*

(b) Setting 2: D and W (W related to D and Z)*

(c) Setting 3: W only (W related to Z only)*

(d) Setting 4: W only (W related to D and Z)*

*Variables included in selection model (dependence structure). The horizontal and vertical lines correspond to true $\widetilde{r}$ and θ_Z respectively.

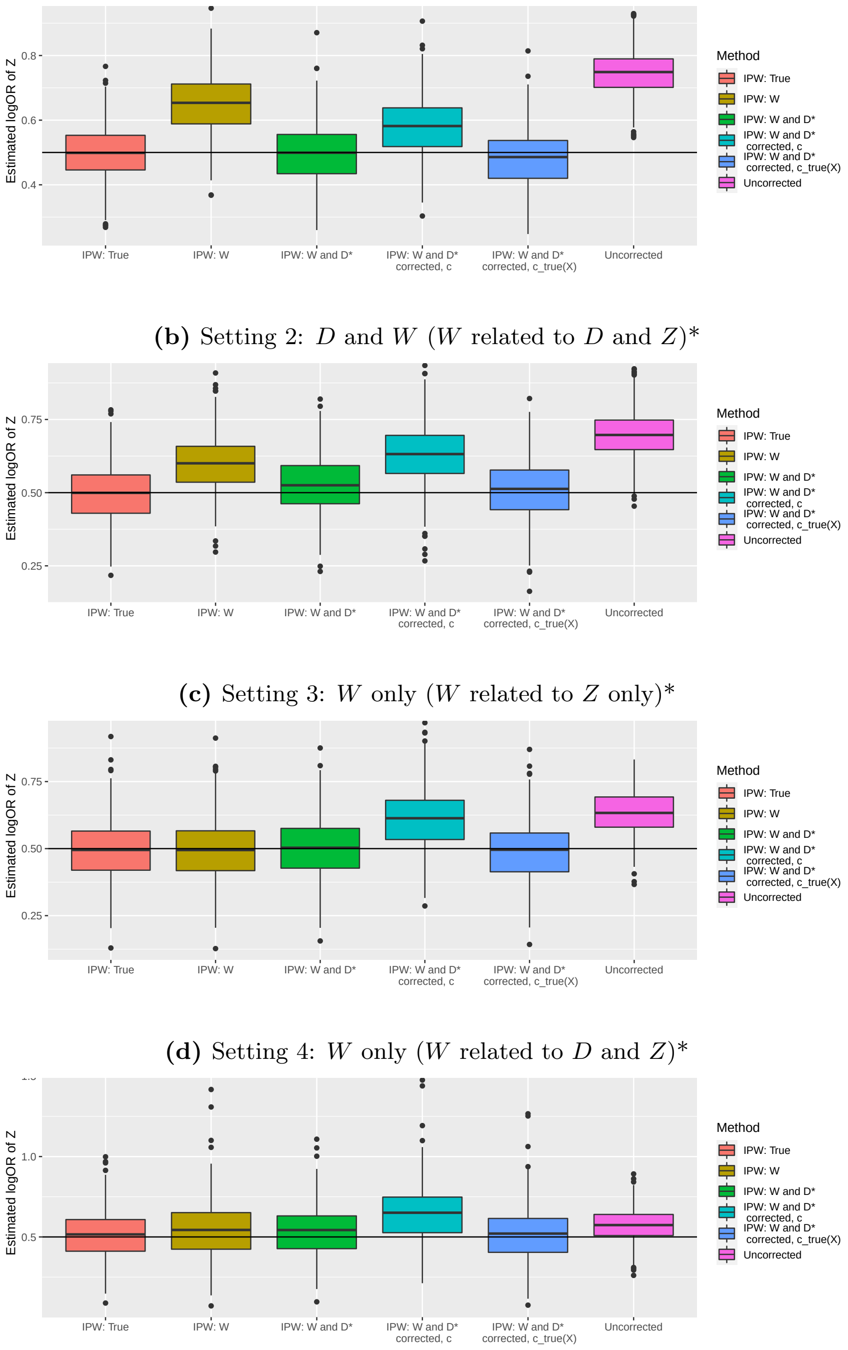
B.8 Simulation part 3: estimating selection probabilities given sensitivity

Now, we turn our attention to estimation of selection probabilities. Ideally, we would obtain an estimate of $P(S = 1|D, W)$. Beyond the challenges due to unobserved components of W, it is particularly challenging to estimate this probability when D is misclassified in the non-probability EHR sample. In **Section 5**, we propose instead attempting to address selection bias using $P(S = 1|W, D^*)$ instead of $P(S = 1|D, W)$.

In this section, we compare θ_Z using the non-logistic link function method using different estimated selection probabilities. For these explorations, we use the true values of $\widetilde{c}$ and $c_{true}(X)$ to estimate the selection probabilities and θ_Z where applicable. We suppose we generate an external probability sample from the population. We estimate selection probabilities using the following methods: (1) $P(S = 1|W)$ estimated using *Eq. S4*, (2) $P(S = 1|D, W)$ ignoring that disease status is misclassified using *Eq. S4*, (3) $P(S = 1|D^*, W)$ correcting for misclassification and using $\widetilde{c}$ and *Eq. 10*, and (4) $P(S = 1|D^* = 1, W)$ correcting for misclassification and using $c_{true}(X)$ and *Eq. 10*.

Figure B.14, we provide boxplots of estimated θ_Z across 500 simulated datasets, where weights for selection bias adjustment are estimated using various approaches. We find that θ_Z is poorly estimated when IPW weights ignore D or D^* and selection depends on D (Settings 1 and 2). In contrast, estimation of θ_Z with weights estimated ignoring the misclassification of D ("IPW: W and D^*") does a reasonable job at correcting for the selection bias, particularly when W is not independently related to D. The estimation strategy in *Eq. 10* performs poorly when we use marginal sensitivity $\widetilde{c}$ in place of $c_{true}(X)$. This occurs in simulation settings where X and Z are associated given D, but this is less of a problem when X and Z are conditionally independent (simulations not shown). We generally see good performance of the estimator in *Eq. 10* when we use true $c_{true}(X)$ to account for the misclassification.

Figure B.14: Estimated θ_Z using different IPW weights across 500 simulated datasets (using weighted non-logistic link function method and correct sensitivity)



(a) Setting 1: D and W (W related to Z only)*

(b) Setting 2: D and W (W related to D and Z)*

(c) Setting 3: W only (W related to Z only)*

(d) Setting 4: W only (W related to D and Z)*

*Variables included in selection model (dependence structure).“IPW: True” corresponds to IPW with the true selection model. “Uncorrected” analysis gives the estimated log-odds ratio ignoring selection and misclassification. “IPW: W” and “IPW: W and D^*” correspond to IPW weights $P(S=1|W)$ and $P(S=1|D,W)$ estimated using *Eq. S4* and ignoring misclassification. The other methods estimate $P(S=1|D^*,W)$ using *Eq. 10*.

C Implementation

C.1 R package *SAMBA*

Accompanying this paper, we have developed an R package called *SAMBA* (sampling and misclassification bias adjustment) for implementing the proposed methods. Methods implemented include estimation of $\widetilde{c}$ and $c_{true}(X)$ with and without selection bias adjustment and estimation of θ using the methods in **Section 3 and 5** in the main paper. We assume that IPW/calibration weights ω used for selection bias adjustment are estimated separately by the user, perhaps using the methods developed in this paper. Current implementation assumes perfect specificity, and future code developments will extend to the setting of imperfect specificity. We demonstrate how we can use SAMBA to perform the proposed analyses through the following pseudo-code:

Downloading R package:

```
devtools::install_github("umich-cphds/SAMBA",build_vignettes = TRUE, build_opts = c("--
    no-resave-data", "--no-manual"))
library(SAMBA)
```

Estimating $\widetilde{c}$ and $c_{true}(X)$:

```
estimated_sensitivity = sensitivity(X = sensitivity model predictors,
    Dstar = observed disease indicator,
    r = marginal sampling ratio if desired,
    prev = assumed relationship between disease and X)
```

Estimating θ:

```
### Approximation of D*|Z  (Sections 3.1 [unweighted] and 5.1 [weighted])
approx = approxdist(Z = disease model predictors,
    Dstar = observed disease indicator,
    weights = IPW or calibration weights if desired,
    c_marg = marginal sensitivity)

### Non-logistic link function method (Sections 3.2 [unweighted] and 5.2 [weighted])
nonlog = nonlogistic(Z, Dstar, weights,
    c_X = patient-specific sensitivity estimates)

### Observed data likelihood maximization (Sections 3.3 [unweighted] and 5.3 [weighted])
loglik = obsloglik(Z, X, Dstar,
    start = starting values for (theta, beta),
    beta0_fixed = fixed beta0 if desired,
    weights)
```

For more details about this package, we refer readers to the instructive vignette.

```
browseVignettes(‘SAMBA’)
```

C.2 Automating methods for large-scale association studies

In the main paper, we focus on the setting with a single disease D of interest and a single predictor set, Z. In modern EHR data analysis, we are often interested in studying many associations at once. Two common study designs are genome-wide association studies (GWAS), where we relate a single D to many different Z's, and phenome-wide association studies (PheWAS), where we relate many different diseases (many D's) to a single Z. Increasingly, researchers are also interested in studying associations across both the phenome and genome (many D's and Z's).

GWAS: For GWAS, we can adjust for phenotype misclassification and selection bias using a *single* set of sampling weights and sensitivity estimates, since the disease outcome is the same for each of the associations of interest. Given estimates of sensitivity and weights ω, we can then estimate θ_Z for each Z of interest using the methods discussed in this paper. We discuss three general methods: (1) approximation of the $D^*|Z$ distribution, (2) regression modeling with a non-logistic link function, and (3) joint estimation of sensitivity and disease model parameters. Given the large numbers of associations of interest and the comparative slowness of estimation, we do not recommend method (3) in the GWAS setting. The first two methods, however, can be easily implemented and scalable to a large number of association tests.

We first consider the setting where we are only doing adjustment for misclassification and not for selection bias. In this case, we are looking at the methods in **Section 3**. With sensitivity $\widetilde{c}$ or $c_{true}(X)$ estimated, both methods (1) and (2) are simple to implement. Method (1), in particular, will be very fast to implement genome-wide, since it involves a simple transformation of the uncorrected point estimates. Therefore, it does not require any models to be re-fit after the uncorrected analysis is performed. The main limitation of method (1) is that it requires strong assumptions about the sensitivity to hold. In particular, we require that $c(Z)$ can be reasonably approximated by constant $\widetilde{c}$, which occurs if X is independent of Z given D. This is a strong assumption, which may not always hold. When this assumption does hold, however, this method will result in corrected and uncorrected point estimates that differ but p-values that are the same. When the p-values are of sole interest, therefore, application of method (1) bias correction ignoring selection bias will have no impact on p-values. Method (2) can be applied in the more general setting where $X^\dagger$ is independent of Z given D. This allows adjustment factors in the disease model to be related to sensitivity. Method (2) p-values and point estimates will differ relative to uncorrected analysis. Compared to method (1), method (2) will be slower, but it will be on the order of standard logistic regression. Therefore, method (2) should be reasonably scalable to many association tests when sensitivity (and sampling weights if used) are already estimated.

Now, we consider the setting where we are doing adjustment for selection bias or misclassification *and* selection bias. Similar In this case, the uncorrected and corrected p-values will be different, and the point estimates will also be impacted. Either method (1) or method (2) can be implemented, and the comparison between methods is similar to the setting ignoring selection bias adjustment.

PheWAS: For PheWAS, a separate set of sensitivities and sampling weights are estimated for each association of interest. If we want to perform 2000 tests, for example, we will need to estimate sensitivity and sampling weights (if we adjust for both misclassification and selection) for 2000 different diseases.

Suppose first that sensitivities ($\widetilde{c}$ and $c_{true}(X)$) and weights ω have already been estimated for each association of interest in the PheWAS. Methods (1) and (2) above can then be applied across all associations as in the GWAS setting described previously. Method (3) may be more feasible to implement for thousands of parallel tests in a PheWAS rather than millions in a GWAS, but estimation will be slower than for the other two methods. Therefore, the results on scalability described for GWAS above apply here.

The primary challenge for applying the proposed methods for PheWAS is in estimating sensitivity and sampling weights, which will differ for each association test. Sampling weights, in particular, are challenging to specify even when we have a single association of interest, and scaling this estimation phenome-wide would be very difficult. Currently, our proposed methods will be very difficult to apply phenome-wide when both misclassification and selection are being accounted for when sampling weights are not known. Instead, we will focus on the setting where we **assume selection is ignorable** and want to estimate θ and sensitivities as in **Section 3**.

Firstly, we can estimate sensitivity jointly with θ through maximizing the observed data log-likelihood as in method (3) above, and we will not need to separately estimate sensitivity and can just implement method (3) for each association of interest. Two other strategies were proposed for estimating sensitivity are as follows: (a) $\widetilde{c} = \frac{P(D^*=1)}{P(D=1)}$ and (b) estimation of $c_{true}(X)$ using *Eq. 6* and given $P(D = 1|X)$.

The primary challenge for automating (a) is that it requires us to known the population marginal disease rate for all diseases of interest. These rates may be easy to obtain for many common diseases (e.g. cancer statistics from SEER or recent statistics from NHANES), but it may be difficult to obtain $P(D = 1)$ for *all* diseases of interest in the phenome. Suppose we focus our attention to diseases for which the population disease rates are known. In this case, $\widetilde{c}$ can be easily estimated for all associations of interest and applied to estimate θ_Z using method (1).

Additionally, suppose we have gold standard known θ_Z for some D and Z. We can use the expression in *Eq. 5* and an estimated association using our misclassified EHR-derived D^* to back out a reasonable value for $\widetilde{c}$ for that disease as follows:

$$\theta_{Z,goldstandard} \approx \theta_Z^{uc}\left[\frac{\widetilde{c}(1-P(D^*=1))}{\widetilde{c}-P(D^*=1)}\right] \implies \widetilde{c} = \frac{\theta_{Z,goldstandard}P(D^*=1)}{\theta_{Z,goldstandard}-\theta_Z^{uc}P(D^*=0)}$$

If we have such gold standard information (e.g. associations with gender) for many diseases, we can use this information to estimate $\widetilde{c}$ for many diseases of interest. One example source for such gold standard associations might be the NHGRI GWAS Catalog, which compiles estimated associations between diseases and genotype information across a broad spectrum of diseases. If we can duplicate those associations for diseases of interest in our EHR dataset, we can use that information to estimate $\widetilde{c}$ for each disease.

Suppose instead that we want to estimate $c_{true}(X)$ and apply method (2). Estimation of $c_{true}(X)$ requires $P(D = 1|X)$, which can be very difficult to specify for a large number of diseases. In our data analyses in MGI, for example, we obtained an estimate for cancer using SEER statistics. This method, therefore, may be difficult to implement phenome-wide at this time.

References

Lauren J Beesley, Lars G Fritsche, and Bhramar Mukherjee. An analytic framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. *Statistics in Medicine*, 39(14):1965–1979, 2020.

Raymond J Carroll. *Measurement Error in Nonlinear Models: A Modern Perspective.* CRC Press, 2006.

Victor Castro, Yuanyuan Shen, Sheng Yu, Sean Finan, Cindy Ta Pau, Vivian Gainer, Candace C. Keefe, Guergana Savova, Shawn N. Murphy, Tianxi Cai, and Corrine K. Welt. Identification of subjects with polycystic ovary syndrome using electronic health records. *Reproductive Biology and Endocrinology*, 13(116):1–8, 2015.

Aba Diop, Aliou Diop, and Jean-François Dupuy. Maximum likelihood estimation in the logistic regression model with a cure fraction. *Electronic Journal of Statistics*, 5(0):460–483, 2011.

S. W. Duffy, J. Warwick, A. R.W. Williams, H. Keshavarz, F. Kaffashian, T. E. Rohan, F. Nili, and A. Sadeghi-Hassanabadi. A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology and Community Health*, 58(8):712–717, 2004.

Michael Elashoff. An EM Algorithm for Estimating Equations. *Journal of Computational and Graphical Statistics*, 13(1):48–65, 2004.

Michael R Elliot. Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights. *Survey Practice*, 2(3):1–7, 2009.

David A Freedman. On The So-Called “Huber Sandwich Estimator” and “Robust Standard Errors”. *The American Statistician*, 60(4):299–302, 2006.

Sebastien Haneuse and Michael Daniels. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why? *eGEMs*, 4(1):1–17, 2016.

Jing Huang, Rui Duan, Rebecca A. Hubbard, Yonghui Wu, Jason H. Moore, Hua Xu, and Yong Chen. PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*, 25(3):345–352, 2018.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 1989.

Tzon-Tzer Lu and Sheng-Hua Shiou. Inverses of 2 by 2 Block Matrices. *Computers and Mathematics with Applications*, 43(1):119–129, 2002.

John M Neuhaus. Bias and Efficiency Loss Due to Misclassified Responses in Binary Regression. *Biometrika*, 86(4):843–855, 1999.

John M Neuhaus and Nicholas P Jewell. A Geometric Approach to Assess Bias Due to Omitted Covariates in Generalized Linear Models Author. *Biometrika*, 80(4):807–815, 1993.

Ori Rosen. Mixture of Marginal Models. *Biometrika*, 87(2):391–404, 2000.

Jennifer A. Sinnott, Wei Dai, Katherine P. Liao, Stanley Y. Shaw, Ashwin N. Ananthakrishnan, Vivian S. Gainer, Elizabeth W. Karlson, Susanne Churchill, Peter Szolovits, Shawn Murphy, Isaac Kohane, Robert Plenge, and Tianxi Cai. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics*, 133(11):1369–1382, 2014.