Biometrics 000, 000–000 DOI: 000
October 2019

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

Summary: Health research using electronic health records (EHR) has gained popularity, but misclassification of EHR-derived disease status and lack of representativeness of the study sample can result in substantial bias in effect estimates and can impact power and type I error. In this paper, we develop new strategies for handling disease status misclassification and selection bias in EHR-based association studies. We first focus on each type of bias separately. For misclassification, we propose three novel likelihood-based bias correction strategies. A distinguishing feature of the EHR setting is that misclassification may be *related to patient-varying factors*, and the proposed methods leverage data in the EHR to estimate misclassification rates *without gold standard labels*. For addressing selection bias, we describe how calibration and inverse probability weighting methods from the survey sampling literature can be extended and applied to the EHR setting.

Addressing misclassification and selection biases simultaneously is a more challenging problem than dealing with each on its own, and we propose several new strategies. For all methods proposed, we derive valid standard error estimators and provide software for implementation. We provide a new suite of statistical estimation and inference strategies for addressing misclassification and selection bias simultaneously that is tailored to problems arising in EHR data analysis. We apply these methods to data from The Michigan Genomics Initiative (MGI), a longitudinal EHR-linked biorepository.

Key words: biobank, electronic health records, non-probability sampling, outcome misclassification, selection bias

This paper has been submitted for consideration for publication in *Biometrics*

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/biom.13400

1 Introduction

Health research using data from large observational databases such as electronic health records (EHR) has become increasingly popular (Beesley et al., 2019). Longitudinal, time-stamped EHR data allow researchers to study a wide array of diseases across patients' entire course of medical care, and linkages to other data sources such as census, death records, prescription claims, or genomic data provide a data-rich environment for health research. Additionally, EHR data are often collected without a specific hypothesis in mind, allowing many researchers to use the same dataset to study many scientific questions in a convenient and cost-effective way. In some applications, these results can then be more easily translated into improvements in patient care through return of results and real-time risk prediction. This provides opportunities for immediate, actionable translation of generated knowledge. Unlike curated and well-designed population-based studies, these databases are rarely originally intended for research use, and patient recruitment processes may not be well understood. Without properly accounting for design issues (e.g. who is in the sample, how data were measured), association analyses using these data are naturally susceptible to bias (Beesley et al., 2020). With larger datasets at researchers' fingertips, the impact of bias relative to variance is becoming more pronounced. In particular, these biases do not disappear with increased sample size, resulting in a large potential for "incorrect" inference with inflated type 1 error and suboptimal coverage. This phenomenon is known as the "big data paradox" (Meng et al., 2018), and statistical strategies for correcting these biases are needed. We focus on two common sources of bias for EHR data analysis: (1) misclassification of derived disease status (information bias) and (2) lack of representativeness (selection bias). We consider a common problem where one is interested in relating a binary disease phenotype D to predictors Z.

EHR-derived disease variables (phenotypes) can be misclassified for many reasons. Researchers often define disease status based on diagnosis codes recorded in the EHR for billing purposes, which provide a restricted snapshot of a patient's complete disease

history. Even the most sophisticated phenotyping methods are limited by the information available in the EHR. Over-reporting may be a concern for self-reported symptoms such as pain or fatigue, and preferential coding or up-coding may occur in response to insurance incentives (O'Malley et al., 2005). Under-reporting may often occur, since secondary conditions may be inconsistently recorded, past medical history may be incomplete, and symptoms between visits may be missed. For academic databases, patients may visit the hospital for short-term treatment and return to local providers for continued care. Taken together, these factors can lead to a large degree of misclassification relative to patients' true disease history, and we hypothesize that under-reporting of disease is the primary source of misclassification for many EHR phenotypes as a result of limited duration and comprehensiveness of follow-up. Several researchers have explored misclassification in EHR or claims data assuming *constant* sensitivity and specificity (Sinnott et al., 2014; Lange et al., 2015; Hubbard et al., 2015). However, a key feature of misclassification for EHR-derived phenotypes is that we expect more diagnoses to be missed for patients followed for a shorter period of time and for fewer visits, so *misclassification may depend on patients' individual observation patterns*. This problem has been discussed in the literature on EHR data analysis (e.g. Goldstein et al., 2016; Phelan et al., 2017). Even so, statistical literature handling this covariate-related misclassification is sparse. Neuhaus (1999) presented analytic expressions for bias under covariate-related misclassification, and Beesley et al. (2020) provided a sensitivity analysis approach tailored to the EHR setting. Ad hoc strategies including adjusting for number of encounters or clinic type have also been proposed (Goldstein et al., 2016; Phelan et al., 2017). In general, however, existing work considering covariate-related misclassification is limited, necessitating new statistical methods that can address this more complex misclassification mechanism.

EHR data are also susceptible to bias due to a lack of representativeness with respect to some population of interest, e.g. the US population. It can be difficult to understand the mechanism driving patient interactions with the health care system, which may be

related to many patient factors including overall health and access to care. When ignored, selection can negatively impact association analyses (Beesley et al., 2019). Patient selection can be addressed using survey techniques if the selection strategy is known, but this is *unknown* in the EHR setting. Researchers have partially accounted for selection bias by adjusting for factors such as referral status and clinic type (Phelan et al., 2017; Goldstein et al., 2016). Haneuse and Daniels (2016) developed a framework for modeling selection in EHR data as a series of selection steps. This strategy can be very useful for characterizing selection mechanisms generating an analytical sample from a bigger EHR database. However, these methods do not address the systematic differences between people that are and are not included in the EHR itself. To bridge this gap, strategies in the survey sampling literature for dealing with unknown selection probabilities (termed non-probability sampling) such as calibration weighting and inverse probability of selection weighting can be applied (Bower et al., 2017; Baker et al., 2013). Little work has been done to describe how such methods can be implemented in the EHR setting.

In this paper, we develop new, practical strategies for handling phenotype misclassification and selection bias in EHR-based association studies. We first focus on each type of bias separately. For misclassification, we propose three novel likelihood-based bias correction and inference strategies. These strategies allow us to estimate the rate of misclassification incorporating covariate relationships and require minimal external information and *no gold standard labels*. For addressing selection bias, we describe how calibration and inverse probability of selection weighting methods from the survey sampling literature can be modified and applied in the EHR setting. Addressing both sources of bias at once is more challenging, and we propose several new estimation and inference strategies. For all strategies proposed, we derive valid standard errors and provide software for implementation (R package, *SAMBA*). Through simulation, we demonstrate the ability of these methods to reduce or eliminate bias and correctly estimate standard

errors. We apply our proposed methods to data from The Michigan Genomics Initiative (MGI), a longitudinal EHR-linked biorepository within Michigan Medicine.

2 Model, notation, and conceptual framework

Let binary D represent a patient's true disease status. Suppose we are interested in the relationship between D and person-level information, Z. Z may contain genetic information or any other characteristics of interest. We call this the *disease mechanism*.

We consider a large EHR database with the goal of making inference about some *defined target population*. For example, we might define our target population as the US adult population between ages 50-65. This population may differ from our *source population* (e.g. people in the catchment area of the health system). We will assume that inference about $D|Z$ is transportable between the source and target populations (Dahabreh and Hernán, 2019). To simplify our discussion but without further loss of generality, we will imagine our target population is the same as the source population and will use these terms interchangeably. Let S indicate whether a given person in the *source/target population* is included in our data, where the probability of inclusion may depend on disease status, D, and additional covariates, W. Let $W^\dagger$ denote variables in W that are not adjusted for in the disease model (not in Z). We will use the terms "sampled" or "selected" interchangeably to refer to patients included in our EHR dataset. We may often expect the sampled and non-sampled people to have different rates of disease, and other factors such as age, residence, access to care, and general health state may also impact inclusion. We call this mechanism the *selection mechanism*. In reality, inclusion in the analytical dataset may be impacted by multiple selection phases as illustrated for MGI in **Figure A.3**. Here, we focus on the aggregate mechanism governing inclusion. In practice, we will rarely have W fully measured, and we consider theoretical W.

Instances of the disease are recorded in the EHR. Factors such as patient age, length of follow-up, and number of hospital visits may impact whether we *observe/record* the

disease for a given person. Let D^* be the *observed* disease status. D^* is a potentially misclassified version of D with corresponding sensitivity and specificity. We call the mechanisms generating D^* given D the *observation mechanisms.* Let X denote patient and provider-level predictors related to sensitivity, and let $X^\dagger$ denote the variables in X not included in Z. Let Y denote factors related to specificity. Later on, we will assume D^* has perfect specificity (no over-reporting) or that specificity is constant in Z. **Figure 1** shows the conceptual model, which is expressed mathematically in *Eq. 1*.

[Figure 1 about here.]

Conceptual Model (*Eq. 1*)

$$\begin{aligned}
\text{Disease Mechanism}: &\quad \text{logit}(P(D=1|Z;\theta)) = \theta_0 + \theta_Z Z \\
\text{Selection Mechanism}: &\quad P(S=1|D,W;\phi) \\
\text{Observation Mechanisms}: &\quad \text{logit}(P(D^*=1|D=1,S=1,X;\beta)) = \beta_0 + \beta_X X \\
&\quad \text{logit}(P(D^*=1|D=0,S=1,X;\beta)) = \psi_0 + \psi_Y Y
\end{aligned}$$

In our statistical development, we will often refer to the following functions of the observation and selection model parameters

$$\begin{aligned}
c_{true}(X) &= P(D^*=1|D=1,S=1,X;\beta) \qquad (Eq.\ 2)\\
c(Z) &= P(D^*=1|D=1,S=1,Z;\beta) = \int c_{true}(X) f(X^\dagger|Z,D=1,S=1) dX^\dagger \\
\widetilde{c} &= P(D^*=1|D=1,S=1;\beta) = \int c(Z) f(Z|D=1,S=1) dZ \\
r(Z) &= \frac{P(S=1|D=1,Z;\phi)}{P(S=1|D=0,Z;\phi)} = \frac{\int P(S=1|D=1,W;\phi) f(W^\dagger|Z,D=1) dW^\dagger}{\int P(S=1|D=0,W;\phi) f(W^\dagger|Z,D=0) dW^\dagger} \\
\widetilde{r} &= \frac{P(S=1|D=1;\phi)}{P(S=1|D=0;\phi)} = \frac{\int P(S=1|D=1,Z;\phi) f(Z|D=1) dZ}{\int P(S=1|D=0,Z;\phi) f(Z|D=0) dZ}
\end{aligned}$$

The first expression represents the generating sensitivity mechanism. The subsequent expressions show the average sensitivity as a function of Z and the overall marginal sensitivity, $\widetilde{c}$, both of which are implicit functions of β. The fourth expression represents

the sampling ratio with respect to D as a function of Z, and constant $\widetilde{r}$ represents the ratio of marginal sampling probabilities (here, called the marginal sampling ratio). We can define specificities $b(Z) = P(D^* = 0|D = 0, S = 1, Z; \psi)$ and $\widetilde{b}$ similarly.

A common approach is to model $D^*|Z, S = 1$ (analysis model) and interpret results under the target model, $D|Z$. To explore settings in which this approach produces bias, we relate the parameters in the conceptual and analysis models. In **Supporting Section A.1, we prove the following key relationship:**

$$P(D^* = 1|Z, S = 1) = \frac{1 - b(Z) + [c(Z)r(Z) - \{1 - b(Z)\}]\, P(D = 1|Z)}{1 + [r(Z) - 1]\, P(D = 1|Z)}. \qquad (Eq.\ 3)$$

Eq. 3 is an extension of Neuhaus (1999) allowing for covariate-related misclassification and incorporating selection. The contribution of misclassification and selection reduces to $c(Z)$, $b(Z)$, and $r(Z)$ in *Eq. 2*, where $c(Z)$ and $b(Z)$ represent misclassification and $r(Z)$ represents selection. Under distinctness of β, ψ, and ϕ in *Eq. 1*, $c(Z)$, $b(Z)$, and $r(Z)$ are independent functions of model parameters given Z. These three factors work together to generate bias in $P(D^* = 1|Z, S = 1)$ relative to $P(D = 1|Z)$. We explore this bias in **Supporting Section A.2**. Briefly, we will have bias in estimating θ_Z anytime we have misclassification. We will also have bias if $r(Z) \neq \widetilde{r}$ (sometimes, if $r(Z) \neq 1$). A special case is when $D|Z$ follows a logistic regression as in *Eq. 1*. In this case, we can show that

$$\log\left[\frac{P(D^* = 1|Z, S = 1) - \{1 - b(Z)\}}{c(Z) - P(D^* = 1|Z, S = 1)}\right] = \theta_0 + \theta_Z Z + \log\left[r(Z)\right]. \qquad (Eq.\ 4)$$

The left-hand side of the equation takes a GLM form with a different (non-logistic) link function, and the right-hand side contains an offset term as a function of the sampling ratio. If we knew $c(Z)$, $b(Z)$, and $r(Z)$, we could estimate θ by fitting *Eq. 4* to the observed data. For the remainder of this paper, **we assume that we have perfect specificity** $(b(Z) = 1)$ **or that** $b(Z)$ **is a known constant**, $\widetilde{b}$. In this setting, we provide strategies for estimating θ when $c(Z)$ and $r(Z)$ are unknown, all guided by *Eq. 4*.

3 Accounting for phenotype misclassification assuming ignorable selection

Suppose that patient selection is ignorable for θ. In other words, assume that $r(Z) = 1$, which occurs when selection does not depend on D given Z. This might happen if, for example, our target population is our internal hospital population. In this section, we propose strategies for estimating θ in *Eq. 4* accounting for unknown $c(Z)$.

3.1 Method 1: approximating $D^*|Z$ distribution

Suppose $c(Z)$ is independent of Z, so $c(Z) = \widetilde{c}$. This will be the case if X is independent of Z given D. This is a strong assumption that may be unrealistic for some EHR data analyses. For example, disease risk factors may be included in Z and related to misclassification through enhanced disease surveillance. Suppose further that $b(Z) = \widetilde{b}$ is known. If we know prevalence $P(D = 1)$, then we can estimate sensitivity as $\widetilde{c} = \frac{P(D^*=1)-[1-\widetilde{b}]P(D=0)}{P(D=1)}$. If $\widetilde{b} = 1$, we can estimate $\widetilde{c} = \frac{P(D^*=1)}{P(D=1)}$. In **Supporting Section A.3**, we use Taylor series approximations to relate true θ_Z to the uncorrected parameter θ_Z^{uc} as follows:

$$\theta_Z \approx \theta_Z^{uc}\left[\frac{\{\widetilde{c}-(1-\widetilde{b})\}\{1-P(D^*=1)\}P(D^*=1)}{\{\widetilde{c}-P(D^*=1)\}\{P(D^*=1)-(1-\widetilde{b})\}}\right]. \qquad (Eq.\ 5)$$

Replacing θ_Z^{uc} with an estimate, this expression recovers an existing estimator for *binary* Z in Duffy et al. (2004). We show we can apply *Eq. 5* more generally. This expression is convenient, because it can be applied when only summary statistics for θ_Z^{uc} are available.

3.2 Method 2: direct estimation of θ using a non-logistic link function

Suppose instead that $c(Z)$ is not constant in Z, so misclassification depends either directly on Z or on predictors related to Z given D. Suppose further that $b(Z) = \widetilde{b}$ is known. We can estimate θ using $\log\left[\frac{P(D^*=1|Z)-\{1-\widetilde{b}\}}{c(Z)-P(D^*=1|Z)}\right] = \theta_0 + \theta_Z Z$, which is a generalized linear model with a non-logistic link function. The question then becomes how to estimate $c(Z)$.

In **Supporting Section A.4**, we show that we can replace $c(Z)$ with estimated $c_{true}(X)$ if either (1) $X^\dagger$ is independent of Z given D or (2) $X^\dagger$ is independent of D given Z. The latter case may rarely hold, because $X^\dagger$ may contain information such as the length of

follow-up related to D. However, the former assumption may often be reasonable. As shown in **Supporting Section A.5**, we can estimate $c_{true}(X) = \text{expit}\left[\beta_0 + \beta_X X\right]$ using

$$\log\left[\frac{P(D^*=1|X) - \{1-\widetilde{b}\}P(D=0|X)}{P(D=1|X) + \{1-\widetilde{b}\}P(D=0|X) - P(D^*=1|X)}\right] = \beta_0 + \beta_X X, \qquad (Eq.\ 6)$$

assuming $P(D=1|X)$ is known. In practice, we will approximate $P(D=1|X)$ as in **Supporting Section A.5**. Importantly, *Eq. 6* may not always have a solution for a given estimate of $P(D=1|X)$, and we can instead estimate $c_{true}(X) = \min\left(\frac{P(D^*=1|X)-[1-\widetilde{b}]P(D=0|X)}{P(D=1|X)}, 1\right)$.

3.3 Method 3: joint estimation of β and θ using observed data log-likelihood

Rather than estimating sensitivity and θ in a two-step process, we can *jointly* estimate θ and sensitivity parameter β. For this estimation, we **assume perfect specificity** ($\widetilde{b} = 1$). Let $\perp$ represent conditional independence. If either (1) $X^\dagger \perp Z|D$ or (2) $X^\dagger \perp D|Z$, we can estimate (θ, β) using the following observed data log-likelihood: $\sum_i D_i^* \log\left[\frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}}\frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}\right] + (1-D_i^*)\log\left[1 - \frac{e^{\beta_0+\beta_X X_i}}{1+e^{\beta_0+\beta_X X_i}}\frac{e^{\theta_0+\theta_Z Z_i}}{1+e^{\theta_0+\theta_Z Z_i}}\right]$. This is a zero-inflated logistic regression. We jointly estimate θ and β by maximizing this log-likelihood through either a Newton-Raphson algorithm or expectation-maximization algorithm. We may run into numerical problems tied to weak identifiability in practice, which can be reduced by fixing a model parameter. We observed good performance when β_0 was fixed at $\text{logit}(\widetilde{c})$ for $\widetilde{c} = \frac{P(D^*=1)}{P(D=1)}$. Details are presented in **Supporting Section A.6**.

4 Accounting for patient selection under perfect classification

Suppose instead that D is perfectly observed (so $D^* = D$) and that we have some unobserved mechanism governing patient selection. We have that $\text{logit}\left[P(D=1|Z, S=1)\right] = \theta_0 + \theta_Z Z + \log\left[r(Z)\right]$, where $r(Z)$ is defined in *Eq. 2*. When $r(Z)$ is known, we can estimate θ by fitting this model. In the setting of case-control sampling, $r(Z)$ is a constant and does not impact estimation of θ_Z. When $r(Z)$ is a function of Z, however, failure to account for $r(Z)$ can result in bias. Estimation of $r(Z)$ is challenging, and researchers have developed many strategies for estimating θ without requiring $r(Z)$. Here, we describe two strategies

for obtaining **weights** for selection bias adjustment, and we extend these methods to incorporate selection composed of many intermediate sampling stages. In practice, we will not have W fully available, so our goal will be to *reduce* bias due to selection.

4.1 Method 1: inverse probability of selection weighting using external data

When sampling probabilities are known or estimable, inverse probability of selection weighting (IPW) can be applied to correct for selection bias. In this approach, we can estimate θ by fitting a weighted regression for $D|Z$, where each patient's data is weighted by $\omega \propto \frac{1}{P(S=1|D,W)}$. Estimation of $P(S = 1|D, W)$ for EHR data is generally difficult. However, we can borrow results from the non-probability sampling literature and leverage limited external data from the population of interest to obtain *rough* estimates.

Suppose we have individual-level data on D and W for an *external* sample of people from the *target population*. Example data sources from the US adult population include National Health and Nutrition Examination Survey (NHANES); the NCI Surveillance, Epidemiology, and End Results (SEER) program; and the US Census. We can estimate the selection probability for our *internal* EHR dataset as follows. Let S_{ext} indicate inclusion in the external data and S_{all} indicate inclusion in either the internal or external data. We suppose $P(S_{ext} = 1|D, W)$ for the external sample is *known*. When only sampling weights are available, we can estimate $P(S_{ext} = 1|D, W)$ by fitting a regression model, e.g. beta regression, for the weights in the external data (Elliot, 2009). Define $p_{jk} = P(S = j, S_{ext} = k|W, D, S_{all} = 1)$. Following Elliot (2009) and **Supporting Section A.7**,

$$P(S = 1|D, W) = P(S_{ext} = 1|D, W)\frac{p_{11} + p_{10}}{p_{11} + p_{01}}$$
$$\text{in large populations, } \approx P(S_{ext} = 1|D, W)\frac{P(S = 1|W, D, S_{all} = 1)}{1 - P(S = 1|W, D, S_{all} = 1)}. \qquad (Eq.\ 7)$$

We can estimate p_{11}, p_{01}, and p_{10} using a multinomial regression for inclusion in the external data only, internal data only, or both. In large populations where we have little or no overlap between the internal and external datasets ($p_{11} \approx 0$), we can apply the

second expression. When overlap between the two datasets is unknown, we can apply *Eq. S4*, which relies on estimated joint distributions in the external and internal data.

We may also want to incorporate multiple selection stages into the modeling of the aggregate selection mechanism $P(S = 1|D, W)$, as in Haneuse and Daniels (2016). In **Supporting Section A.9**, we re-write $P(S = 1|D, W)$ as a product of selection stage models, and we propose a patchwork strategy for accounting for each selection stage using the available information, which may be individual-level data or just summary statistics.

Eq. 7 assumes $P(S_{ext} = 1|D, W)$ is defined using *our* target population. In **Supporting Section A.7.4**, we explore the more general setting where the external data population (population A) and our target population (population B) are different. We show that we can apply *Eq. 7* if the joint distributions of D and W are the same in the two populations. Furthermore, if population B is a subset of population A, we can estimate selection into population B if we have additional information about population *differences* as follows:

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

where $P(P_B = 1|D, W, P_A = 1)$ is the proportion of people in population A that are also in population B as a function of D and W.

4.2 Method 2: calibration weighting using external summary statistics

Calibration weighting uses *summary statistics* from the population to re-weight the internal data so that the weighted distributions match the population. Several versions of calibration weighting exist. We will focus on two types: (1) poststratification, where the joint distribution of D and W is available, and (2) raking, where only marginal distributions are available. Under poststratification, we define weights $\omega \propto \frac{f(W,D)}{f(W,D|S=1)}$, which incorporates summary information from the population along with estimated relationships from the EHR data. Construction of raking weights involves an iterative algorithm to produce weights ω that recover the marginal distributions in the population.

5 Jointly addressing phenotype misclassification and patient selection

When $c(Z)$ and selection weights ω are known, adjustment for both sources of bias is a simple extension of **Section 3** to incorporate weighting. However, sensitivity and weights ω will rarely be known, and estimation of these quantities simultaneously is difficult. Firstly, misclassification complicates the estimation of selection weights, and **Section 4** cannot be applied directly. Secondly, sensitivity estimates in **Section 3** often rely on differences between observed and population disease rates, which will be impacted by selection. Each source of bias complicates estimation of terms used to correct for the other source of bias, and additional thought is needed. As shown in **Figure 2**, we propose a series of intermediate steps through which these quantities can be estimated. Fixing these quantities, we then describe how we can estimate θ following *Eq. 4*.

[Figure 2 about here.]

Step 1: Estimating the marginal sampling ratio

We first specify the marginal sampling ratio, $\widetilde{r}$. This can be treated as a fixed hyperparameter. We can use the data, known $\widetilde{b}$, the population disease rate $P(D=1)$, and our prior beliefs about $\widetilde{c}$ to inform plausible $\widetilde{r}$ as follows (**Supporting Section A.8**):

$$\widetilde{r} = \frac{P(D^*=1|S=1) - [1-\widetilde{b}]}{\widetilde{c} - P(D^*=1|S=1)} \frac{1-P(D=1)}{P(D=1)}. \qquad (Eq.\ 8)$$

Step 2: Estimating marginal or patient-specific sensitivities

Given $\widetilde{r}$, we estimate either marginal or patient-varying sensitivity. Using $P(D=1|S=1) = \frac{\widetilde{r}P(D=1)}{\widetilde{r}P(D=1)+P(D=0)}$, we can estimate $\widetilde{c}$ using $\widetilde{c} = \frac{P(D^*=1|S=1)-[1-\widetilde{b}]P(D=0|S=1)}{P(D=1|S=1)}$, noting that this could give $\widetilde{c} > 1$. We can estimate $c_{true}(X)$ using the approximate relationship:

$$\log\left[\frac{P(D^*=1|X,S=1) - \{1-\widetilde{b}\}P(D=0|X,S=1)}{\widetilde{b}P(D=1|X,S=1) + \{1-\widetilde{b}\} - P(D^*=1|X,S=1)}\right] \approx \beta_0 + \beta_X X, \qquad (Eq.\ 9)$$

where $P(D=1|X,S=1)$ is replaced with $\frac{\widetilde{r}P(D=1|X)}{\widetilde{r}P(D=1|X)+P(D=0|X)}$ (**Supporting Section A.5**). *Eq. 9* may have no solution for $P(D=1|X,S=1)$ incompatible with the data,

and we may directly estimate $c_{true}(X) \approx \min\left(\frac{P(D^*=1|X,S=1)-[1-\widetilde{b}]P(D=0|X,S=1)}{P(D=1|X,S=1)}, 1\right)$, where $P(D^*=1|X,S=1)$ is estimated using the EHR data.

Step 3: Estimating sampling or calibration weights

Given $\widetilde{c}$ or $c_{true}(X)$, we can estimate weights ω for selection bias adjustment. Since D is not available due to misclassification, we propose defining inverse probability weights using $P(S=1|W,D^*)$, with W replaced with available data in practice. We have that

$$P(S=1|W,D^*) = \frac{f(D^*|S=1,W)}{f(D^*|W)}P(S=1|W). \qquad (Eq.\ 10)$$

$P(D^*=1|S=1,W)$ can be directly estimated using the internal data, and we can estimate $P(S=1|W)$ using *Eq. 7* or *Eq. S4*. Combining sensitivity $c=\widetilde{c}$ or $c_{true}(X)$ with estimated $P(D=1|W)$ from the external data, we approximate $P(D^*=1|W) \approx cP(D=1|W)+[1-\widetilde{b}]P(D=0|W)$. We can define poststratification weights as $\omega \propto \frac{f(D^*,W)}{f(D^*,W|S=1)}$.

Step 4: Estimating θ given sampling/calibration weights ω and sensitivity

5.1 Method 1: approximation of $D^*|X$ accounting for selection

Suppose we assume $r(Z)=\widetilde{r}$. This may be reasonable if $W^\dagger \perp Z|D$ and the covariates of interest in Z are not in W. Suppose further that $X \perp Z|D$. Then, *Eq. 5* can be used to correct for both sources of bias (**Supporting Section A.3**). Intuitively, the impact of the selection enters that estimator through the observed rate of disease in the sample. In general, $r(Z)$ may rarely be constant, and we have $\theta_Z \approx \theta_Z^{\omega,uc}\left[\frac{\{\widetilde{c}-(1-\widetilde{b})\}\{1-p^*\}p^*}{\{\widetilde{c}-p^*\}\{p^*-(1-\widetilde{b})\}}\right]$, where $\theta_Z^{\omega,uc}$ is estimated from fitting a model for $D^*|Z$ on the *sampled* patients and *weighting* by ω, and p^* is the $\omega-$*weighted* average of D^* in our sample (**Supporting Section A.3**).

5.2 Method 2: non-logistic link function method with weighting

Suppose $c(Z)$ is a function of Z. We again remove the contribution of $r(Z)$ in *Eq. 4* by weighting estimation by ω. We estimate θ by fitting an ω-weighted version of the model $\log\left[\frac{P(D^*=1|Z)-\{1-\widetilde{b}\}}{c(Z)-P(D^*=1|Z)}\right] = \theta_0+\theta_Z Z$ to the patients in the sample. Assuming $X^\dagger \perp Z|D$ or $X^\dagger \perp D|Z$, we can replace $c(Z)$ with estimated $c_{true}(X)$ from Step 2.

5.3 Method 3: joint estimation using observed data log-likelihood incorporating weights

When $\widetilde{b} = 1$, we can jointly estimate θ and β accounting for selection bias by maximizing a $\omega-$weighted version of the log-likelihood in **Section 3.3** through a weighted expectation-maximization algorithm. Details are available in **Supporting Section A.6**.

6 Standard error estimation for bias-corrected estimates

In **Supporting Section A.10**, we detail how to estimate standard errors for each method assuming perfect specificity. The imperfect specificity case is similar. Here, we summarize that discussion. When we ignore selection bias, variance estimation is straightforward. For the method in **Section 3.1**, we can estimate standard errors for $\hat{\theta}_Z$ given $\widetilde{c}$ as $\text{Var}(\hat{\theta}_Z) \approx \text{Var}(\hat{\theta}_Z^{uc}) \left[\frac{\widetilde{c}\{1-P(D^*=1)\}}{\widetilde{c}-P(D^*=1)}\right]^2$. When we estimate θ as in **Section 3.2**, the corresponding information matrix can be inverted to obtain standard errors. We can obtain a covariance matrix for θ and β from **Section 3.3** by inverting the expected observed data information matrix. We prove that methods in **Section 3** will result in *larger* standard errors relative to uncorrected analysis on average (**Supporting Section A.10**).

Methods for selection bias adjustment in **Section 4** involve fitting a weighted regression model. Corresponding standard errors can be estimated using a Huber-White sandwich estimator as implemented in the R package *survey* (Freedman, 2006). We obtain standard errors for methods in **Sections 5.1 and 5.2** similarly. To estimate standard errors for the method in **Section 5.3**, we propose a sandwich estimator based on weighting the observed data score and information matrices as detailed in **Supporting Section A.6**.

These standard errors are mainly estimated fixing sensitivity and/or selection bias weights ω. However, rigorous standard errors should also incorporate uncertainty from estimating these quantities. To account for this residual uncertainty, we could apply bootstrap methods. We compare the impact of ignoring this uncertainty in simulations. We find that this does not impact variance estimation too much, but there may be some underestimation of variance when we ignore uncertainty in estimating selection weights.

7 Simulations

We present simulations for evaluating the proposed methods in terms of bias and standard error estimation. We divide this simulation study into three parts. In the first part, we focus on the setting with outcome misclassification and ignorable patient selection. In the second part, we focus on selection and assume we have no misclassification. After evaluating these two simpler cases, we then explore the setting with both sources of bias. Unless otherwise stated, all simulations **assume perfect specificity** ($\widetilde{b} = 1$).

In all settings, we generate 500 datasets with 5000 members and $P(D = 1) \approx 10\%$. In part 1, we impose outcome misclassification under different covariate-related sensitivity mechanisms ($\widetilde{c} \approx$ 40-50%) and different relationships between X, Z, and D. In part 2, we sub-sample about 50% of patients under different sampling mechanisms. In part 3, we sub-sample patients *and* impose misclassification, where X is related to Z given D ($\widetilde{c} \approx 65\%$). We apply our methods to correct bias in estimated θ_Z. Details about data generation and implementation can be found in **Supporting Section B.1**.

7.1 Simulation results

Figure 3 presents the biases in the estimated log-odds ratio of Z across 500 simulated datasets for the first two scenarios. **Figure 4** presents the bias for the third scenario.

Misclassification Only: Uncorrected analysis produces bias in all settings considered, with relative biases reaching 40% when $X^\dagger$ and Z are related (given D). The method in **Section 3.1** performs well in settings where X and Z are independent (so $c(Z) = \widetilde{c}$) but performs poorly when X and Z are related. The method in **Section 3.2** performs well as long as $X^\dagger$ is not related to *both* D and Z. In this case, we see some residual bias, which comes from the substitution of $c_{true}(X)$ for $c(Z)$. Still, this bias may be substantially lower than bias in uncorrected analysis. We would expect this method to perform well in all settings if the correct $c(Z)$ were known. When β_0 is fixed at a reasonable value, the method in **Section 3.3** performs well as long as $X^\dagger$ is not related to *both* Z and D.

Even when theoretically justified, the observed data log-likelihood maximization without fixed β_0 struggles, particularly when $X^\dagger$ is related to D given Z, indicating difficulty in estimating θ and β jointly without incorporating external information (e.g. $P(D=1)$).

Selection Only: Uncorrected analysis produces biases reaching 25% except when W is independent of D and is the only driver of selection. These biases can grow larger with stronger covariate effects on selection. We compare weighting strategies for correcting this bias. When the IPW model is correctly structured, we can estimate θ_Z with low bias. This is true even when the true selection probabilities are estimated using external data as in *Eq. 7*. Poststratification on W and D has good performance in all settings. Poststratification on W and raking performed similarly. These methods perform poorly when selection depends on D, and we see residual bias when selection depends on W with W related to D. This is a result of binning continuous W during weight estimation.

Both Selection and Misclassification: Bias of uncorrected analysis ranges from about 15% to 50%. Methods that *only* correct for misclassification can result in residual bias, sometimes even be larger than in uncorrected analysis (e.g. 70%). When we also account for selection, however, we see little bias for methods in **Sections 5.2 and 5.3**. The method in **Section 5.1** performs poorly since X is related to Z given D in this example.

Other Metrics for Inference: **Figure 5** provides empirical and estimated variances. Estimated variances tend to be similar to empirical variances. Ignoring uncertainty due to estimation of selection weights seems to be a bigger problem than ignoring uncertainty due to estimation of sensitivity. Coverage rates of 95% confidence intervals tend to be low (even 5%) for uncorrected analyses. In contrast, coverages tend to be near nominal for methods that fully correct bias. In **Figure B.1**, we show that misclassification bias-adjusted p-values are similar to unadjusted p-values when Z and X are independent (assuming ignorable selection). However, when Z and X are associated, the corrected and uncorrected p-values differ, and uncorrected type I error can be highly inflated.

Sensitivity to $P(D=1|X)$ and $\widetilde{r}$: In **Web Appendices B.3 and B.7**, we explore

the sensitivity of estimated $c_{true}(X)$ to the choice of $P(D = 1|X)$ and $\widetilde{r}$. Misspecification of $P(D = 1|X)$ or $\widetilde{r}$ can adversely impact sensitivity estimation, but the impact on estimated θ_Z tends to be small relative to bias from uncorrected analysis.

Imperfect specificity: In **Figure B.8**, we demonstrate that incorrectly assuming perfect specificity can result in residual bias, which can sometimes even be greater than bias from uncorrected analysis. Our methods in **Sections 3.1 and 3.2** correct this bias when $\widetilde{b}$ is known and the method's assumptions about $c(Z)$ hold. We might expect the impact of ignoring over-reporting to be a function of both $\widetilde{b}$ and $\widetilde{c}$.

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

8 Illustrative example: correcting for imposed misclassification in MGI

The Michigan Genomics Initiative (MGI) is an EHR-linked biorepository containing $> 40,000$ patients with International Classification of Disease diagnosis information (Fritsche et al., 2018). We define a binary cancer phenotype based on whether each MGI patient ever received any cancer diagnosis code. We view this phenotype as true D and study the relationship between D and gender (Z). We **impose misclassification** (generate D^*) under different covariate-related sensitivities corresponding to different relationships between X and Z ($\widetilde{c} \approx 70\%$) assuming perfect specificity. In **Figure 6**, we apply **Section 3** to correct bias in the gender odds ratio. In all settings, bias is evident in uncorrected analysis and is strong when X is related to gender. The method in **Section 3.1** performs poorly unless X is independent of Z. When sensitivity depends on smoking (related to gender and disease), assumptions for all methods are violated and residual bias is seen. Sensitivity depending directly on gender creates estimation difficulty and resulting bias. For the method in **Section 3.2**, this bias goes away when $c(Z)$ is known.

[Figure 6 about here.]

9 Discussion

Data analyses using electronic health records (EHR) data are susceptible to bias, which can negatively impact the accuracy and generalizability of statistical inference. In this paper, we focus on two common sources of bias: (1) misclassification of derived disease variables and (2) lack of representativeness. To address these key problems, we propose a variety of bias-correction strategies. We derive valid standard errors and provide an R package, *SAMBA*. A key advancement is the development of strategies to handle covariate-related misclassification. Our methods leverage each patient's follow-up history and external disease information to estimate the rate of misclassification *without requiring gold standard disease status labels*. We also explore strategies for dealing with the harder problem of selection bias. Correction for selection bias is extremely difficult for EHR data, and we describe how we can apply weighting methods in the survey sampling literature to at least partially address selection. As in Haneuse and Daniels (2016), our methods can accommodate multi-stage selection often present for EHR data, but our methods further bridge the gap between patients that are and are not included in the EHR. A key limitation of these methods is the need for high-quality external information, including external summary statistics or individual-level data from the population of interest.

Among the methods for handling misclassification, the method in **Section 5.2** is particularly attractive and easy to implement. Estimating sensitivity under that method requires some external summary information, but simulations demonstrate good performance even with imperfect summary information and under some assumption violations (**Supporting Section B.3** and **Figure 3a**). When only disease prevalence is available, the method in **Section 5.3** with fixed sensitivity model intercept can perform well, but it can be more sensitive to assumption violations, so these assumptions must be carefully considered. Poststratification emerges as an appealing approach for handling selection bias since it relies on summary statistics from the population rather than individual-level

data. We recommend the combination of poststratification and the method in **Section 5.2** as a starting point for analysts interested in applying these methods.

Simulations assume perfect specificity, but **Sections 3.1 and 3.2** can also be applied when specificity is a known constant less than 1. Disease model estimates may be sensitive to these specificity assumptions. In general, sensitivity estimation could be improved by incorporating external validation data when available. Throughout, we assume D^* is binary, and we explore non-binary phenotyping in **Supporting Section A.11**. We focus our attention on a single disease D and adjustment factors Z, but these methods could be applied to study many disease-covariate combinations. Automation strategies are discussed in **Supporting Section C.2**. Overall, this paper provides useful strategies and software for handling outcome misclassification and selection bias in EHR data analysis.

Acknowledgments

We thank Drs. Brummett, Abecasis, and Kheterpal along with many collaborators and staff at MGI and MGI participants for donating their biosamples for research. This work was supported by The University of Michigan Comprehensive Cancer Center core grant supplement 5P30-CA-046592, NSF DMS 1712933 and The University of Michigan precision health award U067541. We thank Alexander Rix for his work developing *SAMBA*.

Data Availability Statement

Michigan Genomics Initiative data are available after institutional review board approval to select researchers. See `https://precisionhealth.umich.edu/our-research/michigangenomics/` for details.

References

Baker, R., Brick, J. M., Bates, N. A., et al. (2013). Report of the AAPOR Task Force on Non-Probability Sampling. Technical report.

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

Beesley, L. J., Salvatore, M., Fritsche, L. G., et al. (2019). The Emerging Landscape of Epidemiological Research Based on Biobanks Linked to Electronic Health Records. *Statistics in Medicine* **39,** 773–800.

Bower, J. K., Patel, S., Rudy, J. E., and Felix, A. S. (2017). Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. *Curr Epidemiol Rep* **4,** 346–352.

Dahabreh, I. J. and Hernán, M. A. (2019). Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology* **34,** 719–722.

Duffy, S. W., Warwick, J., Williams, A. R., et al. (2004). A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology and Community Health* **58,** 712–717.

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

Freedman, D. A. (2006). On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician* **60,** 299–302.

Fritsche, L. G., Gruber, S. B., Wu, Z., et al. (2018). Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *The American Journal of Human Genetics* **102,** 1–14.

Goldstein, B. A., Bhavsar, N. A., Phelan, M., and Pencina, M. J. (2016). Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology* **184,** 847–855.

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

Hubbard, R. A., Benjamin-Johnson, R., Onega, T., Smith-Bindman, R., Zhu, W., and Fenton, J. J. (2015). Classification accuracy of claims-based methods for identifying providers failing to meet performance targets. *Statistics in Medicine* **34,** 93–105.

Lange, J. M., Hubbard, R. A., Inoue, L. Y., and Minin, V. N. (2015). A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics* **71,** 90–101.

Meng, W., Adams, M. J., Hebert, H. L., Deary, I. J., McIntosh, A. M., and Smith, B. H. (2018). A Genome-Wide Association Study Finds Genetic Associations with Broadly-Defined Headache in UK Biobank (N = 223,773). *EBioMedicine* **28,** 180–186.

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

O'Malley, K. J. O., Cook, K. F., Price, M. D., Wildes, R., Hurdle, J. F., and Ashton, C. M. (2005). Measuring Diagnoses: ICD Code Accuracy. *Health Services Research* **40,** 1620–1639.

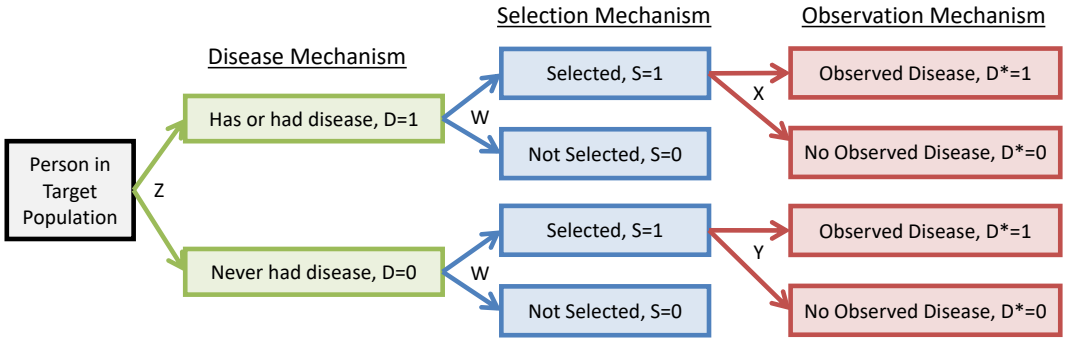
Phelan, M., Bhavsar, N., and Goldstein, B. A. (2017). Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference. *eGEMs* **5,** 22.

Sinnott, J. A., Dai, W., Liao, K. P., et al. (2014). Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics* **133,** 1369–1382.

Supporting Information

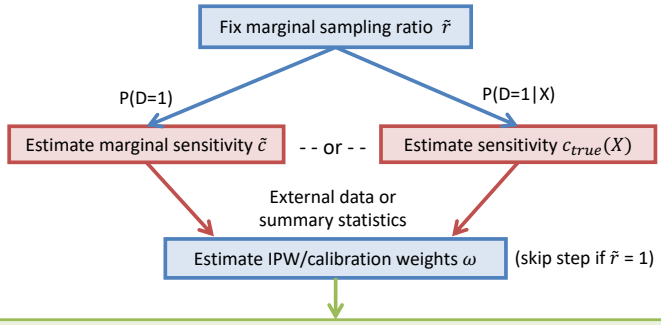
Web appendices referenced in Sections 2-7 and 9 are available with this paper at the *Biometrics* website on Wiley Online Library. A vignette demonstrating use of the R package *SAMBA* (available at `https://github.com/umich-cphds/SAMBA`) is also provided.

Figure 1: Diagram of the assumed data structure*



*This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

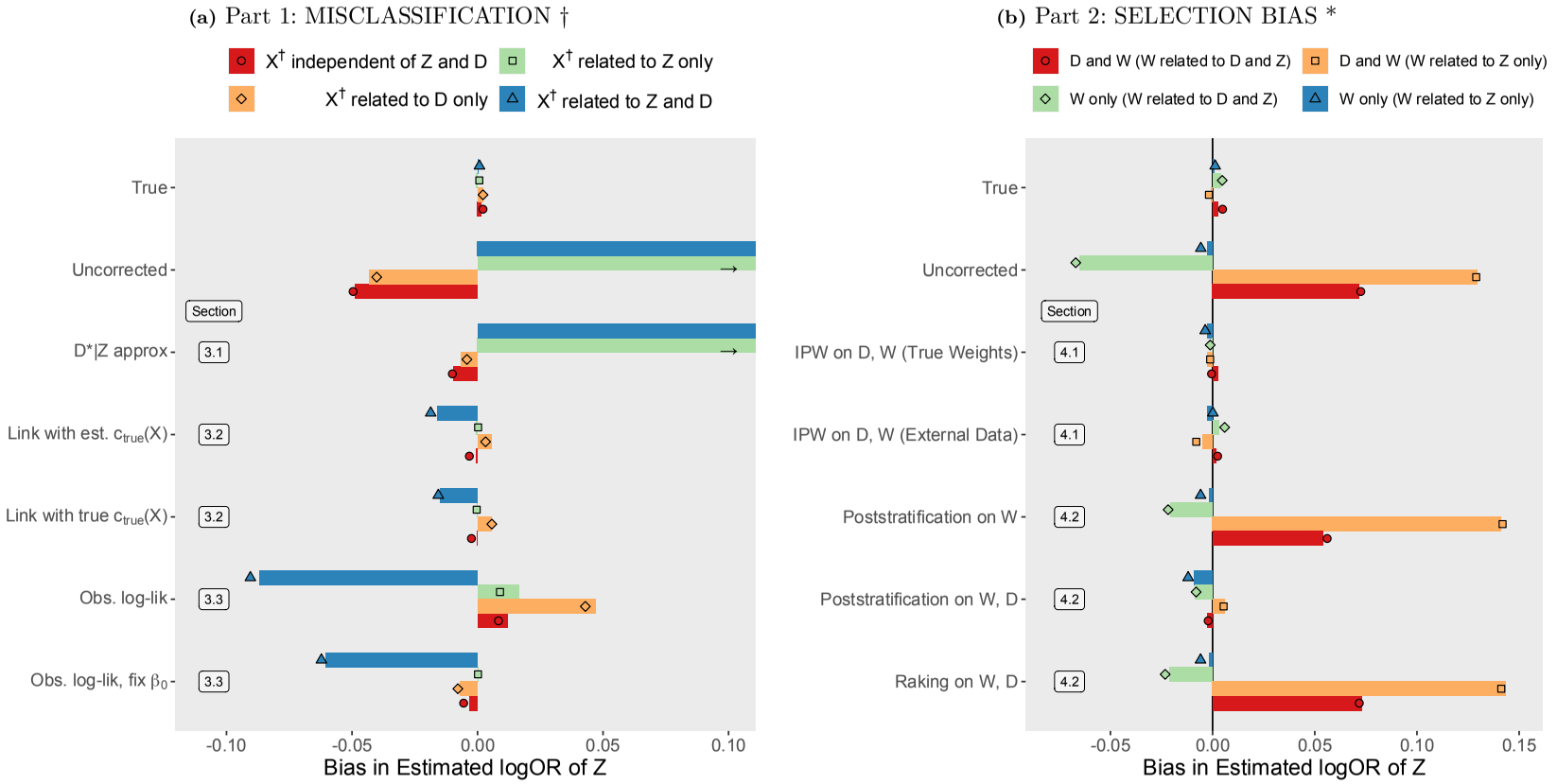
Figure 2: Flowchart of data analysis accounting for both misclassification and patient selection*



Estimate disease model parameters			
METHOD	Approx. D*|Z	Non-Logistic Link Method	Observed Data Log-Likelihood
SECTIONS	3.1 and 5.1	3.2 and 5.2	3.3 and 5.3
SPECIFICITY	Known constant $\tilde{b}$	Known constant $\tilde{b}$	Perfect ($\tilde{b} = 1$)
ASSUME	$X \perp Z \mid D$	$X^\dagger \perp Z \mid D$ or $X^\dagger \perp D \mid Z$ (neither if $c(Z)$ known)	$X^\dagger \perp Z \mid D$ or $X^\dagger \perp D \mid Z$

* The notation $u \perp v|w$ corresponds to conditional independence between random variables (or sets of random variables) u and v given w. Labels along arrows correspond to external information used in the estimation. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

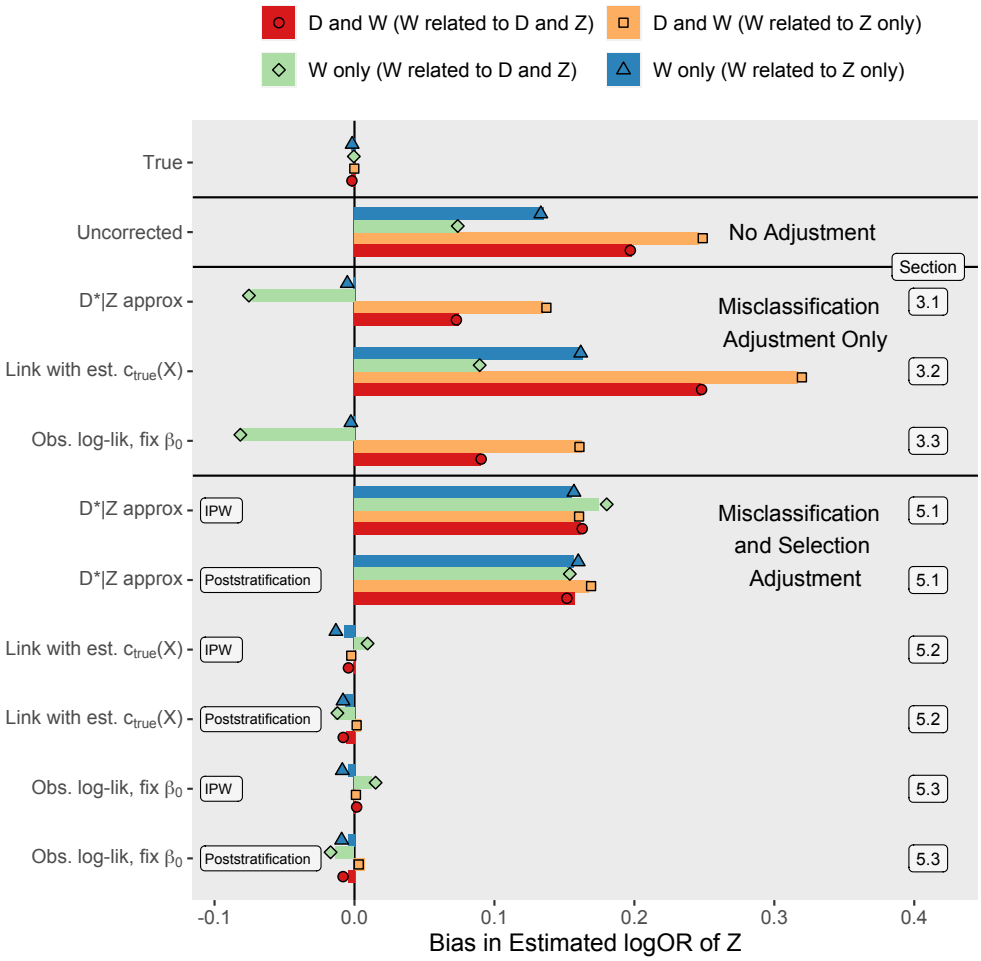
Figure 3: Bias in estimated log-odds ratio of Z across 500 simulations under selection bias *or* phenotype misclassification. Bars (points) represent the average (median) difference between estimates and the truth of $\theta_Z = 0.5$.



† The "$D^*|Z$ approx." method uses estimated $\widetilde{c}$ assuming true $P(D = 1)$ is known. "Link" represents the non-logistic GLM fit using either true or estimated $c_{true}(X)$ in place of $c(Z)$. $c_{true}(X)$ was estimated assuming $P(D = 1|X)$ was known. For the observed log-likelihood method with fixed β_0, β_0 was set to estimated logit($\widetilde{c}$). In all simulations, we assume $\widetilde{b} = 1$ is known.

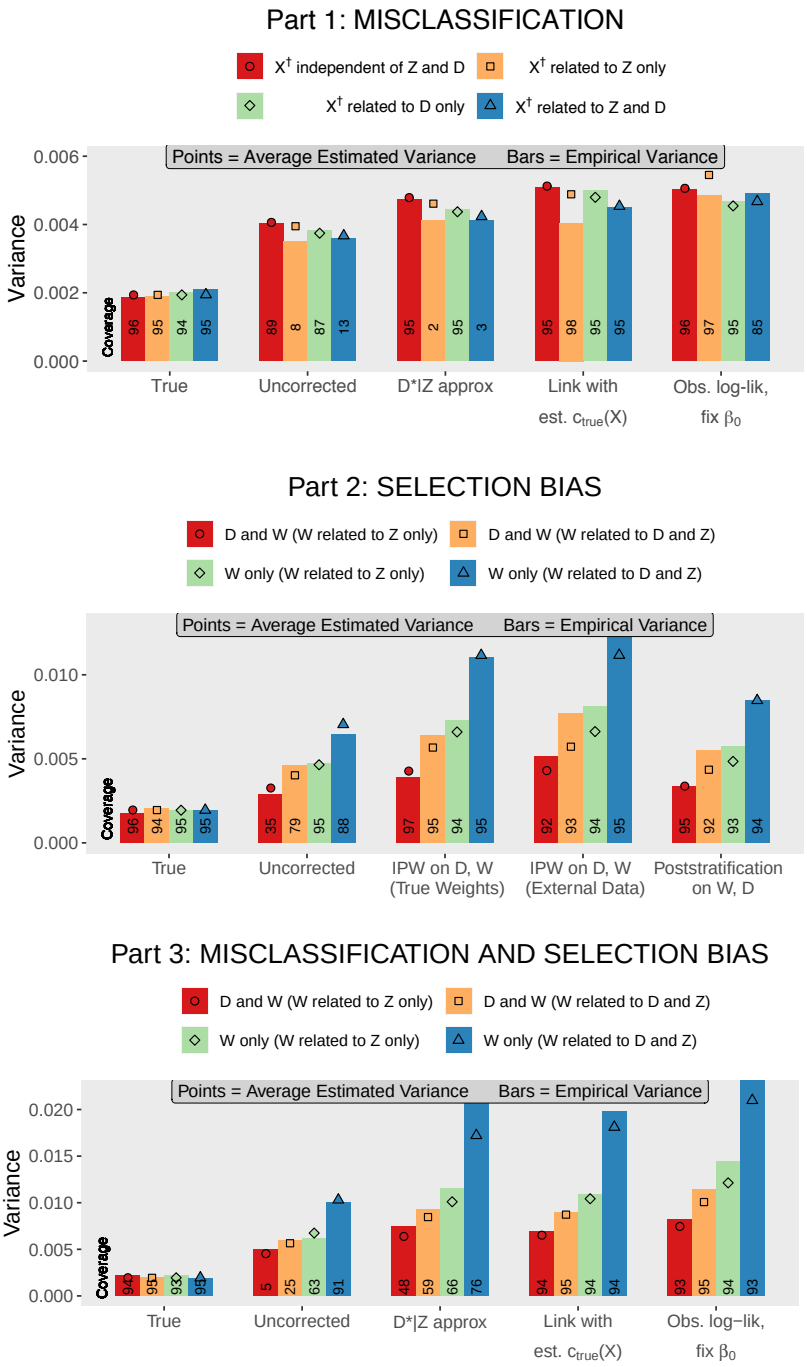
* Labels correspond to variables included in the selection model and associations between variables. "True Weights" indicates weighting using the true selection model. "External Data" indicates weights estimated using *Eq. 7*. For all calibration weighting (poststratification and raking), W was binned into intervals of roughly 0.5. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Figure 4: (Part 3) Bias in estimated log-odds ratio of Z across 500 simulations under selection bias *and* phenotype misclassification.† ** Bars (points) represent the average (median) difference between estimates and the truth of $\theta_Z = 0.5$.



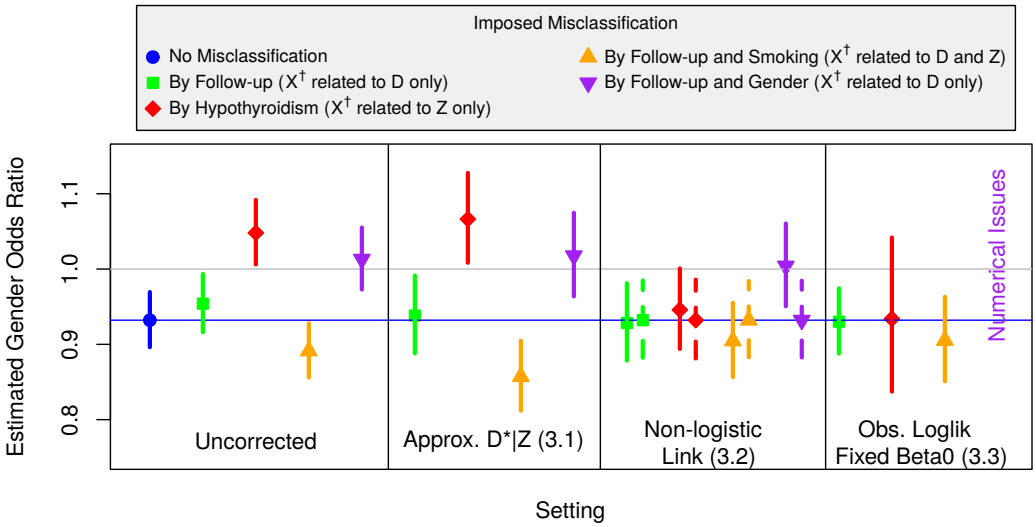
† The "$D^*|Z$ approx." method uses estimated $\widetilde{c}$ assuming true $P(D = 1)$ is known. "Link" represents the non-logistic GLM fit using estimated $c_{true}(X)$ in place of $c(Z)$ assuming $P(D = 1|X)$ was known. For the observed log-likelihood method, β_0 was set to estimated logit($\widetilde{c}$). For all methods, $\widetilde{c}$ and $c_{true}(X)$ were estimated assuming true $\widetilde{r}$ was known and true $\widetilde{b} = 1$. ** Labels correspond to variables included in the selection model and associations between variables. IPW was implemented using the *true* selection probabilities. We obtain similar results using *estimated* probabilities. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Figure 5: Comparison of empirical and median estimated variances for the log-odds ratio of Z across 500 simulations†



† The "$D^*|Z$ approx." method uses estimated $\widetilde{c}$ assuming true $P(D = 1)$ is known. "Link" represents the non-logistic GLM fit using estimated $c_{true}(X)$ in place of $c(Z)$ and assuming $P(D = 1|X)$ was known. For the observed log-likelihood method, β_0 was set to estimated logit($\widetilde{c}$). In Part 2, *True Weights* indicates weighting using the true selection model and "External Data" indicates weights estimated using *Eq. 7*. For poststratification, W was binned into intervals of roughly 0.5. Labels correspond to variables included in the selection model and associations between variables. In Part 3, $\widetilde{c}$ and $c_{true}(X)$ were estimated assuming true $\widetilde{r}$ was known. For Part 3, we show results for IPW weighting using the *true* selection probabilities. In all simulations, we assume $\widetilde{b} = 1$ is known. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Figure 6: 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. Dashed lines indicate use of the true sensitivity, $c(Z)$. Methods from **Sections 3.1-3.3** are applied with known $\widetilde{b} = 1$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.