Web Appendix to "Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification"

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 $P(D^* = 1|S = 1, D = 0, Z) = 0$. We have

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

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

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

or equivalently,

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

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 and $r(Z)$ reflects contributions of the sampling mechanism. Notably, if we set $r(Z) = 1$, we have

$$P(D^* = 1|Z, S = 1) = c(Z)P(D = 1|Z)$$

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

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

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

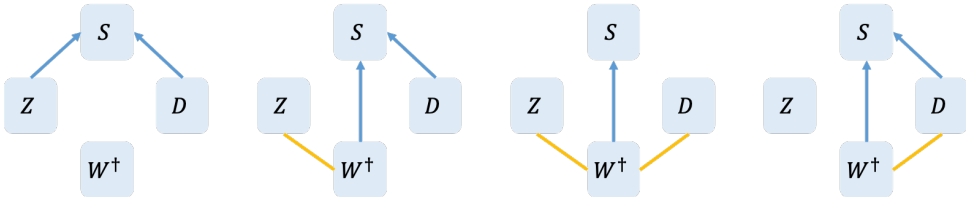
This produces the expression in *Eq. 4*.

A.2 Bias under naive 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 $c(Z)=1$, 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.1**. 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 (assuming $c(Z)=1$). Solid lines indicate associations, and arrows indicate drivers of patient selection.*



*Final setting will generally only result in small or negligible bias. See Neuhaus and Jewell (1993).

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 $\log\left[\frac{P(D^*=1|Z,S=1)}{c(Z)-P(D^*=1|Z,S=1)}\right]=\theta_0+\theta_Z Z$. Fitting a standard logistic regression will result in some bias in estimating θ for any $c(Z)\neq 1$. If both $c(Z)$ and $r(Z)$ are not equal to 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. (2018), 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 $\tilde{c}$, and sampling ratio, $\tilde{r}$. In that paper, we made additional restricting assumptions on X and W that, ultimately, boil down to the following: (1) $r(Z) = \tilde{r}$ and (2) $c(Z) = \tilde{c}$. In this particular setting, we showed that we can approximate θ_Z^{uc} as

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

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 $p^* = P(D^*=1|S=1) = \sum_{d=0,1} P(D^*=1|D=d|S=1)P(D=d|S=1) = P(D^*=1|D=1|S=1)P(D=1|S=1)$. We rewrite the above equation as

$$p^* = \tilde{c}P(D=1|S=1) = \tilde{c}\frac{\tilde{r}P(D=1)}{\tilde{r}P(D=1)+P(D=0)}$$

$$\implies P(D=1) = \frac{p^*}{p^*+(\tilde{c}-p^*)\tilde{r}} \qquad \implies P(D=0) = \frac{(\tilde{c}-p^*)\tilde{r}}{p^*+(\tilde{c}-p^*)\tilde{r}}$$

putting these together, we have

$$\begin{aligned}
\hat{\theta}_Z^{uc} &\approx \left[\frac{\frac{1}{1+e^{\theta_0+\theta_Z\bar{Z}}}}{\frac{e^{\theta_0+\theta_Z\bar{Z}}}{1+e^{\theta_0+\theta_Z\bar{Z}}}(1-\tilde{c})\tilde{r}+\frac{1}{1+e^{\theta_0+\theta_Z\bar{Z}}}}\right]\theta_Z \\
&\approx \left[\frac{(\tilde{c}-p^*)\tilde{r}}{p^*(1-\tilde{c})\tilde{r}+(\tilde{c}-p^*)\tilde{r}}\right]\theta_Z \\
&= \left[\frac{\tilde{c}-p^*}{\tilde{c}-p^*\tilde{c}}\right]\theta_Z \\
\implies \theta_Z &\approx \theta_Z^{uc}\frac{\tilde{c}(1-p^*)}{\tilde{c}-p^*}
\end{aligned}$$

This is the exact same structure as the estimator in Duffy et al. (2004), except this estimator is justified for Z that is non-binary and for $\tilde{r} \neq 1$ as well. One notable feature of the above estimator is that it does not depend on $\tilde{r}$. Under the restrictive assumptions on $r(Z)$ and $c(Z)$ 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. We note that, under selection ignorable for θ (i.e. $\tilde{r} = 1$), we get the same estimator as above, which gives us *Eq. 5*.

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

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

In reality, $\tilde{c}$ is unknown. However, we can obtain an estimate of $\tilde{c}$ 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. We have

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

$$\approx E\left(\hat{\theta}_Z^{uc}\right)^2 (1-p^*)^2 Var\left(\frac{\tilde{c}}{\tilde{c}-p^*}\right) + Var\left(\hat{\theta}_Z^{uc}\right)(1-p^*)^2 E\left(\left[\frac{\tilde{c}}{\tilde{c}-p^*}\right]^2\right)$$

Using Taylor series and other approximations, we have

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

This is now a function of known values along with $E(\tilde{c})$ and $Var(\tilde{c})$. We can insert our prior uncertainty about $\tilde{c}$ 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 ($\tilde{r} = 1$), we get expression

$$\theta_Z \approx \theta_Z^{uc} \frac{\tilde{c}(1-p^*)}{\tilde{c}-p^*}$$

where now $p^* = P(D^* = 1|S = 1) = P(D^* = 1)$ 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. We provide two conditions under which this replacement is appropriate. Here, we provide some support for these assertions.

First, we note that

$$c_{true}(X)P(D=1|Z,X^\dagger) = P(D^*=1|Z,X^\dagger)$$

Under a logistic regression model, this relationship implies

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

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

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

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

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

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

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

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. 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 observe that

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

since $D^* = 1$ implies $D = 1$. If we assume a logistic regression model structure for sensitivity as in *Eq. 1*, we have

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

This expression allows 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.

Suppose first that D is independent of X. In this case, we might 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. 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 can replace $P(D = 1|X)$ with known relationship $P(D = 1|X_{sub})$. This will allow us to estimate β.

Importantly, the above expression may not always have a solution for a given estimate $P(D = 1|X_{sub})$, and it may produce inaccurate sensitivity estimates when $P(D = 1|X_{sub})$ is poorly specified. An alternative strategy for estimating $c_{true}(X)$ is to fit a standard regression model for $P(D^* = 1|X)$ and use $c_{true}(X) = \min\left(\frac{P(D^*=1|X)}{P(D=1|X)}, 1\right)$ using estimates for both the numerator and denominator. In our experience, this estimator tends to be more robust to misspecification of $P(D = 1|X)$.

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

$$P(D = 1|S = 1, X) = \frac{P(S = 1|X, D = 1)P(D = 1|X)}{\sum_d P(S = 1|X, D = d)P(D = d|X)} \approx \frac{\tilde{r}P(D = 1|X)}{\tilde{r}P(D = 1|X) + P(D = 0|X)}$$

where we approximate $\frac{P(S=1|X,D=1)}{P(S=1|X,D=0)}$ with $\tilde{r} = \frac{P(S=1|D=1)}{P(S=1|D=0)}$. Using logic as above, we also have that

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

A.6 Jointly estimating θ and β

In this section, we describe how we can jointly estimate θ and β to deal with misclassification.

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 **Web Appendix 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. 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).

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 function and expected information matrix.

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

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 $\tilde{c}$ 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:

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

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:

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

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:

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

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

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

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, we clarify the expression used to estimate $P(S = 1|D, W)$ for obtaining IPW weights in **Section 4.1**. Assuming that no patients are included in both the internal and external datasets, we have that

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

and

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

Putting those pieces together, we have

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

We also have that

$$\begin{aligned}
P(S = 1|D, W, S_{all} = 1) &= \frac{P(S = 1, D, W|S_{all} = 1)}{P(D, W|S_{all} = 1)} \\
&= \frac{P(D, W|S = 1)P(S = 1|S_{all} = 1)}{P(D, W|S_{all} = 1)} \\
&= \frac{P(D, W|S = 1)P(S = 1|S_{all} = 1)}{\sum_d P(D, W|S_{all} = 1, S = d)P(S = d|S_{all} = 1)} \\
&= \frac{P(D, W|S = 1)P(S = 1|S_{all} = 1)}{P(D, W|S = 1)P(S = 1|S_{all} = 1) + P(D, W|S_{ext} = 1)P(S = 0|S_{all} = 1)} \\
\implies \frac{P(D, W|S = 1)}{P(D, W|S_{ext} = 1)} &= \frac{P(S = 1|D, W, S_{all} = 1)}{1 - P(S = 1|D, W, S_{all} = 1)}\frac{P(S = 0|S_{all} = 1)}{P(S = 1|S_{all} = 1)}
\end{aligned}$$

Therefore, we could also express

$$\begin{aligned}
P(S = 1|D, W) &= P(S_{ext} = 1|D, W)\frac{P(S = 1)}{P(S_{ext} = 1)}\frac{P(S = 1|D, W, S_{all} = 1)}{1 - P(S = 1|D, W, S_{all} = 1)}\frac{P(S_{ext} = 1|S_{all} = 1)}{P(S = 1|S_{all} = 1)} \\
&= P(S_{ext} = 1|D, W)\frac{P(S = 1|D, W, S_{all} = 1)}{1 - P(S = 1|D, W, S_{all} = 1)}
\end{aligned}$$

In practice, W may not be available for either the internal or external datasets, and a subset, W_{sub} might be used in its place. We would effectively be approximating $P(S = 1|D, W)$ using available $P(S = 1|D, W_{sub})$ in that case.

A.7.2 Assuming phenotype misclassification

Now, we suppose that we have phenotype misclassification, so D is not observed. In this case, the best we can do is estimate $P(S = 1|D^*, W)$. 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 method in *Eq. 7* 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 $P(D^* = 1|W) \approx cP(D = 1|W)$ and assuming $P(D = 1|W)$ is known or estimated using the external probability sample. Here, sensitivity c may be $\tilde{c}$ or

$c_{true}(X)$. 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.7.3 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:

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

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.8 Proof of *Eq. 8*

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

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

Now, we also have that

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

$$\implies P(D=1|S=1) = \frac{P(D^*=1|S=1)}{\tilde{c}}$$

$$\implies P(D=0|S=1) = \frac{\tilde{c} - P(D^*=1|S=1)}{\tilde{c}}$$

Putting these pieces together, we have

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

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 C.1** provides a visualization of the various mechanisms generating patient inclusion in MGI.

In light of these complicated selection stages, the strategy of modeling overall selection using a single model as in *Eq. 7* may be insufficient. Instead, we define a set of intermediate selection indicators corresponding to different inclusion mechanisms. In the MGI example, let S_1 might 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, and let S_4 indicate whether the patient was included in MGI, where $S_k = 1$ only if $S_{k-1} = 1$. Define $S_0 = 1$ for all patients in the population. In this updated notation, the overall sampling indicator S corresponds to S_4 and $P(S = 1|D, W) = \prod_{k=1}^{4} P(S_k = 1|D, W, S_{k-1} = 1)$.

Individual-level data available

Suppose first we have individual level data for all patients such that $S_1 = 1$, and the amount of individual-level data may differ for patients in subsequent samples (e.g. $S_2 = 1$, $S_3 = 1$, etc.). We expect to have the most individual-level data in the MGI dataset (patients with $S_4 = 1$). At a given stage $k > 1$, we model $P(S_k = 1|D, W, S_{k-1} = 1)$ using available covariates W_{k-1}, effectively approximating $P(S_k = 1|D, W, S_{k-1} = 1)$ with $P(S_k = 1|D, W_{k-1}, S_{k-1} = 1)$. This approach is used in Haneuse and Daniels (2016).

Unlike the mechanisms considered in Haneuse and Daniels (2016), $P(S_1 = 1|W, D, S_0 = 1) = P(S_1 = 1|W, D)$ cannot be directly estimated using the EHR data since we do not have individual-level data available for all patients in the population. However, if we have individual-level data on D and subset W_0 of W for a probability sample from the population, we can estimate $P(S_1 = 1|W_0, D)$ using *Eq. 7*, treating the $S_1 = 1$ patients as our internal sample. This allows us to bridge the gap between our large EHR (e.g. Michigan Medicine) and the population of interest.

Summary statistics available

Suppose instead that we know the joint distribution of D and W for the general population of interest and for our $S_4 = 1$ sample. We also note that

$$P(S_k = 1|D, W, S_{k-1} = 1) = \frac{P(D, W|S_k = 1)P(S_k = 1|S_{k-1} = 1)}{P(D, W|S_{k-1} = 1)} \qquad (Eq.\ S2)$$

Taking the product of *Eq. S2* over k, corresponding inverse probability of selection weights would be $\omega \propto \frac{P(D,W)}{P(D,W|S=1)}$. This expression recovers the poststratification weights discussed previously, where the contributions of intermediate sampling steps are multiplied out of the expression. In practice, we may not have true W available, and we will approximate ω using available variables in W, W_{sub}.

Mixed information across selection stages

Rather than having summary statistics or individual-level data uniformly across all selection stages, we may have individual-level data for some stages of the selection mechanism and only summary statistics for others. In this case, we can still express $P(S = 1|D, W) = \prod_{k=1}^{4} P(S_k = 1|D, W, S_{k-1} = 1)$, and we will obtain each $P(S_k = 1|D, W, S_{k-1} = 1)$ using the data type available at that selection stage. If individual-level data is available on the S_{k-1} sample, we can apply the methods above for individual-level data. If only summary statistics are available for the S_{k-1} sample, we can estimate $P(S_k = 1|D, W, S_{k-1} = 1)$ using *Eq. S2*. In this way, we can piece together estimates for each $P(S_k = 1|D, W, S_{k-1} = 1)$ based on the available information at each selection stage. This will allow us to obtain an estimate for $P(S = 1|D, W)$.

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. In general, we appeal to existing results in the maximum likelihood estimation and survey sampling literature.

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 $D^*\|Z$ distribution (Section 3.1) • $\text{Var}(\hat{\theta}_Z) \approx \text{Var}(\hat{\theta}_Z^{uc}) \left[\frac{\tilde{c}(1-P(D^*=1))}{\tilde{c}-P(D^*=1)}\right]^2$ where $\hat{\theta}_Z^{uc}$ is the uncorrected log-odds ratio. • $\text{Var}(\hat{\theta}_Z) \approx \hat{\theta}_Z^{uc\,2} \left[\frac{P(D^*=1)[1-P(D^*=1)]}{E(\tilde{c})-P(D^*=1)}\right]^2 \text{Var}(\tilde{c}) + \text{Var}\left(\hat{\theta}_Z^{uc}\right) \left[\frac{E(\tilde{c})[1-P(D^*=1)]}{E(\tilde{c})-P(D^*=1)}\right]^2$
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 $D^*\|Z$ 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 $\text{Var}(\hat{\theta}_Z) \approx \text{Var}(\hat{\theta}_Z^{uc}) \left[\frac{\tilde{c}(1-P(D^*=1))}{\tilde{c}-P(D^*=1)}\right]^2$. Since $\tilde{c}$ 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:

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

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:

$$I_{link}(\theta) = \sum_i \frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}+e^{\theta_0+\theta_Z Z}} \frac{e^{\theta_0+\theta_Z Z}}{(1+e^{\theta_0+\theta_Z Z})^2}(1,Z)^{\otimes 2} \qquad (Eq.\ S3)$$

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:

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

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

$$I_{obs}(\theta,\beta) = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}$$

Assuming D is invertible, we have that

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

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, $\text{trace}(BD^{-1}B^TA^{-1})$ will be positive. The question remains whether it will be greater than or less than 1. We generally expect $\text{trace}(BD^{-1}B^TA^{-1})$ 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. S3*. 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.

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 start 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 is then generated using the following relation: $\text{logit}\left(P(D=1|Z)\right) = -2 + 0.5Z$.

In *simulation part 1*, we then generate D^* using the sensitivity relation $\text{logit}\left(P(D^*=1|X, D=1)\right) = \beta_0 + X$ and assuming perfect specificity. We consider 5 different scenarios for β_0 and the association between X and Z as shown in **Table B.1**. The values of β_0 correspond to marginal sensitivities of roughly 0.4, 0.65, 0.8, and 0.95.

In *simulation part 2*, we define $D^* = D$. We allow 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 impose 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 consider 4 different simulation scenarios as shown in **Table B.1**. The ϕ values were chosen to give roughly a 50% selection probability on average.

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 $\sigma_{zx} = 0$ as in Setting 2 of simulation part 1. Many other simulation settings were explored with similar results, but these will not be presented here.

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. $\tilde{c}$ was estimated as $\frac{P(D^*=1)}{P(D=1)}$. $c_{true}(X)$ was estimated using the method in *Eq. 6* and assuming known $P(D=1|X)$.

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 (denoted "Population IPW") or estimated using a probability sample from that population and applying *Eq. 7*. 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**. $\tilde{c}$ and $c_{true}(X)$ were estimated using $\tilde{r}$ fixed at the simulation truth and using $\tilde{c} = \frac{P(D^*=1|S=1)}{P(D=1|S=1)} = P(D^*=1|S=1)\frac{\tilde{r}P(D=1)+P(D=0)}{\tilde{r}P(D=1)}$ or *Eq. 9* respectively. We used the correct IPW weights for these simulations rather than sample-estimated weights. Results are very similar when we estimate IPW weights using *Eq. 10*. Implementation of the observed data log-likelihood maximization method assumed fixed intercept $\beta_0 = \text{logit}(\tilde{c})$.

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 **Web Appendix A.10**.

Table B.1: Simulation set-up

Setting	Part 1		Part 2				
	β_0	σ_{zx}	ϕ_0	ϕ_D	ϕ_W	σ_{zw}	σ_{dw}
1	-0.4	0	-0.6	1	0.5	0.4	0
2	0.65	0	-0.6	1	-0.5	0.4	1
3	1.4	0	-0.2	0	-0.5	0.4	0
4	2.9	0	-0.1	0	0.5	0.4	1
5	-0.4	0.5	-	-	-	-	-

Part 3

B.2 Simulation 1: phenotype misclassification with ignorable sampling

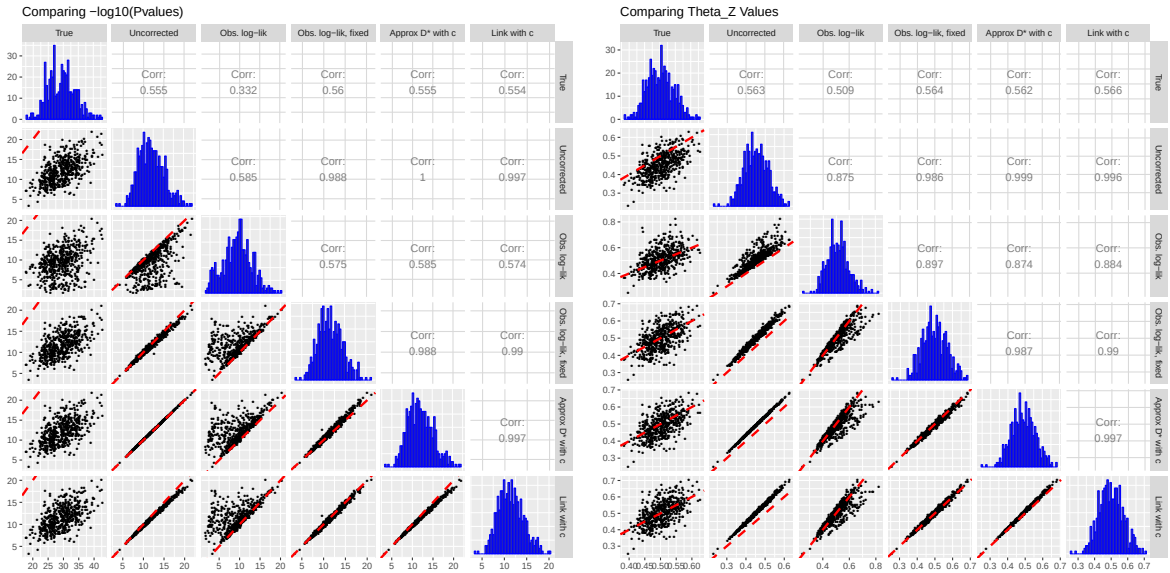
In **Section 7**, we present a set of three simulation studies exploring the performance of our proposed methods for handling (1) phenotype misclassification, (2) selection bias, and (3) both misclassification and selection bias. In this and the following two sections, we provide additional explorations into these simulation study results and additional evaluation of our proposed estimators for sensitivity and sampling/calibration weights. Our focus in *this* section is the setting where we have phenotype misclassification and can ignore the sampling mechanism.

B.2.1 Impact of correcting for misclassification on p-values

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 simulations when the outcome D^* has $\tilde{c} \approx 0.4$ and Z and X are independent (Setting 1 in **Table B.1**).

Figure B.1: Estimated p-values and θ_Z across 500 simulations after imposing phenotype misclassification with roughly 40% average sensitivity assuming X and Z are independent

(a) -log10(P-values) **(b)** Estimated θ_Z



The left panel of **Figure B.1** demonstrates that, with the exception of the strategy where we jointing estimate θ and β (which results in losses of efficiency relative to methods where we assume sensitivity is known), the p-values for the uncorrected and corrected analysis are nearly identical when X and Z are independent. As shown in the right panel, however, the resulting θ estimates often differ slightly. This indicates that, while misclassification is important to

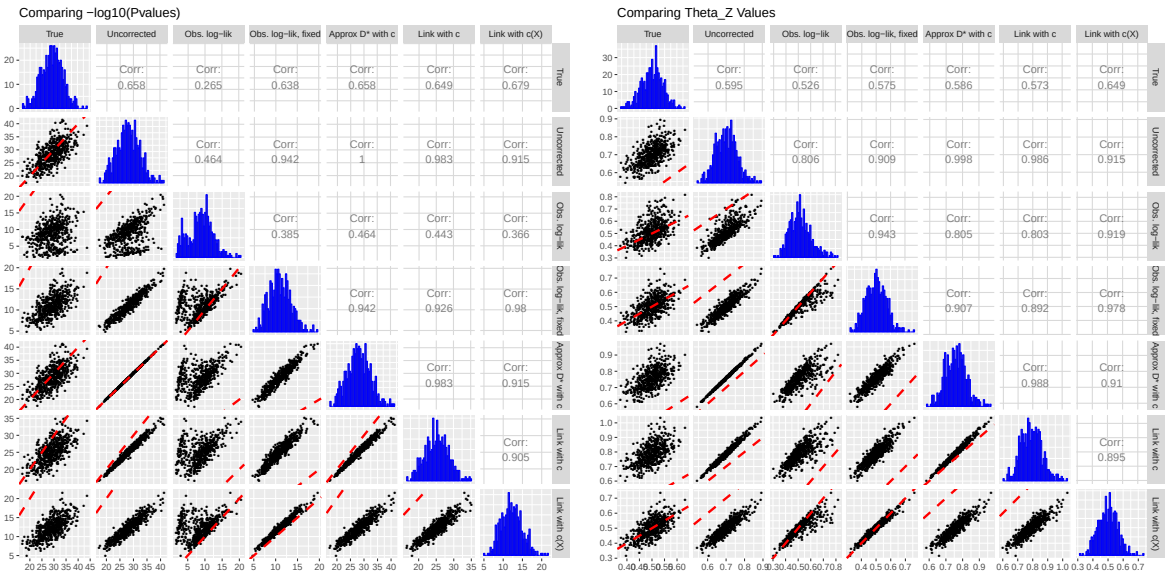
address for estimating θ, misclassification may be ignored in some cases (in particular, when we can assume independence between Z and X) when the primary interest is in estimating p-values.

Figure B.2 shows the same plots in the setting where X and Z are correlated (Setting 5 in **Table B.1**). In this case, both the estimated p-values and θ_Z values can differ substantially between the bias-corrected and uncorrected methods. An exception is the method in which we approximate $D^*|Z$. This method ignores covariate relationships in the sensitivity and therefore does not correctly handle the misclassification in this setting. This simulation demonstrates that, when X and Z are expected to be correlated, we can have a potentially substantial impact on both p-values and parameter estimates. While the p-values in the corrected and uncorrected data analyses are very highly correlated, the magnitudes of the estimated p-values are different.

These observations are particularly useful for PheWAS studies, where we compare p-values resulting from regression modeling of many different phenotypes, each of which may have 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. When X and Z may be related, however, accounting for misclassification across diseases can be important.

Figure B.2: Estimated p-values and θ_Z across 500 simulations after imposing phenotype misclassification with roughly 40% average sensitivity assuming X and Z have correlation 0.5

(a) -log10(P-values) **(b)** Estimated θ_Z



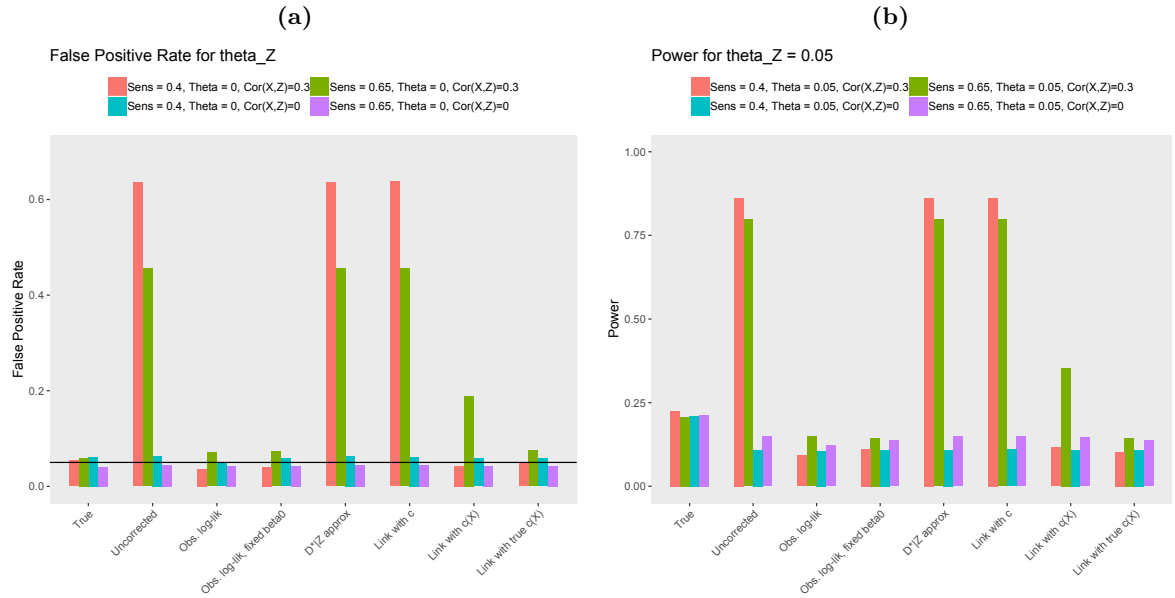
B.2.2 Power and type I error

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 and the correlation between X and Z. **Figure B.3** shows the results across 500 simulated datasets corresponding to 95% confidence intervals.

Figure B.3a shows the type I error rates. When X and Z are uncorrelated, we see nominal type I error rate across simulation settings considered, where the horizontal line in **Figure B.3** 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 and Z are correlated, however, we can see that the type I error for several of the methods is extremely large. This is due to bias resulting from the misclassification related to Z. For the methods that correct for misclassification by allowing a dependence between sensitivity and covariates, type I error rates generally return to nominal. An exception is the setting where X and Z are correlated, sensitivity is roughly 0.65, and we correct for misclassification using the non-logistic link function method with estimated $c(X)$. This may correspond to a setting where $c(X)$ is more difficult to estimate, and therefore, some residual bias remains.

Figure B.3b 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 settings where X and Z are uncorrelated, power is generally extremely low due to the small value of θ_Z. However, we can see some small differences across sensitivities, where higher sensitivity might have a slight edge in terms of higher power. When X and Z are correlated, power is substantially larger for methods that have bias away from the null. For methods that correct for the bias due to misclassification, the power naturally goes down as corrected point estimates move toward the null. Again, we see slightly inflated power in one particular setting for the non-logistic link function method with estimated $c(X)$.

Figure B.3: Estimated false positive rates and power across 500 simulations after imposing phenotype misclassification



B.2.3 Evaluating sensitivity estimates

Switching topics, we now focus on estimation of the sensitivity parameters themselves. In the main paper, we propose several strategies for estimating either marginal sensitivity or individual-level sensitivity using the observed data and some minor additional information about the population of interest. These methods are as follows:

Method 1: Estimate "crude" marginal sensitivity as $\tilde{c} = P(D^* = 1|D = 1) = \frac{P(D^*=1)}{P(D=1)}$
Method 2: Estimate $c_{true}(X)$ using link function method assuming $P(D = 1|X)$ is known: Fit the following model for $D^*|X$

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

Method 3: Estimate $c_{true}(X)$ through joint estimation of β and θ.

In **Figure B.4**, we evaluate the estimated $\tilde{c}$ from the above equations across 500 simulations in two simulation settings from **Table B.1**. In both simulation settings, the average sensitivity is roughly 0.4. In Setting 1, X and Z are independent, and they are correlated in Setting 5. In both settings, these estimates are well-centered around the true marginal sensitivity (in red).

In **Figure B.5**, we show the average $c_{true}(X)$ estimates across various methods and for 10 simulated datasets after applying either the non-logistic link function method with fixed $P(D = 1)$ or $P(D = 1|X)$ after applying the observed data log-likelihood maximization method (with or without a fixed β_0). In both simulation settings, the non-logistic link method and observed data log-likelihood maximization method with fixed intercept at $logit(\tilde{c})$ perform well. However, the sensitivity estimates from the observed data log-likelihood with arbitrary β_0 are very variable and do a poor job of recovering the truth for any one simulated dataset. This suggests that incorporating information into the observed data log-likelihood method about $P(D = 1)$ through fixed intercept at $\text{logit}(\tilde{c})$ can improve our ability to estimate the individual-level sensitivity substantially.

In **Figure B.6**, we show the estimated values for β_X for several methods and across 500 simulations. These estimates tend to be 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.

Figure B.4: Estimated $\tilde{c}$ across 500 simulations

(a) Setting 1 **(b)** Setting 5

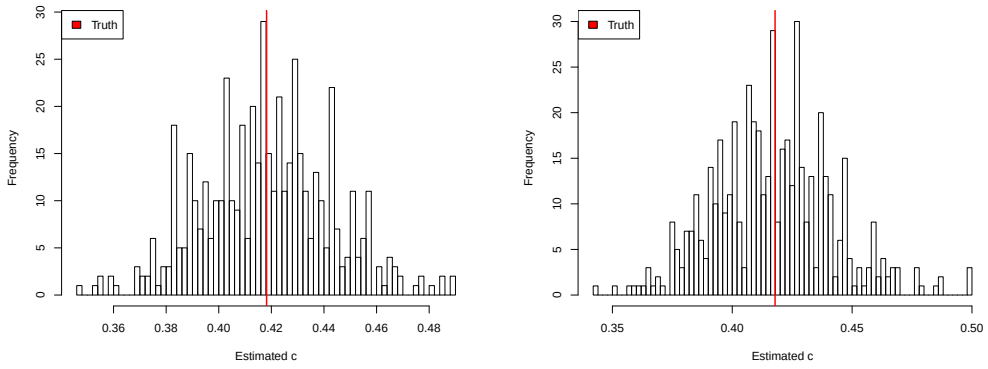


Figure B.5: Estimated $c_{true}(X)$ for 10 simulations.

(a) Setting 1 **(b)** Setting 5

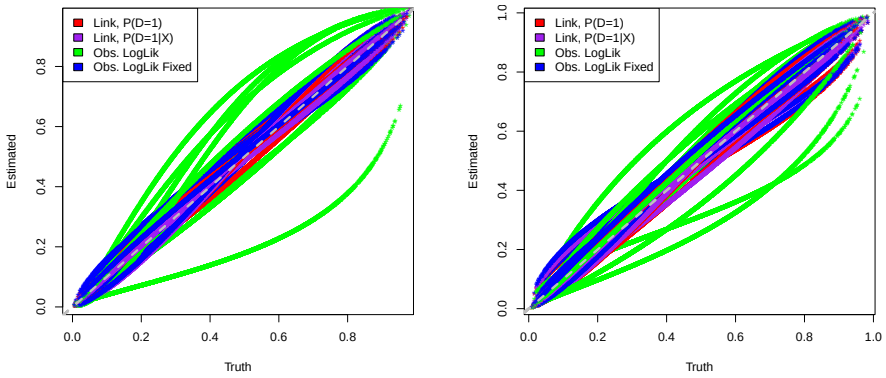
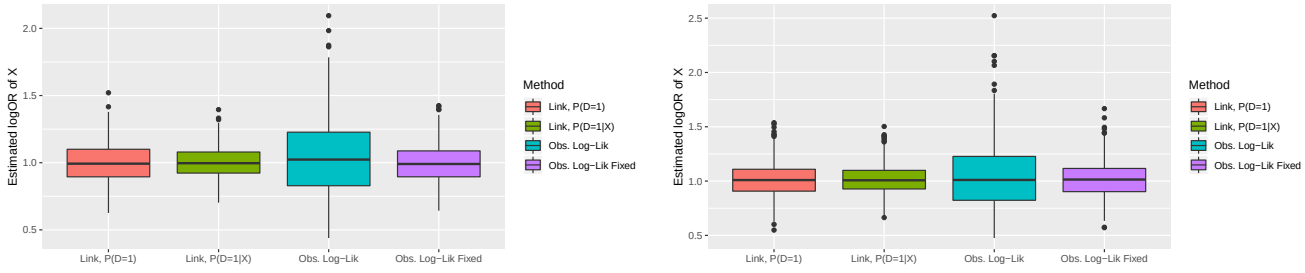


Figure B.6: Estimated β_X across 500 simulations (true $\beta_X = 1$)

(a) Setting 1 **(b)** Setting 5



B.3 Simulation 2: non-ignorable sampling with perfect phenotype classification

B.3.1 Evaluating estimator of sampling probabilities

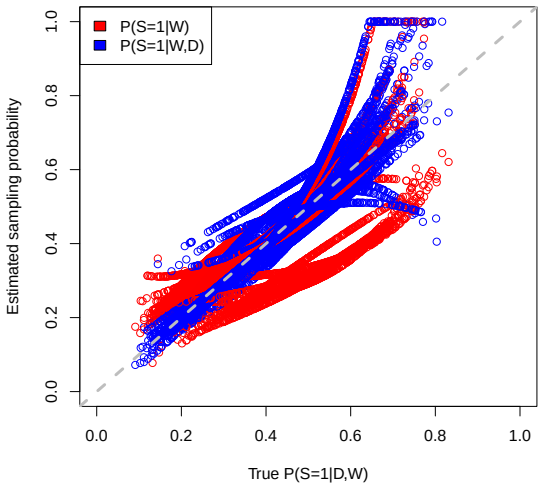
In this section, we demonstrate the ability of the method in Elliot (2009) to reasonably recover the sampling probability tied to the non-probability sample via simulation. Given the non-probability sample and a simple random sample drawn from the same population, we use the following equations to estimate the selection probabilities into the non-probability sample

$$\begin{aligned} P(S=1|D,W) &= P(S_{ext}=1|D,W)\frac{f(D|S=1,W)f(W|S=1)P(S=1)}{f(D|S_{ext}=1,W)f(W|S_{ext}=1)P(S_{ext}=1)} \\ &= P(S_{ext}=1|D,W)\frac{P(S=1|W,D,S_{all}=1)}{1-P(S=1|W,D,S_{all}=1)} \end{aligned}$$

We estimate $P(S=1|D,W)$ based on distributions of W and $D|W$ as above. When we simulate data such that selection depends on D and W and W is independent of D (Setting 1 in **Table B.1**), we obtain the estimated selection probabilities shown in **Figure B.7**. In this case, the true selection probability depends on both D and W, and modeling conditional only on W (to estimate $P(S=1|W)$) does not quite hit the mark. When we use the above method to estimate $P(S=1|W,D)$, however, we often do a reasonable job of recovering the sampling probabilities. We note that the above methods are not guaranteed to produce probabilities less than 1, and we apply an additional thresholding step that assigns all probabilities greater than 1 to the value 1. This explains the bending behavior we see for some of the predicted probability lines

in **Figure B.7**.

Figure B.7: Estimated sampling probabilities for 10 simulated datasets



B.4 Simulation 3: non-ignorable sampling and phenotype misclassification

In this section, we explore the performance of our proposed strategies for estimating sensitivity and sampling/calibration weights when both phenotype misclassification and potential selection bias are present. We explore a simulation setting where sensitivity is roughly 65% and in which both W and Z are marginally related to D. Sampling truly depends on both D and W, which corresponds to Setting 2 in **Table B.1**.

B.4.1 Determining the sampling ratio

To determine $\tilde{r}$ and $\tilde{c}$ jointly, we note that

$$\tilde{r} \approx \frac{P(D^* = 1|S = 1)}{\tilde{c} - P(D^* = 1|S = 1)} \frac{1 - P(D = 1)}{P(D = 1)} \qquad (Eq.\ S4)$$

We can use this relationship to plot a curve relating $\tilde{r}$ and $\tilde{c}$ as shown for 500 simulations in **Figure B.8**. True values for $\tilde{c}$ and $\tilde{r}$ are plotted as red lines, and we can see that the predicted curves intersect the true values. Using this plot, we can estimate either $\tilde{r}$ or $\tilde{c}$ by fixing a value for the other. In general, we can use our understanding of the problem and data to determine reasonable choices for $\tilde{r}$. Additionally, this can be used as a sensitivity analysis/tuning parameter, and analysis can be repeated for various values of $\tilde{r}$.

B.4.2 Estimating sensitivities

In this section, we evaluate the proposed methods for estimating sensitivity when we have both phenotype misclassification and selection bias. Suppose first that we estimate $\tilde{c}$. One strategy is to estimate $\tilde{c}$ given $\tilde{r}$ by inverting the equation in *Eq. S4*. Alternatively, we could also estimate $\tilde{c}$ using fixed ω assuming non-sampled patients would have the same marginal sensitivity as the sampled patients (not known in main paper). We can express $\tilde{c}$ as follows:

Method 1: Given $\tilde{r}$, we estimate $\tilde{c} = P(D^* = 1|S = 1)\frac{\tilde{r}P(D=1)+P(D=0)}{\tilde{r}P(D=1)}$

Method 2: Given ω, we estimate $\tilde{c} = \frac{p^*}{P(D=1)}$ where $p^* = \frac{\sum_i D_i^* \omega_i}{\sum_i \omega_i}$

Figure B.8: Predicted $\tilde{r}$ across different potential marginal sensitivities across 500 simulations

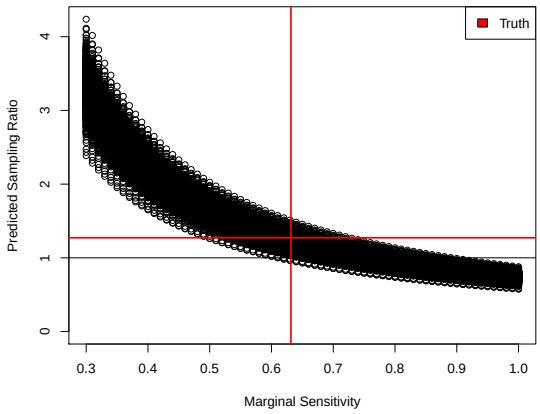
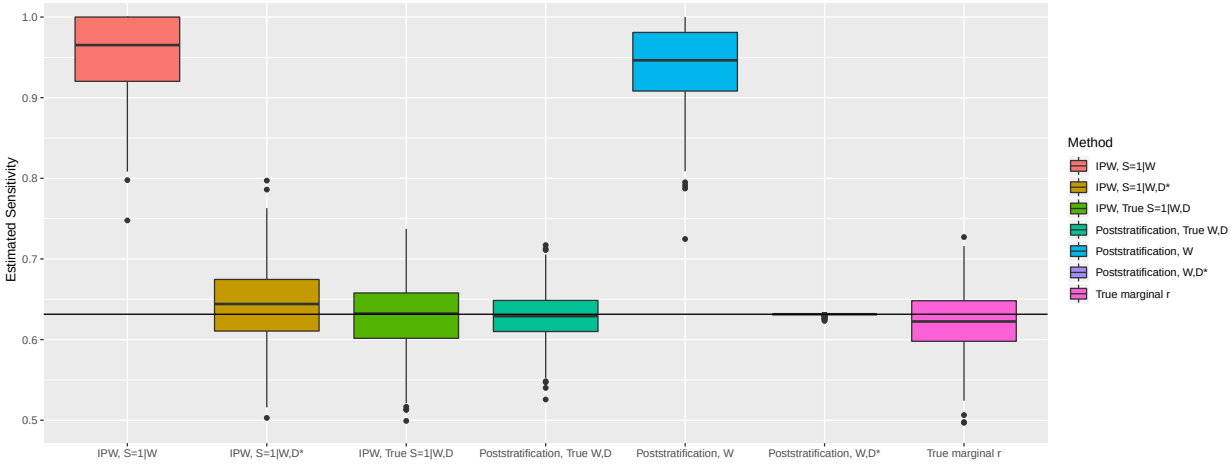


Figure B.9 provides the estimated marginal sensitivity across 500 simulations and applying various estimation methods, where the horizontal line represents the truth. We evaluate these expressions for $\tilde{c}$ either assuming that we know true $\tilde{r}$ or that various forms of ω are available. We consider ω specified using the true $P(S = 1|W, D)$ model, using estimated $P(S = 1|W, D^*)$, and using estimated $P(S = 1|W)$ (ignoring D entirely). We can see that the average estimated sensitivities across weighting methods that incorporate outcome information tend to be very close to the truth, and these estimates are better when true $P(S = 1|W, D)$ is used. When outcome information is *not* used to define ω, we can see substantial bias in the resulting estimate of $\tilde{c}$ (assuming sampling does indeed depend on D). We notice that the estimates of $\tilde{c}$ using poststratification weights given W and D^* have very small variability. This is because we used true $\tilde{c}$ to obtain the weights, which were then used to estimate $\tilde{c}$. In practice, we do not expect this same accuracy to be seen. Estimated $\tilde{c}$ using true $\tilde{r}$ also performs well.

Figure B.9: Estimated marginal sensitivity for 500 simulations:



Suppose instead that we want to estimate $c_{true}(X)$. We can estimate this as follows:

Method 1: Given $\tilde{r}$, we fit the following model for $D^*|X, S = 1$

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

where $P(D=1|X)$ or $P(D=1|X) = P(D=1|X_{sub})$ is assumed to be known.
Method 2: Given ω, we fit a weighted version of the following regression model

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

to the non-probability sample using weights ω.
Method 3: Given ω, maximize the weighted observed data log-likelihood.

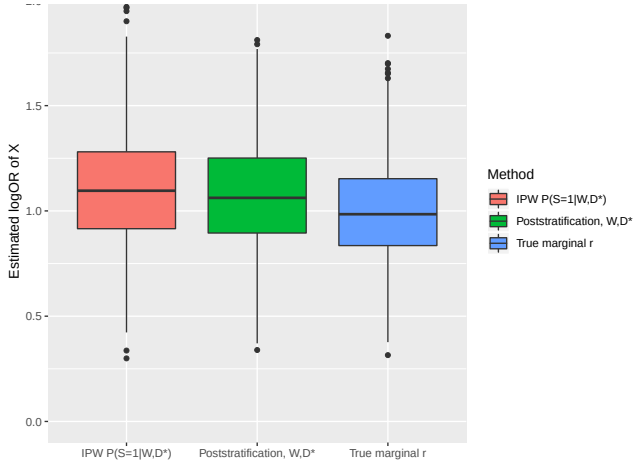
Method 2 for estimating $c_{true}(X)$ is not presented in the main paper but could also be used if ω were known under the assumption that $c_{true}(X)$ is the same for sampled and non-sampled patients.

Figure B.10 shows the results for Methods 1-2. In the left panel, we provide boxplots of the estimated β_X obtained using the above methods across 500 simulations. These values are reasonably well-centered around the true value of 1, but we note that the weighting methods using D^* have slightly worse performance. This is likely due to the replacement of $P(S=1|W,D)$ with $P(S=1|W,D^*)$ for weighting.

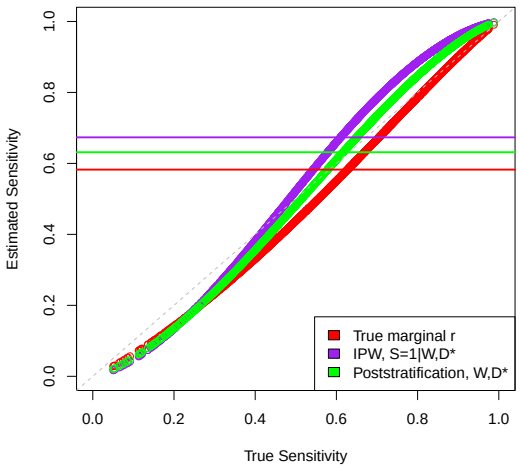
The right panel shows the estimated individual-level sensitivity values $c_{true}(X)$ for a single simulated dataset using various methods. Horizontal lines correspond to the estimated $\tilde{c}$, and the gray dotted line indicates equality between the true and estimated sensitivity values. This plot demonstrates that the above methods can do a good job of recovering the true sensitivity values when $\tilde{r}$ or ω is well-specified.

Figure B.10: Properties of estimated $c_{true}(X)$

(a) Estimated β_X values across 500 simulations



(b) Estimated $c_{true}(X)$ for one simulation. Horizontal bars represent the corresponding estimate for $\tilde{c}$

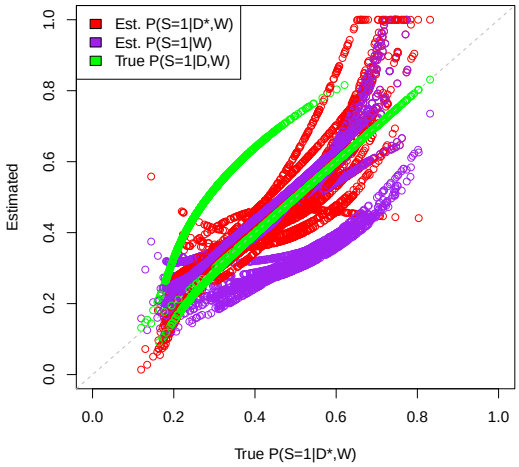


B.4.3 Estimating sampling weights

Determining ω is less straightforward than specifying $\tilde{r}$. Given true $\tilde{c}$ or $c_{true}(X)$, we estimate $P(S=1|W)$ and $P(S=1|W,D^*)$ using the method in **Web Appendix A.7**. **Figure B.11** provides estimated values for $P(S=1|D^*,W)$ and $P(S=1|W)$ for 10 simulated datasets. The simulation truth here is that sampling depends on both D and W. We can see that $P(S=1|W)$ does a poor job at recovering the target distribution. However, our proposed approach can do a reasonable job at estimating the true $P(S=1|D^*,W)$. Both of our estimated probabilities, however, do not fully capture the true $P(S=1|D,W)$. Therefore, there may be a potential for resulting bias in estimated θ. We saw evidence of a small negative impact of replacing

$P(S = 1|D, W)$ with $P(S = 1|D^*, W)$ on estimated $\tilde{c}$ previously.

Figure B.11: Estimated sampling probabilities for 10 simulated datasets



C Data analysis in MGI

C.1 MGI at a glance

Table C.1 provides descriptives for the MGI patients used in our analysis. We are particularly interested in length of follow-up and the number of visits to Michigan Medicine, since we hypothesize these variables may provide insight into phenotype misclassification. Analyses were limited to unrelated patients of recent European descent.

Many patient selection mechanisms went into generating our analytical dataset. If we define our target population as the US adult population, then we have a mechanism by which patients from the population enter the Michigan Medicine EHR. From there, patients must go to a MGI recruitment clinic, be approached for consent, and then consent to have their data included in MGI (which requires participants to donate biosamples and allows researchers EHR access). These various mechanisms are summarized in **Figure C.1**.

Table C.1: Descriptives of MGI Dataset, N = 40,101

	N (%) or mean (min-max)
Age at first diagnosis (years)	48.8 (1-95)
Age at last diagnosis (years)	56.7 (18-103)
Female	21,021 (52.4)
Cancer diagnosis	21,345 (53.2)
Follow-up time (years)	8.0 (0-40.2)
Number of visits	84 (1-1,323)
Number of unique phecodes	68.5 (1-608)

Figure C.1: Visualization of data generation mechanisms for MGI. Each layer corresponds to a stage of patient selection, and the sixth layer corresponds to misclassification of the true disease status, likely related to each patient's observation process.

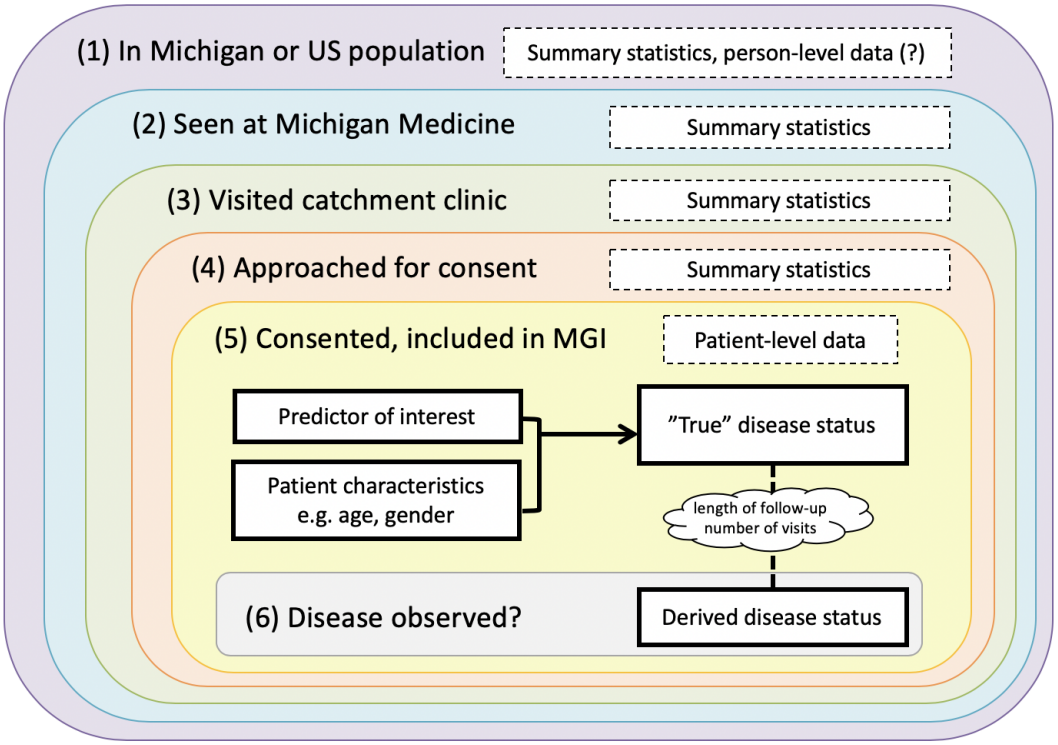


Table C.2: Relationship between MGI examples and conceptual model

Example	D	D^*	Z	X	W
1	latent disease status*	disease phecode	age**, gender	age, follow-up (years), log(visits per follow-up year)	n/a
2a	cancer phecode†	corrupted phecode	gender	follow-up (years), gender	n/a
2b	latent cancer status	cancer phecode	gender	age, follow-up (years), log(visits per follow-up year)	D, age, gender
3	latent AMD‡ status	AMD phecode	genotype, age, batch, gender, PCs 1-4	age, follow-up (years), log(visits per follow-up year)	D, age

* cancer of any type, colorectal cancer, melanoma, diabetes, and hypothyroidism
** age at last diagnosis in EHR
† takes value 1 if patient has any phecode corresponding to cancer diagnosis
‡ age-related macular degeneration

C.2 MGI example 1: factors related to sensitivity

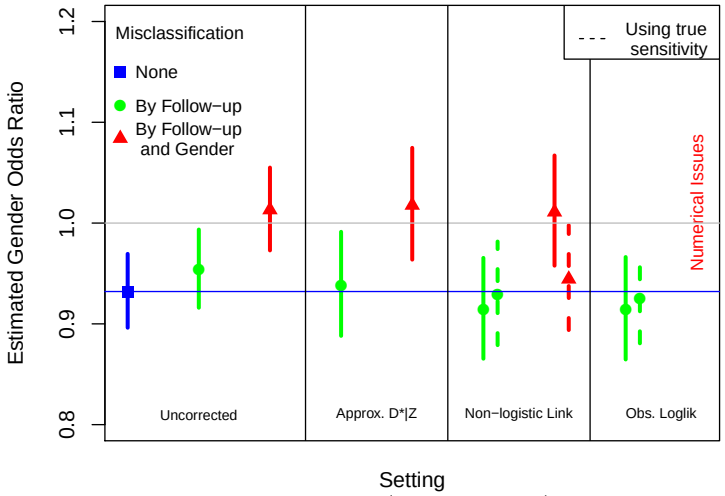
In this section, we explore factors related to sensitivity for several diseases in MGI, including cancer of any type, colorectal cancer, diabetes, hypothyroidism, and melanoma. For this analysis, we do not attempt to account for potential selection bias. We speculate that higher sensitivity may be related to longer follow-up time and a greater number of unique visit days in the EHR. All chosen diseases are associated with greater age, so we incorporated age into the model for sensitivity as well. For each of several diseases, we apply the methods in **Section 3.3** to estimate β and the corresponding patient-specific sensitivity estimate, $c_{true}(X)$. We adjust for gender and age in the disease model.

Figure 6a shows the resulting estimated β for various EHR-derived disease variables in MGI, and **Figure 6b** shows the distributions of estimated sensitivity across patients in MGI for different diseases. There are several surprising results. Firstly, the estimated distribution of sensitivity for diabetes indicated poor sensitivity on average. Recall, many patients in MGI were recruited prior to surgery. We expect it to be routine for doctors to inquire about diabetes status prior to surgery. While diabetes status may well be recorded in doctors' notes, it may not always be recorded in the EHR as a *diagnosis code.* We would expect higher estimated sensitivity values across the board if we defined EHR-derived disease status using more advanced phenotyping practices. Interestingly, the estimated sensitivity values for overall cancer diagnosis were generally high compared to the other diseases. We speculate that this may be due to routine collection of information about historical cancer diagnoses.

C.3 MGI example 2a: association between cancer diagnosis and gender

Suppose we treat EHR-derived cancer diagnosis as the *truth*, D.Given D, we then impose misclassification (generate D^*) using two different mechanisms: (1) patients with longer follow-up are more likely to have observed disease and (2) patients with longer follow-up and female patients are more likely to have observed disease, each resulting in an average sensitivity of about 70%. We apply methods in **Section 3** to correct resulting bias in the gender odds ratio. Results are shown in **Figure C.2**.

Figure C.2: Estimated MGI cancer and gender odds ratio after imposed misclassification and correction (reference category = male)*



*Solid lines indicate estimation using no bias correction ("uncorrected") or using estimated sensitivity (other methods). Dashed lines indicate use of the true sensitivity. Methods from **Sections 3.1-3.3** are applied.

C.4 MGI example 2b: association between cancer diagnosis and gender

In this section, we walk through an exploration into the relationship between cancer and gender using data from MGI. First, we attempt to estimate the degree of misclassification in cancer diagnosis and its relationship to patient-specific characteristics. Then, we explore the patient selection mechanism from the population using data from NHANES and summary statistics from the US Census and SEER. Finally, we put these pieces together and apply the methods discussed in the main paper to estimate the relationship between cancer and gender, adjusting for misclassification and the selection mechanism. An overview of the external data sources used in this analysis can be found in **Table C.3**. Analyses were limited to unrelated patients of recent European descent.

Table C.3: External data sources used for selection bias adjustment

Weight Type	External Information	Data Source
Poststratification	Age distribution in US	2010 US Census
Poststratification	Cancer prevalence by age	SEER 2016 Invasive Cancers*
IPW	Individual data on age, gender, cancer status	NHANES 2011-2016*

*Surveillance, Epidemiology, and End Results; National Health and Nutrition Examination Survey

C.4.1 Estimating sensitivity of EHR-derived cancer phenotype

Our estimators for sensitivity primarily rely on prior estimates of either $\tilde{r}$ or IPW/calibration weights ω. Here, we explore sensitivity $c_{true}(X)$ as a function of different fixed values of $\tilde{r}$. We can express $\tilde{r}$ as a function of the data and unknown parameter $\tilde{c}$ as in *Eq. 8*. Using the observed data, we obtain the relationship expressed in **Figure C.3a**.

Given several possible values of $\tilde{r}$ (in particular, 25, 50, 100, and 250), we want to calculate patient-level sensitivities $c_{true}(X)$ using *Eq. 9*. This estimator also requires us to specify $P(D = 1|X)$. We expect true cancer status to be related to both the length of follow-up and the number of visits in the EHR, and we do not know the true relationship between cancer status and these variables. However, we can obtain an estimate of the relationship between age and invasive cancer prevalence through 2016 using SEER (Surveillance, Epidemiology, and End Results) data resources available at `https://seer.cancer.gov/csr/1975_2016/results_merged/topic_prevalence.pdf`. Defining $P(D = 1|X) = P(D = 1|X_{sub})$ to be these prevalence rates by age, we can estimate $c_{true}(X)$ across different values of $\tilde{r}$.

The sorted sensitivity estimates across values of $\tilde{r}$ are shown in **Figure C.3b**, and **Figure C.3c** shows corresponding estimates for β.

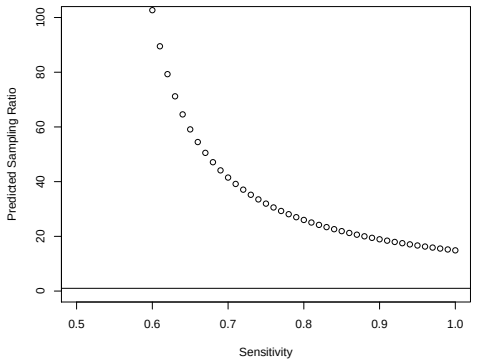
In addition to estimating $c_{true}(X)$ using *Eq. 9*, we can also estimate $c_{true}(X)$ ignoring the potential selection bias using the joint estimation strategy in **Section 3.3** based on the observed data log-likelihood. When selection bias is present, the observed data log-likelihood method will provide a poor estimate of θ, but we hypothesize that the estimated β may still provide some insight into $c_{true}(X)$. This seems reasonable since the impact of selection would enter the likelihood through offset term $r(Z)$ in the disease model. Failure to include $r(Z)$ in the model would be expected to impact θ more strongly than β. We plot the sensitivity estimate from the observed data log-likelihood method in **Figure C.3b** as well. The corresponding estimates for e^{β} are 1.77 (95% CI: 1.70, 1.85) for log(visits/follow-up), 1.13 (95% CI: 1.12, 1.14) for follow-up years, and 1.80 (95% CI: 1.75, 1.87) for age.

Knowing nothing else about $\tilde{r}$, we will use the estimated sensitivity values $c_{true}(X)$ obtained assuming $\tilde{r} = 250$. We chose this value given the similarity between its predicted sensitivity values and values from the observed log-likelihood method, which did not rely on any outside

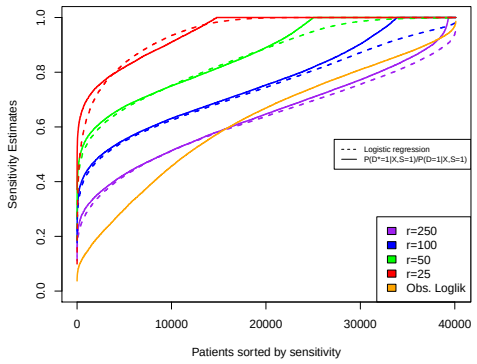
information. In practice, it may be desirable to repeat the target analysis for several different values of $\tilde{r}$.

Figure C.3: Sensitivity estimation

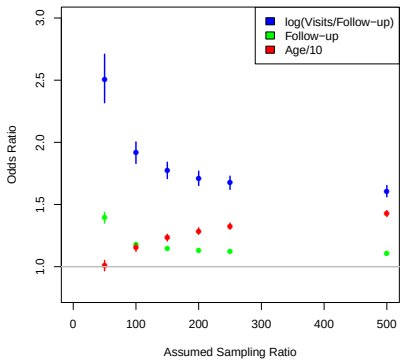
(a) Sampling ratio $\tilde{r}$ as a function of $\tilde{c}$



(b) Estimated sensitivities $c_{true}(X)$ across $\tilde{r}$



(c) Estimated β across $\tilde{r}$ values via logistic regression



C.4.2 Estimating poststratification weights using population summary statistics

We first obtain US Census summary statistics describing the age distribution of the US population from `https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml` and compare the US age distribution to the distribution of patient ages at the first and last times of follow-up in MGI. **Figure C.4** shows the results. We can see that the age distribution in MGI strongly differs from the US population. In particular, MGI patients tend to be older than patients in the general population.

We then obtain US overall invasive cancer prevalence rates by age from SEER at `https://seer.cancer.gov/csr/1975_2016/results_merged/topic_prevalence.pdf`. We compare those rates with observed cancer rates in MGI in **Figure C.4**, where these rates in MGI are either based on the age of last diagnosis or estimated using a more complicated prevalence estimate that incorporates the length of follow-up for each patient. Using the above quantities, we construct several different versions of the poststratification weights. Here, we define age to be the age at last follow-up.

VERSION 1: By age distribution only, where $\omega \propto f(\text{age})/f(\text{age}|S=1)$.

VERSION 2: By age and cancer, ignoring misclassification

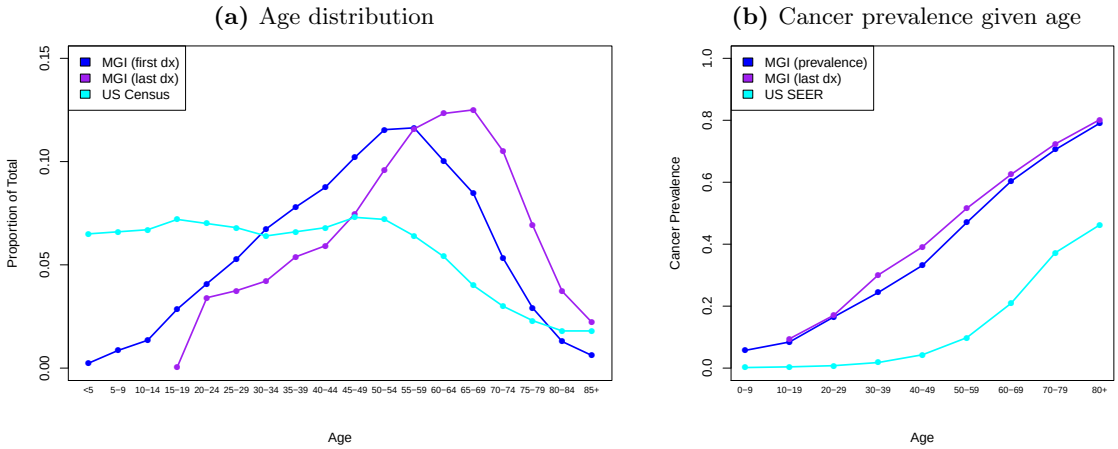
$$\omega \propto \frac{[1-P(D=1|\text{age})]^{1-D^*}\,[P(D=1|\text{age})]^{D^*}\,f(\text{age})}{f(D^*|\text{age},S=1)f(\text{age}|S=1)}$$

VERSION 3: By age and cancer, correcting for misclassification

$$\omega \propto \frac{[1-c_{true}(X)P(D=1|\text{age})]^{1-D^*}\,[c_{true}(X)P(D=1|\text{age})]^{D^*}\,f(\text{age})}{f(D^*|\text{age},S=1)f(\text{age}|S=1)}$$

where $c_{true}(X)$ is estimated as discussed previously.

Figure C.4: Comparing age and cancer distributions in MGI and US*



* Dx = diagnosis. Cancer prevalence in MGI was calculated either using the age of last diagnosis (denoted by "last dx") or using a better prevalence formula that incorporates the length of follow-up for each patient (denoted by "prevalence"). The age distribution in MGI was calculated either using the age at the first or last diagnosis in the EHR.

C.4.3 Estimating sampling probabilities using NHANES

In this section, we use an external probability sample from the US population of interest to obtain sampling probabilities (ω) for inverse probability weighting. In particular, we consider publicly available NHANES (National Health and Nutrition Estimation Survey) data from 2011-2016 consisting of N=28,709 patients with recorded age, gender, and cancer diagnosis history (yes/no). These data are available at `https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011`. We estimate several versions of ω incorporating outcome information in different ways as follows, where age in MGI is defined as the age at last follow-up in the EHR.

VERSION 1: Sampling probability estimated as a function of age and gender using *Eq. 7*

$$P(S=1|\text{age, gender}) = P(S_{ext}=1|\text{age, gender})\frac{P(S=1|\text{age, gender}, S_{all}=1)}{1-P(S=1|\text{age, gender}, S_{all}=1)}$$

VERSION 2: Sampling probability estimated as a function of age, gender, and cancer status but not correcting for misclassification (using *observed* cancer status D^* as if it were D in the non-probability sample in *Eq. 7*)

$$\begin{aligned}&P(S=1|\text{age, gender, cancer})\\ &= P(S_{ext}=1|\text{age, gender, cancer})\frac{P(S=1|\text{age, gender, cancer}, S_{all}=1)}{1-P(S=1|\text{age, gender, cancer}, S_{all}=1)}\end{aligned}$$

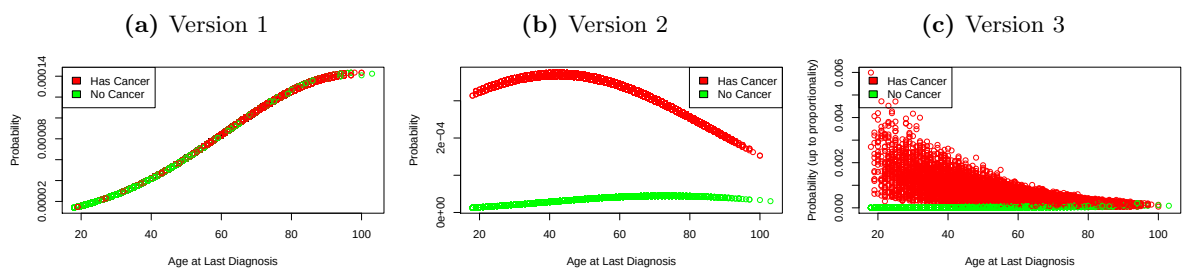
VERSION 3: Sampling probability estimated as a function of age, gender, and cancer status correcting for misclassification using *Eq. 10*

$$\begin{aligned}P(S=1|D^*,W) &= \frac{f(D^*|S=1,\text{age, gender})}{f(D^*|\text{age, gender})}P(S=1|\text{age, gender})\\ P(D^*=1|\text{age, gender}) &\approx c_{true}(X)P(D=1|\text{age, gender})\end{aligned}$$

where S_{all} is an indicator corresponding to the combined MGI and NHANES samples and S_{ext} is an indicator corresponding to inclusion in NHANES. $P(S_{ext}=1|\text{age, gender})$ is not available directly, but we estimate this distribution using beta regression on the provided NHANES sampling weights as suggested by Elliot (2009). $P(D^*=1|S=1,\text{age, gender})$ was estimated using the EHR data, and $P(D=1|\text{age, gender})$ was estimated using NHANES.

Figure C.5 shows the resulting selection probabilities as a function of the age at last diagnosis up to proportionality. The weights ignoring the outcome look strikingly different than the two version of the weights that incorporate the outcome. For each version of the weights, the inverse probabilities are then scaled to sum to the number of patients in MGI for use in estimation of θ later on.

Figure C.5: Estimated sampling probabilities (up to proportionality) using various methods



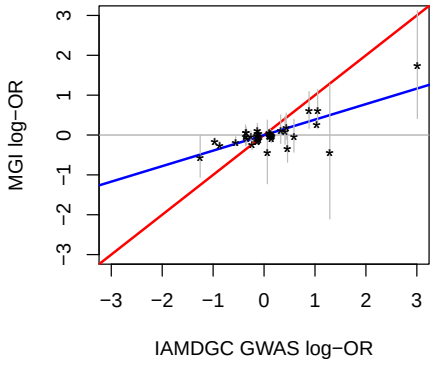
C.5 MGI example 3: correcting GWAS results for age-related macular degeneration

In this section, we apply the proposed methods to correct bias associated with the relationship between genetic factors and age-related macular degeneration (AMD) in MGI. We define whether patients have age-related macular degeneration in MGI as whether they ever received the phenotype code (aggregate of ICD codes) "362.2", which corresponds to "degeneration of macula and posterior pole of retina." This indicator becomes D^*, with latent D representing true AMD status for these patients up to their current age. We note that this definition may include some patients whose macular degeneration was not age-related. In response, we restricted analyses to patients aged 50 and older. We then restricted our focus to a matched subsample of MGI participants, where up to 10 unique patients without AMD were matched to each patient with AMD. Matched controls were identified using exact matching on gender and genotyping array and then applying nearest neighbor propensity score matching on age at last diagnosis and the first four principal components of the genetic data using a matching caliper of 0.25.

For each patient in MGI, a genetic profile spanning millions of genetic loci is available. Using these data and adjusting for age, gender, genotyping array, and the first four principal components of the genetic data, we can estimate the relationship between AMD diagnosis and each one of the genetic loci using logistic regression. These variables represent Z. We are interested in comparing the "naive" AMD log-odds ratio estimates we get for several genetic loci of interest in MGI to corresponding estimates from a well-designed GWAS study.

Reference GWAS results were based on data from 16000 *advanced* AMD cases and 18000 controls as part of the International AMD Genomics Consortium (IAMDGC, `http://amdgenetics.org`). A small fraction of patients in MGI may also have been included in the IAMDGC dataset. We will use these IAMDGC GWAS results (denoted "reference" GWAS) as a comparative gold standard. We selected 44 independent genetic loci most strongly related to AMD in this large reference GWAS based on better-quality data. In **Figure C.6**, we compare the 44 IAMDGC GWAS estimates to the estimates from MGI. Point estimates from MGI appear to be attenuated relative to the estimates seen in the IAMDGC GWAS. There are several explanations. Firstly, we are considering loci that are most strongly related to AMD in the IAMDGC GWAS. These loci will naturally have very strong estimated log-odds ratios, which may be artificially inflated due to the "winner's curse." At the same time, we expect there to be some potential for bias in the MGI point estimates due to misclassification of the AMD phenotype and potential selection bias. Another explanation for smaller effect estimates in MGI is that less advanced AMD cases were also included, as were cases of macular degeneration in older adults that may not have been age-related. In this section, we apply our bias-correction methods to address bias in MGI and compare the resulting estimates with the IAMDGC GWAS results.

Figure C.6: Genome-wide significant loci for AMD and corresponding log-OR in MGI. The diagonal red line corresponds to equality, and the blue line corresponds to a linear regression fit to the point estimates.



C.5.1 Estimating sensitivity

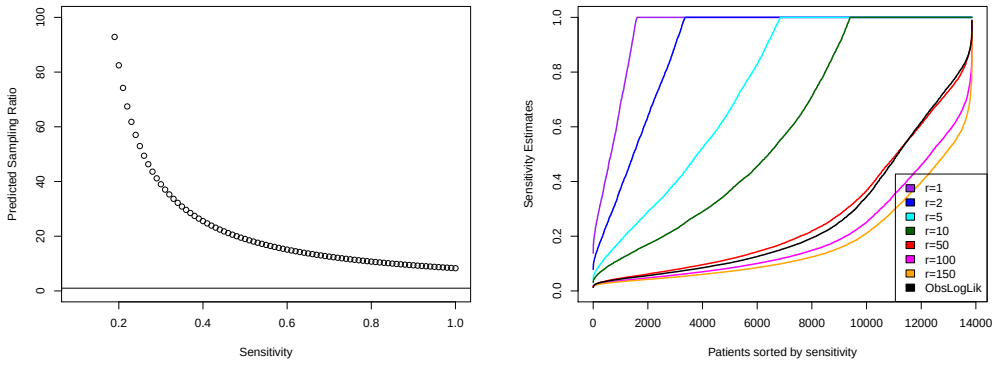
First, we estimate AMD sensitivity as a function of unknown marginal sampling ratio, $\tilde{r}$. We can express $\tilde{r}$ as a function of the data and unknown parameter $\tilde{c}$ as in *Eq. 8*. Using the observed data, we obtain the relationship expressed in **Figure C.7a**.

Given several possible values of $\tilde{r}$ (in particular, we consider 1,2,5,10,50,100, and 150), we calculate patient-level sensitivities $c_{true}(X)$ using $\hat{c}_{true}(X) \approx \min\left(\frac{\hat{P}(D^*=1|X,S=1)}{\tilde{r}P(D=1|X)+P(D=0|X)}, 1\right)$. We assume X contains age at last diagnosis, length of follow-up in years, and the log of the number of visits per follow-up year. We approximate $P(D=1|X)$ with $P(D=1|X_{sub}=\text{age})$. We obtain the relationship between X_{sub} =age and AMD status D from NIH as described below. The sorted sensitivity estimates across values of $\tilde{r}$ are shown in **Figure C.7b**.

In addition to estimating $c_{true}(X)$ fixing $\tilde{r}$, we also estimate $c_{true}(X)$ ignoring the potential selection bias using the joint estimation strategy in **Section 3.3**, where age and gender are included in the disease model. We plot the resulting sensitivity estimate in **Figure C.7b** as well. Knowing nothing else about $\tilde{r}$, we will use the estimated sensitivity values $c_{true}(X)$ obtained assuming $\tilde{r}=50$. We chose this value given the similarity to predicted sensitivity values from the observed log-likelihood method, which did not rely on any outside information. In practice, it may be desirable to repeat the target analysis for several different values of $\tilde{r}$.

Figure C.7: Estimating AMD sensitivity in MGI

(a) Sampling ratio $\tilde{r}$ as a function of $\tilde{c}$ **(b)** Estimated sensitivities $c_{true}(X)$ across $\tilde{r}$



C.5.2 Estimating calibration weights

We estimate poststratification weights combining population summary statistics with our estimates of sensitivity. In defining these weights, we combine estimates of the US age distribution available for 2010 from the US census with US macular degeneration prevalence rates by age for white Americans from the NIH National Eye Institute online summary statistics (also available for 2010). We use these prevalence rates to estimate the following

$$\omega \propto \frac{f(D^*, \text{age})}{f(D^*, \text{age}|S=1)} = \frac{[1-c_{true}(X)P(D=1|\text{age})]^{1-D^*}\,[c_{true}(X)P(D=1|\text{age})]^{D^*}\,f(\text{age})}{f(D^*|\text{age}, S=1)f(\text{age}|S=1)}$$

Here, W_{sub} contains age only and $P(D^*=1|\text{age}, S=1)$ is estimated directly using MGI.

C.5.3 Bias-corrected genetic associations

Using the sensitivity estimates and calibration weights, we apply the methods discussed in **Section 5** to estimate misclassification and selection bias-corrected θ for each one of the 44 genetic loci of interest. We apply three bias-correction strategies (1) accounting for misclassification

only (so setting all weights $\omega = 1$) and (2) accounting for misclassification and also weighting by poststratification weights ω. We then compare these resulting θ estimates with the estimates from the IAMDGC GWAS. We calculate the sum of squared differences between each of the 44 estimates in MGI and the IAMDGC GWAS along with the sum of the absolute differences. We also compare Spearman correlation and Lin's concordance correlation coefficient for the point estimates across the 44 loci. Finally, we present the average ratio of the estimated variance relative to the variance in the IAMDGC GWAS estimate. Results are shown in **Table C.4**.

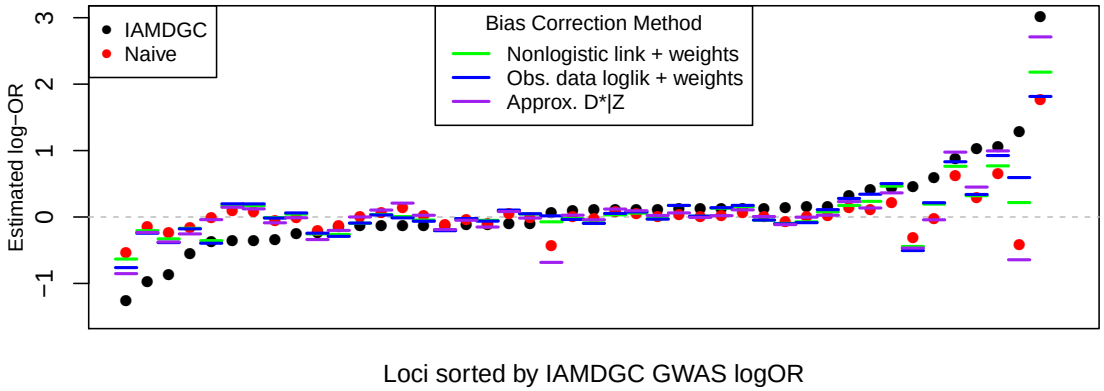
With the exception of the method that assumes constant sensitivity across patients (the method in which we approximate the $D^*|Z$ distribution), the methods adjusting for selection bias through calibration weighting tend to do a better job at recovering the IAMDGC GWAS point estimates. The non-logistic link function method and maximization of the observed data log-likelihood (both adjusting for selection through calibration weighting) produce point estimates closest to the IAMDGC GWAS estimates. **Figure C.8** shows the difference between the resulting estimated log-odds ratio estimates and the IAMDGC GWAS estimates for the 44 genetic loci (along the x-axis). No method uniformly maps the point estimates to the IAMDGC GWAS estimates. This does demonstrate, however, that the point estimates can sometimes differ substantially between the various methods for a given genetic locus, and these differences here are more pronounced for extreme values of the IAMDGC GWAS θ (far left and far right values).

Table C.4: Comparison between IAMDGC GWAS log-OR and bias-corrected log-OR across 44 genetic loci. Shaded boxes indicate the methods with the best performance for each column metric.

	Sum of squared differences	Sum of absolute differences	Spearman Correlation	Lin's CCC*	Average relative variance
IAMDGC GWAS	0	0	1	1	1
Naive analysis	13.77	9.38	0.43	0.61	5.86
Approx. of $D^*\|Z$	12.27	8.31	0.42	0.73	14.10
Approx. of $D^*\|Z + \omega$	12.73	9.48	0.58	0.60	12.49
Non-logistic link	13.09	9.85	0.38	0.68	14.81
Non-logistic link + ω	11.42	6.03	0.63	0.77	15.01
Obs. data log-lik	12.84	8.66	0.41	0.75	18.58
Obs. data log-lik + ω	10.80	5.86	0.63	0.77	22.66

* concordance correlation coefficient

Figure C.8: GWAS log-OR estimates after bias correction



D Implementation

D.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 $\tilde{c}$ 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. 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 $\tilde{c}$ 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,
    param_init = 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')
```

D.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 $\tilde{c}$ 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 $\tilde{c}$, 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 ($\tilde{c}$ 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) $\tilde{c} = \frac{P(D^*=1)}{P(D=1)}$ 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, $\tilde{c}$ 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 $\tilde{c}$ for that disease as follows:

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

If we have such gold standard information (e.g. associations with gender) for many diseases, we can use this information to estimate $\tilde{c}$ 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 $\tilde{c}$ 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. A Modeling Framework for Exploring Sampling and Observation Process Biases in Genome and Phenome-wide Association Studies using Electronic Health Records. *bioRXiv*, pages 1–19, 2018.

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 ? A General Framework for Considering Selection Bias in EHR-Based. *eGEM*, 4(1):1–17, 2016.

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 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.