Received: 27 October 2021 | Revised: 12 August 2022 | Accepted: 13 August 2022

DOI: 10.1002/sim.9579

RESEARCH ARTICLE

Statistics in Medicine WILEY

Case studies in bias reduction and inference for electronic health record data with selection bias and phenotype misclassification

Lauren J. Beesley[1,2] | **Bhramar Mukherjee**[1]

[1]Department of Biostatistics, University of Michigan, Michigan, USA

[2]Information Systems and Modeling, Los Alamos National Laboratory, New Mexico, USA

Correspondence
Lauren J. Beesley, Department of Biostatistics, University of Michigan, Michigan, USA.
Email: lvandervort@lanl.gov

Funding information
Comprehensive Cancer Center, University of Michigan, Grant/Award Number: 5P30-CA-046592; Division of Mathematical Sciences, Grant/Award Number: 1712933; Laboratory Directed Research and Development, Grant/Award Numbers: 20210761PRD1, LA-UR-22-23875; University of Michigan Precision Health, Grant/Award Number: U067541; National Nuclear Security Administration of the U.S. Department of Energy, Grant/Award Number: 89233218CNA000001

Electronic health records (EHR) are not designed for population-based research, but they provide easy and quick access to longitudinal health information for a large number of individuals. Many statistical methods have been proposed to account for selection bias, missing data, phenotyping errors, or other problems that arise in EHR data analysis. However, addressing multiple sources of bias simultaneously is challenging. We developed a methodological framework (R package, *SAMBA*) for jointly handling both selection bias and phenotype misclassification in the EHR setting that leverages external data sources. These methods assume factors related to selection and misclassification are fully observed, but these factors may be poorly understood and partially observed in practice. As a follow-up to the methodological work, we demonstrate how to apply these methods for two real-world case studies, and we evaluate their performance. In both examples, we use individual patient-level data collected through the University of Michigan Health System and various external population-based data sources. In case study (a), we explore the impact of these methods on estimated associations between gender and cancer diagnosis. In case study (b), we compare corrected associations between previously identified genetic loci and age-related macular degeneration with gold standard external summary estimates. These case studies illustrate how to utilize diverse auxiliary information to achieve less biased inference in EHR-based research.

KEYWORDS
electronic health records, inverse probability weighting, Michigan Genomics Initiative, NHANES, non-probability sampling, poststratification, SEER

1 | INTRODUCTION

Electronic health record (EHR) databases allow researchers to study many diseases across patients' course of medical care. However, observational databases such as EHR present many practical challenges for health research, which can negatively impact internal validity and external generalizability of resulting inference. Some issues include poorly measured variables, missing data, confounding, and limited information about patient recruitment mechanisms. Analytical

and design-based strategies for addressing these data limitations are key to obtaining improved inference based on EHR data.

In this article, we focus on two potential sources of bias for EHR analyses: (1) lack of representativeness of the EHR sample relative to some target population (selection bias) and (2) measurement error in EHR-derived outcome/phenotype variables. In particular, we consider the common setting where researchers are interested in using EHR data to study the relationship between a binary disease phenotype, D, to a set of patient characteristics, Z. Selection bias can occur due to lack of representativeness of a defined target population such as the US adult population, and the EHR-derived disease phenotyped may be misclassified for some patients. Many researchers have proposed methods to correct for *either* selection or measurement error in the EHR setting, this prior scholarship does not address how to handle both sources of bias in a single analysis.[1-4] More recently, Beesley and Mukherjee (2022) proposed novel methodology that leverages population summary statistics and internal data (eg, patient visit patterns) to address both sources of bias simultaneously in a single data analysis.[5] As discussed later on, several existing strategies for handling either selection bias for non-probability samples or phenotype misclassification can be viewed as special cases of the methods in Beesley and Mukherjee.[1,3,4,6] In the ideal setting where variables related to selection (collectively, denoted W) and phenotype misclassification (denoted X) are known and observed, Beesley and Mukherjee demonstrates good bias reduction and inferential performance for the proposed class of methods.

In reality, however, drivers of selection and misclassification may not be known, and known drivers may not always be observed (eg, income, residential information, access to health care). Additionally, the methods for handling selection bias proposed in Elliot 2009 and extended in Beesley and Mukherjee rely on access to high-quality external data (or summary statistics) on D and W from the target population (or a probability sample).[6] The availability of such external data will depend on the target population, the outcome of interest, and the complexity of W. External individual-level data may present additional challenges such as missing data, or we may have access to marginal distributions of variables D and some subset of W for the target population but not their joint distribution. Our ability to correct bias will naturally be limited by the observed internal data and external information we have available. Implementation of these methods in messy real-life data analysis is not trivial, and good performance is not guaranteed by proven theoretical results. It is of interest, therefore, to explore how the methods in Beesley and Mukherjee (and by extension, earlier methods in Sinnott et al, Duffy et al, and Elliot) perform for some real-world inferential problems. Our goal in this paper is to provide a general road map for researchers interested in applying these types of bias correction methods in their own data analyses, which is non-trivial even if the theory and the software package exist.

In this article, we demonstrate how the methods discussed in Beesley and Mukherjee can be applied in practical EHR data analysis through two case studies.[5] In doing so, we explore the potential for bias reduction in practice and highlight some limitations. We consider data from the Michigan Genomics Initiative (MGI), a longitudinal EHR and genotype-linked biorepository within The University of Michigan health system. In case study (a), we examine the relationship between cancer diagnosis and gender, accounting for the *strong* enrichment of cancer patients due to ascertainment mechanisms in MGI. This case study addresses bias by leveraging cancer prevalences by age from SEER (Surveillance, Epidemiology, and End Results program by the National Cancer Institute), age distributions from the US Census, and individual-level data from NHANES (National Health and Nutrition Examination Survey by the US Centers for Disease Control and Prevention [CDC]). In case study (b), we consider the relationship between age-related macular degeneration (AMD) and several genetic loci identified by a large population-based genome-wide association study, and summary statistics for disease prevalence by age from the US CDC are used for bias reduction. Comparative gold standard results from International AMD Genomics Consortium data are available for benchmarking different bias reduction approaches.

In Section 2, we introduce the observational database with individual-level EHR data used for our analysis. Section 3 provides an overview of the bias-correction strategies proposed in Beesley and Mukherjee.[5] In Sections 4-5, we apply these methods to obtain corrected point estimates and standard errors for the two case studies. We conclude with a discussion in Section 6. Through this paper, our goal is to illustrate a valuable set of tools in the EHR data analysis toolkit and highlight important considerations to facilitate their implementation.

2 | CASE STUDIES AND THE MICHIGAN GENOMICS INITIATIVE DATABASE

The Michigan Genomics Initiative (MGI) is an EHR-linked biorepository within Michigan Medicine containing over 75 000 patients with both genotype and phenotype information available.[7] Time-stamped ICD (International

Classification of Disease) diagnosis data are available for each patient. A rich ecosystem of additional information is available for each patient, including lifestyle and behavioral risk factors, lab and medication data, geocoded residential information, socioeconomic metrics, and other patient-level, census tract-level, and provider-level characteristics.

We want to use MGI data to study the association between disease status D and predictors Z and generalize to the target US adult population or a subset. Here, the predictor set Z may also contain additional adjustment factors that are not of primary interest. The process by which data are accumulated and the systematic differences between MGI patients and our target population must be considered to achieve this goal. Supplementary Material Figure A.1 provides a visualization of mechanisms by which patients are included in MGI. Patients are recruited among perioperative patients seen at Michigan Medicine, with targeted recruitment primarily through the Department of Anesthesiology. This naturally results in strong enrichment for diseases associated with surgical intervention, such as cancer.[8]

We illustrate how we can address this lack of representativeness relative to the US adult population through two case studies. Case study (a) explores the relationship between gender and cancer diagnosis. Given the strong enrichment for cancer in MGI, the method for handling selection bias for this case study may have a strong impact on resulting inference. In case study (b), we investigate the relationship between age-related macular degeneration (AMD) diagnosis and 43 genetic loci previously identified as risk factors for AMD. We expect AMD diagnosis to be weakly associated with inclusion in MGI after adjusting for age and other comorbidities, and we may expect that our choices regarding handling of selection bias may be less impactful. These case studies are summarized in Table 1.

In case studies (a) and (b), we consider a subset of MGI participants (enrolled 2012 and later). We first characterize some differences between this MGI dataset and our target population. We define the target population as all US adults for case (a) and as US adults aged 50+ of recent European ancestry for case (b). Using these data, we define observed disease variables for several phenotypes of interest (cancer, macular degeneration, coronary artery disease, and diabetes) based on whether or not patients received particular diagnosis codes during follow-up in the Michigan Medicine EHR. Supplementary Material Table A.1 provides descriptives for the patients used in our analysis and compares these in parallel to available summary statistics from the US adult population. Supplementary Material Table A.2 details the sources used to obtain these population summary statistics. We also provide descriptives for adults interviewed and examined for NHANES in 2017-2018, which represents an external probability sample from the US adult population. We generally find that MGI patients tend to be older and have a greater burden of disease compared to patients in NHANES and the US adult population. This is expected in a hospital-based perioperative cohort. In modeling disease risk, therefore, we need to carefully address potential relationships between patient characteristics (W) and inclusion in MGI if we want to use these results to make inference about the US adult population. We will address selection bias by leveraging external summary statistics and also some individual-level data from NHANES.

We are also concerned about the potential for bias due to misclassification in our EHR-derived phenotypes. Of the MGI patients considered, nearly 10% of patients were seen for less than 6 months and nearly 9% were seen for fewer than 10 visits. We may have little confidence in saying a person does not have a given disease if they were seen for very few visits or a very short window of time. Instead, we may have just missed the disease. Therefore, misclassification of our derived phenotypes is a strong concern, particularly given the short follow-up and small number of visits for some MGI patients.

3 | BRIEF OVERVIEW OF METHODS

Notation

In this section, we summarize some of the methods presented in Beesley and Mukherjee and implemented in this paper.[5] Let binary D represent a patient's true disease status and suppose we are interested in the relationship between D and person-level information, Z. We call this the *disease model*. Let D^* denote the EHR-derived disease phenotype, which we will assume is binary. D^* is a potentially misclassified version of D with corresponding sensitivity and specificity. Unless otherwise noted, we will assume specificity = 1, so D^* is assumed to be misclassified only through missed diseases. We call the mechanism generating D^* given $D = 1$ the sensitivity model and let X denote patient and provider-level predictors related to sensitivity. For example, X may contain factors such as patient age, length of follow-up, and number of hospital visits. We suppose we model both $D|Z$ and $D^*|X, D = 1$ using logistic regressions as follows:

$$\begin{aligned} \text{Disease model}: \quad & \text{logit}(P(D = 1|Z;\theta)) = \theta_0 + \theta_Z Z, \\ \text{Sensitivity model}: \quad & \text{logit}(P(D^* = 1|D = 1, S = 1, X;\beta)) = \beta_0 + \beta_X X, \end{aligned} \tag{1}$$

TABLE 1 Descriptions of two case studies

	Case study (a): Cancer and gender	Case study (b): Age-related macular degeneration and SNPs
Data sources		
Internal data	Michigan Genomics Initiative	Unrelated Michigan Genomics Initiative patients aged 50+ of recent European descent
External data	US 2000 Census, SEER 1975-2018, NHANES 2017-18[a]	US 2000 Census, NIH National Eye Institute 2010[a]
Goal of analysis		
Estimand	Association between cancer and gender	Association between age-related macular degeneration and genetic loci
Target population	All adults in US	US adults aged 50+ of recent European descent
Gold standard θ_Z	US SEER[a] prevalences	IAMDGC GWAS[b]
Key properties	Large potential for selection bias and lower anticipated underreporting due to strong disease enrichment in MGI	Large potential for under-reporting and comparatively low anticipated selection bias after covariate adjustment
Notation and modeling		
(Internal data) n	43 339	30 041
D	Presence of cancer	Presence of AMD
D^*	Receipt of cancer code	RECEIPT of macular degeneration code
Z	Gender	Genotype (0/1/2) for 43 SNPs, age, genotype PCs 1-4, gender, batch
Cases, (% of n)	23 587 (54.4%)	1781 (5.9%)
Sensitivity model (X)	Age, number of visits, length of follow-up	Age, number of visits, length of follow-up
Selection model (W *and* D)	Age, cancer diagnosis, diabetes diagnosis, coronary artery disease diagnosis, smoking, body mass index	AGE, AMD diagnosis, cancer diagnosis, diabetes diagnosis, coronary artery disease diagnosis

[a]Details can be found in Supplementary Material Table A.2.
[b]International Age-Related Macular Degeneration Genomics Consortium.

where S is an indicator denoting inclusion in the EHR database and where the probability of inclusion is assumed to be a function of D and additional covariates, W.

Assumptions and transportability

Our interest is in using the EHR data to make inference about a *defined target population* (eg, the US adult population between ages 50-65). This population may differ from the *source population* (eg, people in the catchment area of the health system, Supplementary Material Figure A.1). When the target population contains individuals not in the source population (eg, due to study eligibility), we need to make assumptions about *transportability* between the source and target populations.[9] Transportability is a common challenge in the domain of causal inference, and we clarify that our goal is to make statements about the *association* between D and Z; we do not aim to make causal claims. In Supplementary Material Section C, we explore settings in which we have transportability for the $D|Z$ association between distinct or overlapping source and target populations, where the source population may be poorly defined. Figures 1 and 2 summarizes those assumptions. In practice, these source population assumptions are difficult to verify unless detailed information is available for the source population.

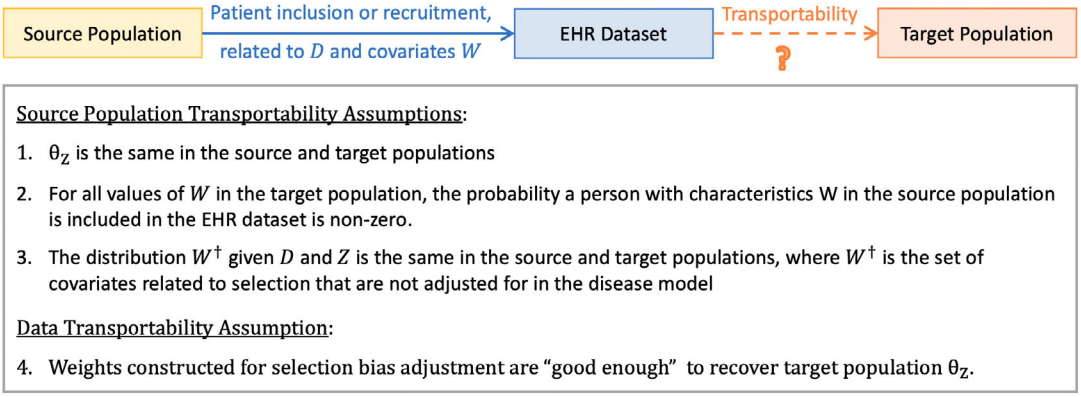

FIGURE 1 Assumptions for making inference about target population different than study source population. Additional details and motivation for these assumptions provided in Supplementary Material Section C. Diagnostics for assessing transportability in practice are discussed in Degtiar and Rose.[10] When interest is in θ_Z but not θ_0, we may ignore Assumption 3 if $W^\dagger \perp Z|D$ in both populations. This is motivated by logistic regression model properties as discussed in Neuhaus[13]

It is not enough, however, to have transportability between the source and target populations. In addition, we must have transportability for the D given Z association between our EHR data and the target population. A lack of transportability can be viewed in terms of selection bias, where we pretend that the EHR dataset is a non-probability sample directly from the target population. A common strategy for addressing selection bias is to construct weights accounting for the lack of representativeness and to perform a weighted analysis using our internal data. In doing so, we make a final transportability assumption that the constructed weights are "good enough" to recover the target population quantity of interest. The benchmark of "good enough" is intentionally vague and should be determined based on the scientific question and bias tolerance.

While source population transportability assumptions are usually difficult to verify, there are some data diagnostics available to explore the quality of the constructed selection weights.[10] One approach detailed in Degtiar and Rose[10] is compare summary statistics for D and W calculated from the weighted internal dataset with population summary statistics. In our case studies, we apply several strategies to estimate inverse probability of selection (IPW) or poststratification weights for selection bias adjustment, and we compare a variety of estimated summaries (eg, mean age) in the internal sample with values for the target population (Supplementary Material Section C.1). When an external probability sample from the target population is also available, a wider variety of diagnostics can be used to assess the reasonableness of transportability, including comparison of constructed weights or propensity scores in the internal and external data.[10,11]

Methods and estimation

Under 1 and assuming the transportability assumptions in Figure 1 hold, Beesley and Mukherjee[5] observes that

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

where $c(Z)$ and $r(Z)$ are defined as follows;

$$\begin{aligned} c_{true}(X)&=\text{expit}\left(\beta_0+\beta_X X\right) \qquad c(Z)=\int c_{true}(X)f(X^\dagger|Z,D=1,S=1)dX^\dagger,\\ r(Z)&=\frac{\int P(S=1|D=1,W;\phi)f(W^\dagger|Z,D=1)dW^\dagger}{\int P(S=1|D=0,W;\phi)f(W^\dagger|Z,D=0)dW^\dagger}, \end{aligned} \tag{2}$$

and where $X^\dagger$ and $W^\dagger$ represent the elements in X and W not included in Z, respectively. The term $c(Z)$ represents sensitivity as a function of Z. $r(Z)$ is the sampling ratio capturing the enrichment of disease in the EHR data as a function of Z, relative to the target population. These two terms will rarely be known in practice, and Beesley and Mukherjee[5] proposes

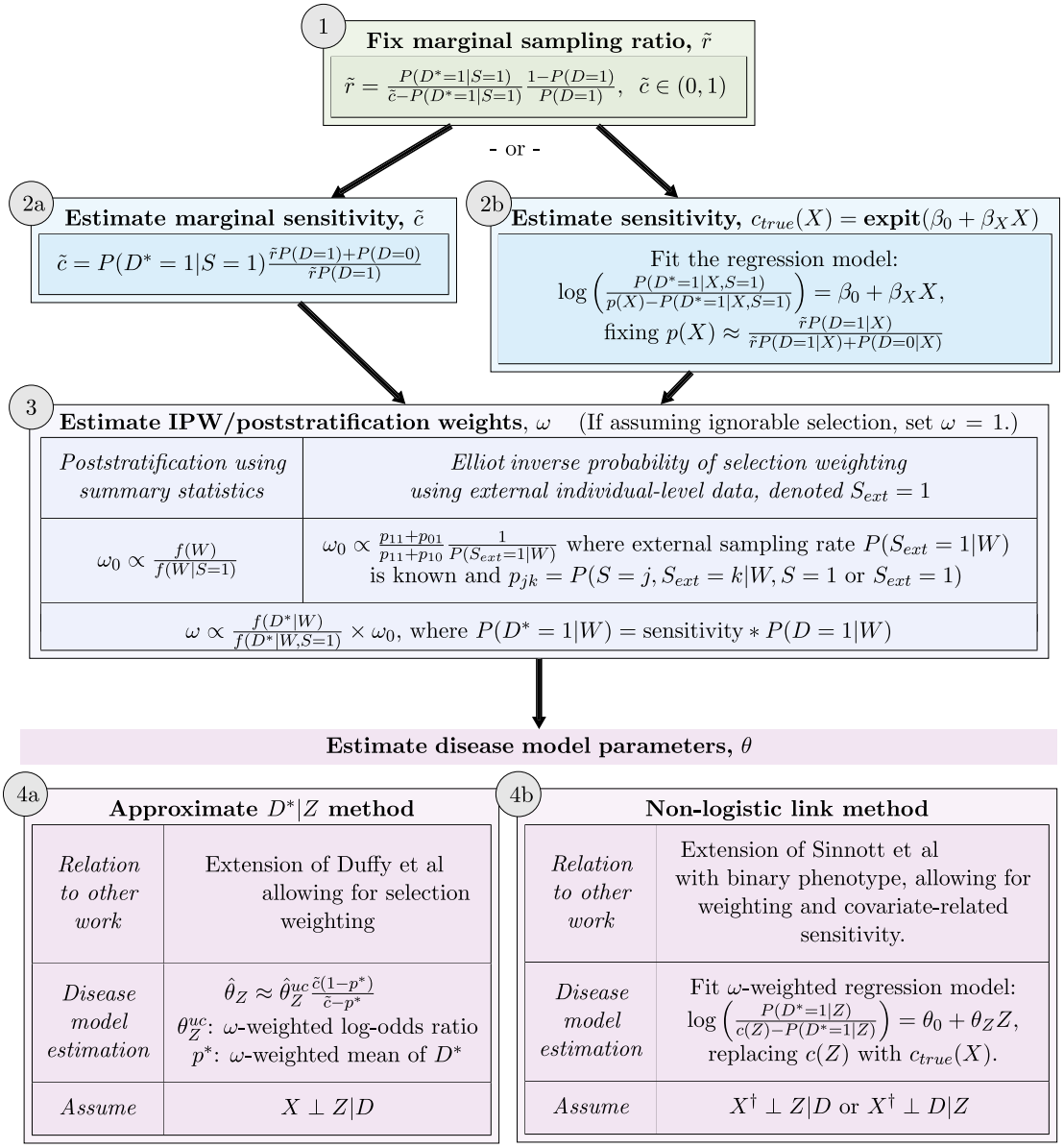

FIGURE 2 Flowchart of sensitivity and disease model parameter estimation methods. $X^\dagger$: predictors in X (sensitivity model) not included in Z (disease model). $c(Z)$: sensitivity $c_{true}(X)$ integrated over the distribution of $X^\dagger$ given $D = 1$ and Z. S_{ext}: indicator of inclusion in external probability sample. For Step ③ inverse probability of selection weighting, p_{jk} can be estimated using logistic (no patients overlap between internal/external data) or multinomial logistic (overlap between internal/external data) regression in the merged internal and external data. When only sampling weights are available for the external dataset, $P(S_{ext} = 1|W)$ can be estimated using beta regression as proposed in Elliot.[6] Standard errors for disease model parameters are obtained using Huber-White sandwich estimators as detailed in Beesley and Mukherjee and implemented in R package *SAMBA*

a multi-step strategy for estimating disease model parameters accounting for unknown $c(Z)$ and $r(Z)$ as illustrated in Figure 2. We summarize these steps as follows:

• *Step ①: Fix marginal sampling ratio.* $\tilde{r} = P(S=1|D=1)/P(S=1|D=0)$. This quantity will not be known, but we will fix it to a reasonable value. Analysis can then be repeated for multiple plausible $\tilde{r}$ values.

• *Step ②: Estimate sensitivity.* If $c(Z)$ is plausibly constant, we estimate sensitivity $\tilde{c}$ using method 2a in Figure 2. Otherwise, we approximate $c(Z)$ using an estimate of $c_{true}(X) = \text{expit}(\beta_0 + \beta_X X)$ via method 2b.

• *Step ③: Estimate weights ω for selection bias adjustment.* When we have individual-level data for D and W in a probability sample from the target population, we can estimate inverse probability of selection weights. When we have summary

statistics of D and W in the target population, we can obtain poststratification weights. In practice, W may not be available, and we will use available elements of W to *reduce* selection bias. In settings where we assume ignorable patient selection, we set $\omega = 1$.

• *Step ④: Estimate disease model parameter* θ_Z. Figure 2 describes two strategies for estimating θ_Z under different assumptions. When sensitivity is assumed constant as a function of Z, we apply a simple method approximating the distribution of $D^*|Z$ (method 4a). Method 4b (called the non-logistic link function method) allows us to estimate disease model parameters, allowing for covariate-related sensitivity. For weighted analysis, standard errors can be obtained using Huber-White sandwich estimators as detailed in Beesley and Mukherjee.[5]

4 | CASE STUDY (A): ASSOCIATION BETWEEN CANCER AND GENDER USING MGI

Goals of analysis

In the first case study, we suppose we want to use data from MGI to make inference about the relationship between cancer (D) and gender (Z) in the US adult target population. The association between cancer and gender is a convenient estimand to study, since the direction of the association is well-understood. SEER data indicate lower lifetime cancer risk (any site, available at https://seer.cancer.gov/csr/previous.html) among women relative to men, with corresponding log-odds ratios of −0.24 (2008-2010), −0.19 (2010-2012), −0.08 (2012-2014), and −0.07 (2014-2016). This known result provides us with the opportunity to benchmark different bias reduction strategies based on the direction of the resulting cancer-gender association estimates.

This case study provides an example of a setting where selection bias could be substantial and where missed diagnoses may be of less concern. MGI is enriched for cancer diagnosis (53% in MGI vs a lifetime risk of 39.5% for US adults [SEER 2017]), and factors such as age, BMI, and smoking status are expected to be related both to cancer diagnosis, gender, and selection into MGI through other comorbidities. There is a *strong* potential for selection bias, and bias-adjusted results may be highly sensitive to the adjustment method. In contrast, we expect the impact of misclassification to be comparatively small, since cancer history may be routinely recorded/reported and this EHR dataset is already highly enriched for cancer diagnoses.

In addressing potential bias due to selection and misclassification in this example, we follow the four-step procedure outlined in Figure 2 and Section 3. Table 1 provides a detailed characterization of the various assumptions made and data used in this analysis. Here, we describe how each of these estimation steps is carried out, incorporating external summary statistics from SEER and individual-level data from NHANES.

Fixing the marginal sampling ratio, $\tilde{r}$

First, we specify a value for the marginal sampling ratio, $\tilde{r} = \frac{P(S=1|D=1)}{P(S=1|D=0)}$, which we can view as a kind of tuning parameter roughly capturing the (unknown) degree of cancer enrichment in the study sample relative to the target population. We can use observed relationships in the data and known disease prevalence in the target population to explore plausible values of $\tilde{r}$ using the equation in Figure 2 as shown in Supplementary Material Figure B.1. To capture many potential scenarios for $\tilde{r}$ compatible with the data, we perform our analysis multiple times using the following fixed values of $\tilde{r}$: 1, 2, 5, 10, 25, 50, and 100. At the extremes, 1 corresponds to no outcome enrichment in the EHR sample, and 100 corresponds to very strong enrichment, with the probability of being included in the study sample being 100x for patients with cancer compared to patients without cancer.

Estimating sensitivity

Fixing $\tilde{r}$, we estimate the sensitivity of the derived cancer phenotype D^* using method 2a (assuming constant sensitivity for all patients) and then using method 2b (assuming sensitivity varies across patients). For method 2b, we define sensitivity model covariates (X) to include age, the length of EHR follow-up, and the log-number of doctor's visits per follow-up year. For method 2a, we assume $P(D = 1) = 0.395$. Method 2b requires us to specify $P(D = 1|X)$ for the target

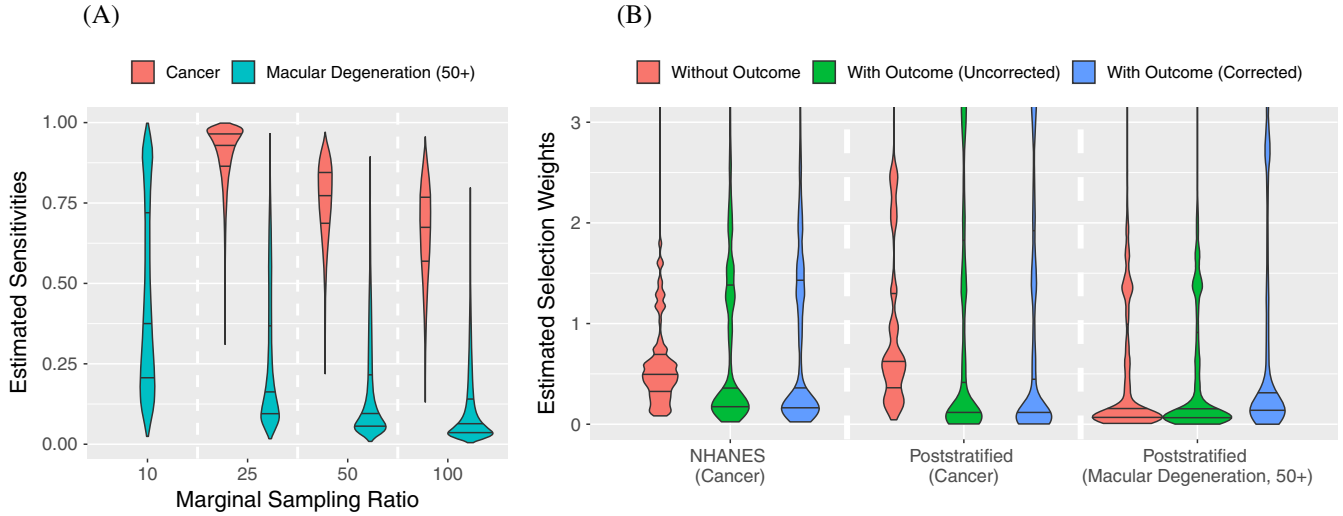

FIGURE 3 Estimated sensitivities and selection adjustment weights [Case studies (a) and (b)]. (A) Estimated sensitivities, (b) estimated weights for selection adjustment. Corrected weights using $\tilde{r} = 25$ are shown. All weights were trimmed at 10. Horizontal black lines correspond to the 25th, 50th, and 75th quantiles

population. We do not know this relationship, but we use SEER summary statistics for the relationship between age and cancer prevalence to approximate $P(D = 1|X) \approx P(D = 1|age)$ as well as we can. For some values of $\tilde{r}$, the method in Figure 2 will provide no solution. As discussed in, Beesley and Mukheree,[5] this method results in no solution when the assumed values for $\tilde{r}$ are incompatible with the observed data. Therefore, we can focus our attention on values of $\tilde{r}$ with estimable $c_{true}(X)$.[5]

Figure 3A shows the distributions of estimated $c_{true}(X)$ in MGI. Median sensitivity estimates for the cancer phenotype are between 0.93 for $\tilde{r} = 25$ to 0.68 for $\tilde{r} = 100$. Estimated sensitivities are somewhat variable across different choices of $\tilde{r}$ and $P(D = 1|X)$ (not known), indicating a need to consider several possible values when the magnitude of the sensitivity estimates themselves are of primary interest. Previous work suggests that the downstream impact of choices for $\tilde{r}$ and $P(D = 1|X)$ on estimated disease model parameters, however, is often small.[5] Supplementary Material Figure B.3 provides the estimates of β associated with X in the sensitivity model. We estimate higher sensitivity with longer follow-up (years) in the EHR (log-odds ratio: 0.14, 95% CI [0.12, 0.16]) and more visits per follow-up time (log-odds ratio: 1.04, 95% CI [0.95, 1.12]).

Handling selection bias

We use two different strategies to address selection bias given estimated sensitivity. In the first strategy, we use summary statistics from SEER, the US Census, and the US CDC to construct poststratification weights. In the second strategy, we use publicly-available data from the NHANES (2017-2018) to construct inverse probability of selection weights. Through these weights, our goal is to account for some of the systematic differences between patients in MGI and patients in the US adult population. In Supplementary Material Table A.1, we demonstrate that MGI is enriched for patients with more comorbidities (eg, diagnosis of coronary artery disease [CAD] or diabetes), and MGI patients tend to be older than the average US adult. This information is incorporated into our weights through W. In this section, we detail how these weights were estimated.

Pulling summary statistics described in Supplementary Material Table A.2, we define three varieties of poststratification weights. Since the joint distribution of disease diagnoses given age is not readily available for the target population, we incorporate multiple diseases assuming independence given age. We define weights *ignoring the cancer outcome* as follows:‘

$$\omega_0 \propto \frac{f(\text{Diabetes}|\text{Age})f(\text{CAD}|\text{Age})f(\text{Age})}{f(\text{Diabetes}|\text{Age}, S = 1)f(\text{CAD}|\text{Age}, S = 1)f(\text{Age}|S = 1)}, \tag{3}$$

where "Diabetes", for example, corresponds to whether the patient received a diabetes diagnosis. Distributions in the numerator come from population summary statistics, and distributions in the denominator are estimated using MGI data. We then *incorporate cancer diagnosis* into weight estimation while correcting for misclassification as follows:

$$\omega \propto \frac{\left[sens \times P(D=1|\text{Age})\right]^{D^*}\left[1-sens \times P(D=1|\text{Age})\right]^{1-D^*}}{\left[P(D^*=1|\text{Age},S=1)\right]^{D^*}\left[1-P(D^*=1|\text{Age},S=1)\right]^{1-D^*}} \times \omega_0, \tag{4}$$

where *sens* is estimated sensitivity ($\tilde{c}$ or $c_{true}(X)$). We substitute population summary statistics (numerator) and MGI estimates (denominator) to obtain these weights. For comparison, we also obtain weights ignoring misclassification by setting $sens = 1$.

To compare weights estimated using different external data sources, we also obtain inverse probability of selection weights using individual-level data from NHANES, incorporating additional information about smoking status, body mass index (BMI), and ethnicity/race (non-Hispanic White, yes/no). Let $S_{ext} = 1$ refer to inclusion in NHANES and $S = 1$ refer to inclusion in our MGI data. We will assume no patients are included in both databases. We first estimate weights *ignoring the cancer outcome* as follows:

$$\begin{aligned}\omega_0 \propto & \frac{1-P(S=1|\text{Age, Diabetes, CAD, BMI, Smoking, Race}, S_{ext}=1 \text{ or } S=1)}{P(S=1|\text{Age, Diabetes, CAD, BMI, Smoking, Race}, S_{ext}=1 \text{ or } S=1)} \\ & \times \frac{1}{P(S_{ext}=1|\text{Age, Diabetes, CAD, BMI, Smoking, Race})}.\end{aligned} \tag{5}$$

The first term accounts for differences between MGI and NHANES and is estimated using logistic regression in the combined MGI and NHANES data. The second term accounts for differences between NHANES and the US adult population. Since NHANES selection weights are provided (but not the selection models themselves), we model NHANES selection using beta regression on the inverted NHANES selection weights.[6] These logistic and beta regression estimates are provided in Supplementary Material Table B.1.

To obtain weights that incorporate cancer diagnosis and also account for phenotype misclassification, we multiply ω_0 from 5 by the following:

$$\frac{[sens \times P(D=1|\text{Covariates})]^{D^*}[1-sens \times P(D=1|\text{Covariates})]^{1-D^*}}{[P(D^*=1|\text{Covariates},S=1)]^{D^*}[1-P(D^*=1|\text{Covariates},S=1)]^{1-D^*}}.$$

We obtain population $P(D = 1|\text{Covariates})$ by fitting a regression model in the NHANES data with covariates age, race, BMI, and smoking status, weighted using the NHANES sample weights. For comparison, we again obtain weights that do not correct for misclassification by setting $sens = 1$.

Estimated poststratification weights and NHANES-based inverse probability of selection weights (after scaling to sum to the number of patients in MGI) are shown for $\tilde{r} = 25$ in Figure 3B. Other values of $\tilde{r}$ are similar. We can see substantial differences in the distribution of weights that do and do not incorporate the cancer outcome. Additionally, weights obtained using NHANES and using SEER/US Census summary statistics tend to be fairly similar in terms of their overall distributions. Weights obtained using these two methods, however, can sometimes differ substantially within individual patients.

A common alternative to weighting used in the survey sampling literature is propensity matching, where the parallel in this scenario would be to match based on probability of being female given other covariates. To compare the results from weights that use external summary information to this internal data-based matching strategy, we construct 1:1 matched sets of males/females using the MGI data. We used nearest neighbor matching, where controls were matched without replacement based propensity score nearest neighbors with a caliper of 0.1. Two matched datasets were constructed, where propensity scores for the first set were based on age, BMI, smoking status, diabetes and CAD diagnosis, while the second set of scores did not adjust for BMI or smoking status. We emphasize that this matching strategy is not designed to obtain inference for a specified target population; rather, this matching strategy is a causal inference-type approach that can address selection bias for estimating θ_Z indirectly by balancing the association between gender and other covariates (eg, age) related to both cancer diagnosis and selection. Therefore, results from this approach do not directly correspond to our estimand of interest, the marginal association between cancer and gender in the US adult population.

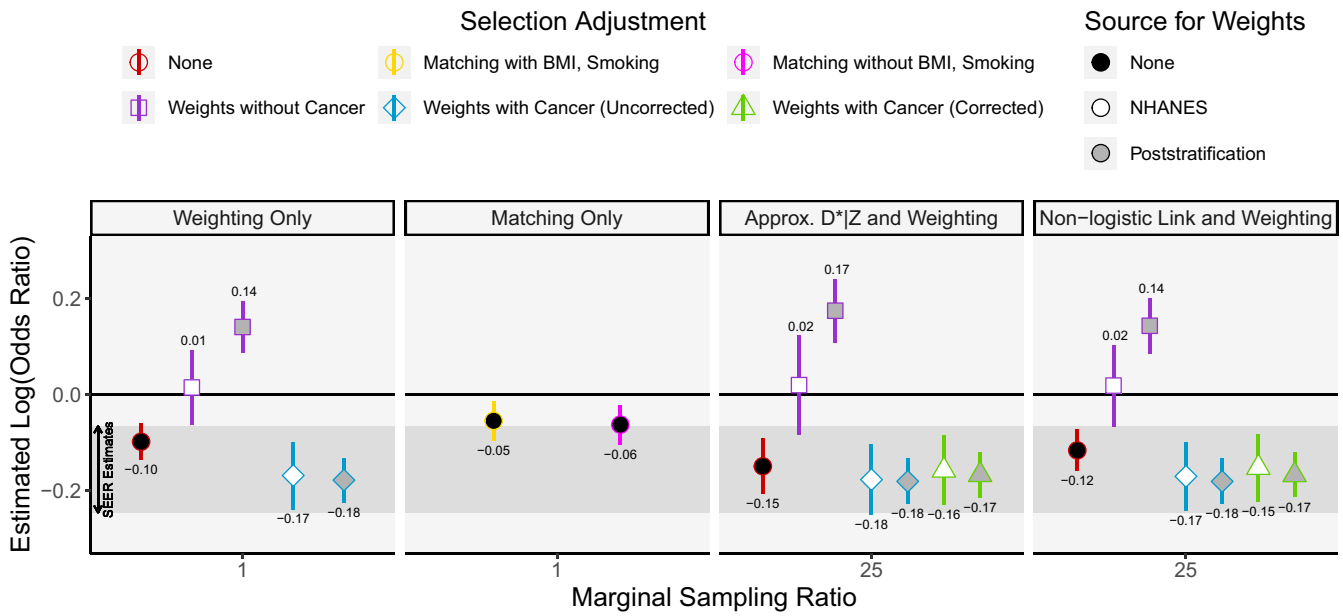

FIGURE 4 Cancer-gender log-odds ratio after applying proposed selection weighting and misclassification adjustment methods (reference = male) [Case study (a)]. Results using a marginal sampling ratio of 25 are shown. Results for sampling ratios of 50 and 100 are similar. The horizontal shaded region corresponds to the range of SEER estimates using data between 2008 and 2016. "Approx. $D^*|Z$" and "Non-logistic Link" correspond to methods 4a and 4b in Figure 2, respectively. The log-odds ratio estimate is printed near each plotted confidence interval

Estimating disease model parameters

Given estimated sensitivity and selection weights, we apply the methods in Figure 2 to estimate the association between cancer and gender (reference = male). Results are shown in Figure 4 and Supplementary Material Table B.2. Uncorrected analysis results in an estimated log odds ratio of −0.10 (95% CI −0.14, −0.06). When we account for misclassification but not selection, we see little qualitative differences in point estimates across methods. This may be due to the fairly high estimated sensitivities for the EHR-derived cancer outcome. Additionally, it may be reasonable to assume that gender (Z) is independent of X given D, so sensitivity $c(Z)$ may be viewed as constant in Z. Assumptions for all three misclassification adjustment methods are satisfied in that case. Interestingly, estimated confidence intervals are narrower for the non-logistic link method (patient-varying sensitivity, interval width: 0.086) than for the approximation method (marginal sensitivity, interval width: 0.115) when we only account for misclassification. This small efficiency gain comes from incorporating covariates X related to D into sensitivity estimation.

We see large differences in the estimated log-odds ratios when we use different strategies to account for selection bias. In particular, weights excluding the cancer diagnosis outcome produce point estimates in entirely the "wrong" direction (eg, a log-odds ratio of 0.14, 95% CI: [0.09,0.19]), reflecting the strong need to incorporate the direct impact of cancer diagnosis on selection when specifying the weights. When we incorporate the cancer outcome in constructing the weights, the resulting point estimates are in the "right" direction (indicating lower rates of cancer diagnosis in women compared to men) for both the NHANES and poststratification weighting strategies (eg, −0.18, 95% CI: [−0.24,−0.13] for NHANES IPW and −0.18, 95% CI: [−0.22,−0.13] for poststratification under approximation method). Additionally, we obtain narrower confidence intervals when we account for selection bias using weights that incorporate the outcome relative to weights that do not incorporate the outcome (eg, widths 0.106 vs 0.093 for poststratification weighting without misclassification adjustment). In this example, we see little impact of correcting for phenotype misclassification in weight development, perhaps due to the high estimated sensitivities for the cancer phenotype. The estimated log odds ratios differ somewhat for weights obtained using poststratification vs NHANES, where poststratification produced stronger cancer-gender associations. In contrast, matching produced very similar point estimates to uncorrected analysis, with little difference observed between matching that did and did not account for BMI and smoking status.

To evaluate whether the differences between NHANES and poststratification weights excluding cancer is due to inclusion of smoking status, BMI, and ethnicity in the NHANES weight construction, we also obtained NHANES weights using only age, diabetes status, and CHD status as predictors. The "Weighting Only" estimate of the cancer-gender log-odds ratio was 0.05, 95% CI: [0.01, 0.10] for weights that did not incorporate cancer. Corresponding weights incorporating cancer diagnosis gave a log-odds ratio of −0.15, 95% CI: [−0.19, −0.11]. Large differences across different weight specifications provide a cautionary tale against ignoring the outcome when estimating selection bias adjustment weights when the outcome is strongly related to selection. Additionally, we get somewhat different results when selection is addressed using different external data sources, and researchers may want to compare results using several different sources in practice. In this example, we may trust results from NHANES over poststratification, since NHANES weights adjust for more patient characteristics and do not assume marginal independence between disease diagnoses.

In Supplementary Material Section C.1, we compare the weighted and unweighted estimates using MGI data with those from the target population for a variety of estimands (eg, mean age). The goal of this analysis is to assess the degree to which the constructed weights can recover target population associations. We find that the IPW weights incorporating the cancer outcome often do a good job at recovering population estimates, while the poststratification weights sometimes perform poorly. This further supports the conclusion that the poststratification weights constructed ignoring the joint distribution between variables may perform poorly for addressing selection bias for these data.

5 | CASE STUDY (B): ASSOCIATION BETWEEN MACULAR DEGENERATION AND GENETIC LOCI USING MGI

Goals of analysis

In case study (b), we want to estimate associations between previously identified genetic loci and age-related macular degeneration (AMD) diagnosis using MGI data, adjusting for age at last visit, gender, and principal components of the genotype data. We define our target population as the US adult population aged 50+ of recent European ancestry (Table 1). AMD is weakly enriched in MGI relative to adults aged 50+ in the US population (Supplementary Material Figure A.2), and we may expect individual genetic loci in Z to be at most weakly associated with selection. Therefore, we may be less concerned with handling of selection bias in this example compared to case study (a). Additionally, we hypothesize that underreporting of disease may be a bigger challenge for case study (b), since we may expect many patients are treated for AMD through local health care providers, and consequently a large number of AMD diagnoses may be missed in the Michigan Medicine EHR. These missed diagnoses may strongly impact estimation of genetic associations. In this second example, we focus on 43 independent genetic loci identified with small P-values ($< 5 \times 10^{-8}$) in a genome-wide association study of over 16 000 advanced AMD cases and 18 000 controls using International AMD Genomics Consortium (IAMDGC) data.[12] Across these 43 loci, MGI and IAMDGC GWAS log-odds ratio point estimates have a Lin's concordance correlation coefficient (CCC) of only 0.61, and uncorrected MGI point estimates generally tend to be closer to the null compared to the IAMDGC estimates (Supplementary Material Figure B.6). The "winner's curse" resulting in inflated IAMDGC point estimates explains some differences, but bias due to selection and misclassification in MGI may also contribute. Below, we apply our methods to explore the extent to which systematic differences in GWAS results between these two studies may be corrected by addressing phenotype misclassification and potential selection bias.

Estimating sensitivity

As with case study (a), we will implement the estimation procedure outlined in Figure 2. Fixing $\tilde{r}$ to different values between 1 and 100, we estimate constant sensitivity and sensitivity as a function of patient-level covariates as in case study (a), restricted to unrelated MGI patients of recent European ancestry. Results are shown in Figure 3A. Sensitivity of the macular degeneration phenotype is estimated to be generally much lower than in case (a) across all values of $\tilde{r}$, with median sensitivity ranging between 0.37 for $\tilde{r} = 10$ and 0.06 for $\tilde{r} = 100$. Higher sensitivities for the cancer phenotype may be related to a more complete disease history for cancer diagnoses relative to macular degeneration diagnoses as entered into the EHR through diagnosis codes. For AMD, there may be some degree of over-diagnosis. Sensitivity estimates assuming imperfect specificity were lower (Supplementary Material Figure B.4) then for perfect specificity.

Handling selection bias

We use summary statistics from SEER, the US Census, and the US CDC to construct poststratification weights. We define weights as if the target population were all US adults of recent European descent. Our analysis then uses data and rescaled weights from patients aged 50+. We obtain three varieties of poststratification weights for the AMD outcome. First, we define weights ω_0 *ignoring the AMD outcome* as in Equation (3) except this time we also incorporate the association between cancer diagnosis and age into both the numerator and denominator. We include cancer diagnosis in the weight definition to account for the strong association between cancer diagnosis and inclusion in MGI, but we do not account for misclassification of the cancer phenotype for this analysis. We then define weights that *incorporate the AMD outcome* using 4, where this time D and D^* correspond to the AMD outcome and *sens* is either the estimated sensitivity (correcting for misclassification) or *sens* = 1 (ignoring misclassification). Resulting weights are shown in Figure 3B for $\tilde{r} = 25$. Other values of $\tilde{r}$ are similar. Unlike case study (a), weights that do and do not incorporate the AMD outcome tend to have similar distributions, reflecting a comparatively small impact of AMD on the probability of inclusion in MGI.

Estimating disease model parameters

We then apply the methods in Figure 2 to obtain bias-corrected point estimates relating macular degeneration diagnosis to 43 genetic loci in MGI. The differences between IAMDGC and MGI point estimates across loci are characterized using three metrics: (i) average absolute difference across 43 pairs of estimates, (ii) Lin's concordance correlation, and (iii) the average absolute percent difference between the MGI and the IAMDGC estimate, relative to the IAMDGC estimate (denoted MAPE; mean absolute percentage error). We also present the average estimated MGI standard errors relative to IAMDGC. Results for the best-performing methods are summarized in Figure 5A. Results for other methods can be found in Supplementary Material Table B.3. We present results using $\tilde{r} = 25$, but other $\tilde{r}$ values with estimable sensitivity (10, 50, 100) are similar. When we correct for selection *or* misclassification, Lin's concordance correlation measure increases from 0.61 (uncorrected) to 0.73 (corrected) or higher. Correcting for both misclassification and selection bias did produce some additional improvement for this metric (Lin's of 0.85). Analyses that accounted for selection (with or without misclassification bias adjustment) resulted in increased (worse) MAPE relative to uncorrected analysis (range 0.85-1.1 vs uncorrected MAPE of 0.81). Unweighted analyses accounting for misclassification but not selection produced similar or better MAPE compared to uncorrected analysis. All bias-correction strategies shown in Figure 5A result in point estimates that are closer to IAMDGC point estimates than in uncorrected analysis on average. Overall, the method approximating the $D^*|Z$ distribution with no selection bias adjustment performs the best among the methods considered in terms of similarity between bias-corrected estimates and IAMDGC estimates. Since selection seems to be at most weakly associated with AMD diagnosis, it is not surprising that methods without selection adjustment generally perform well. Analyses incorporating selection weights had larger standard errors without much gain in terms of bias adjustment, suggesting that selection weighting did not improve inference. In an additional exploration, we calculated these same performance metrics for disease model parameters estimated assuming imperfect specificity as well as imperfect sensitivity. Results for the Approx $D^*|Z$ method without weighting are shown in Supplementary Material Figure B.7. We find that performance improves slightly for specificity set at 0.97 or 0.96 rather than 1, and we see greater and greater discrepancies between MGI and IAMDGC estimates as assumed specificity gets lower.

Figure 5B compares the ranked P-values for each of the 43 genetic loci after bias adjustment to the ranking in IAMDGC. Among the top 5 associations in IAMDGC, the majority are also identified as top associations in MGI. P-values produced by bias correction methods accounting only for misclassification but not selection (no weighting) tend to produce P-values very close or even identical to uncorrected analysis. In Beesley and Mukherjee,[5] we demonstrate that P-values from the non-logistic link function method (ignoring selection) will only differ substantially from uncorrected analysis when $X^\dagger$, representing the factors driving sensitivity not adjusted-for in the disease model, is associated with Z given D. We may be less concerned about the impact of misclassification on P-values when these terms are at most weakly associated (as in this case study). Once selection bias adjustment is incorporated, however, the resulting P-values are impacted, as seen in Figure 5B. In general, selection may often be ignorable when estimating associations with genetic loci. However, we recommend comparing analyses with and without weighting in settings when selection may be more strongly related to Z.

(A)

	Avg. Absolute Deviation	Lin's Concordance Correlation	MAPE	Avg. Relative Standard Error
IAMDGC Top Hits	0	1	0	1
Uncorrected Analysis	0.30	0.61	0.81	2.2
Weighting only	**0.25**	**0.82**	0.85	4.2
Approx. $D^*\|Z$ method [4]	**0.26**	0.76	**0.75**	3.1
Approx. $D^*\|Z$ method + Weighting	**0.26**	**0.85**	0.90	5.6
Non-logistic Link method [3]	**0.28**	0.73	**0.79**	3.1

(B)

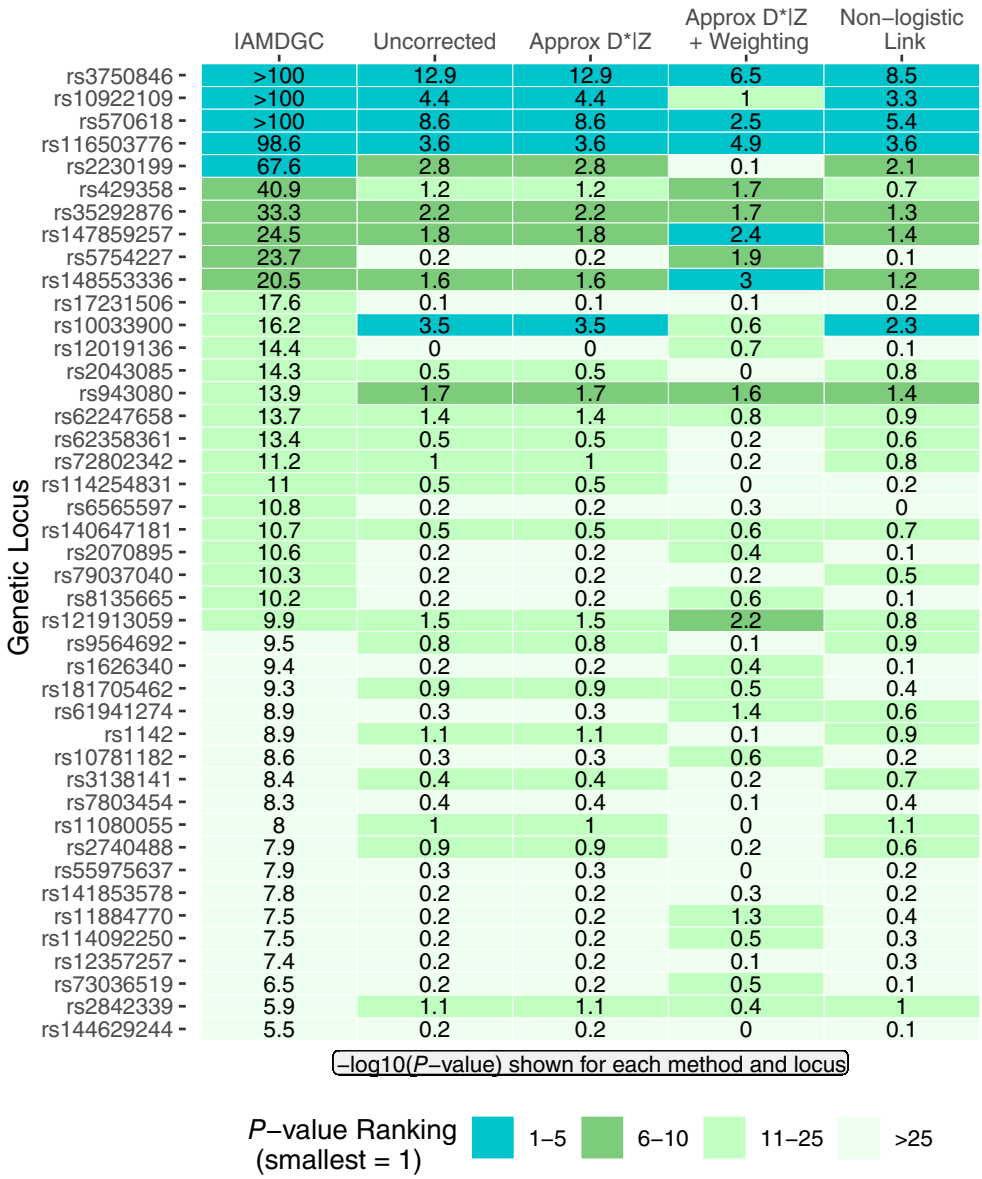

FIGURE 5 Bias-adjusted AMD log-odds ratios across 43 genetic loci and corresponding *P*-values [Case study (b)]. (A) Log-odds ratio summary metrics across 43 genetic loci. (B) $-\log_{10}(P\text{-values})$ and *P*-value rankings for each locus. For Approx. $D^*|Z$ [method 4a] and Non-logistic link function [method 4c] strategies, sensitivity is estimated assuming $\tilde{r} = 25$. Methods with weighting used weights *ignoring the AMD outcome*. Bolded values indicate the best performing methods. Average absolute deviation, average absolute difference between MGI and IAMDGC point estimates (lower is better); Lin's concordance correlation, estimated concordance between MGI and IAMDGC point estimates (higher is better); MAPE (mean absolute percentage error), average absolute difference between 1 and the ratio of MGI and IAMDGC point estimates (lower is better); Avg. relative standard error, ratio of standard errors for MGI and IAMDGC point estimates

6 | DISCUSSION

Many statistical challenges arise in the analysis of electronic health record (EHR) data, including limitations in data quality (ie, measurement error, missing data, etc.), lack of representativeness (ie, who is in the study?), and generalizability (ie, what do results say about my target population?). In Beesley and Mukherjee,[5] we proposed a suite of statistical tools for addressing measurement error and selection bias in disease modeling using EHR data. That work (which can be viewed as an extension of many leading methods in this area)[1,3,4,6] demonstrated good performance of the proposed methods when key factors related to selection and measurement error are observed, but these driving factors may be unknown or only partially measured in practice. In this paper, we illustrate how these statistical bias-correction and inference strategies can be applied in real-world data analysis through two EHR data analysis case studies, and we evaluate their performance.

We consider data from the Michigan Genomics Initiative, a longitudinal EHR-linked biorepository effort within Michigan Medicine. For both of these case studies, comparative gold standard disease associations were used to benchmark the performance of various bias reduction strategies. In case study (a), bias-corrected point estimates for the association between cancer and gender were consistent with associations reported by SEER as long as the cancer outcome was incorporated into development of selection weights. In case study (b), these bias reduction methods resulted in point estimates closer on average to previously identified associations than uncorrected analysis.[12] These case studies demonstrate that the bias correction and inference strategies from Beesley and Mukherjee[5] and others may be useful for *reducing* bias in EHR-based studies even when factors related to selection and misclassification are not well-understood or fully measured.

We emphasize that the goal of *unbiased* estimation in EHR data analysis may be unrealistic given limitations in data availability and many competing sources of bias. Instead, our goal in implementing these methods is to produce *less biased* inference, where here bias is a function of our estimand, the data provenance, and the target population. Given the varying results observed as a function of the analytical approach, these examples highlight the need to tailor the statistical approach to the problem at hand (ie, the same EHR sample may have selection bias in one analysis but not another) and illustrate settings where disease model inference can be sensitive to our strategy for handling bias adjustment. When traversing the rocky terrain of observational data analysis using EHR data, analysts must combine nuanced and thoughtful analysis with knowledge of the scientific context and interpret results from EHR-based health research with an appropriate dose of skepticism. While methods in Beesley and Mukherjee and elsewhere can help account for bias in EHR data analysis, even careful implementation of these will ultimately be limited by the data that are available. In some settings, existing design-based strategies for evaluating and addressing selection biases may outperform sophisticated weighting strategies when high-quality data on key factors related to selection/testing in untested patients are unavailable.

Case study (a) demonstrates a problem that may often arise in analyses comparing different types of selection weights: what do you do when the weights give different conclusions? While we cannot know which adjustment strategy produces the "better" results without knowing the true association, we may trust weights constructed based on adjusted individual-level data more because they may be able to better capture the complex relationships between key variables related to the outcome and/or related to selection. Restated, we may trust the weights constructed using the most reasonable assumptions. In case study (a), we hypothesize that the differences in estimated cancer-gender associations could be the result of two different factors: (1) NHANES weights adjust for smoking status (and body mass index/ethnicity), which may be related to both selection (eg, smokers may be more likely to go to the hospital and be included in MGI) and the cancer outcome of interest and (2) post-stratification assumes independence between the disease diagnoses given age while the NHANES modeling allows us to model the adjusted relationship between disease diagnoses directly and also condition on age, smoking status, and race. Large differences between poststratification and NHANES weights were seen when the weights did not condition on cancer status, but the differences between the data sources were much smaller when weights were constructed conditioning on the outcome. In general, disparate results across weights provide some sense of robustness (or lack thereof) of the disease model estimates to different attempts to account for selection bias. In Supplementary Material Section C.1, we implement several diagnostics for assessing the reasonableness of transportability assumptions as discussed in Degtiar and Rose.[10] Poststratification weights constructed for MGI in both case studies did a poor job of recovering target population summary statistics, while the NHANES-based IPW weights performed comparatively well. When combined with an understanding of the scientific context and bias tolerance, such diagnostics may prove useful for evaluating the reasonableness of weights constructed for selection bias adjustment.

This work provides a roadmap for practical implementation of the methods for handling phenotype misclassification and selection bias in EHR data analysis proposed in Beesley and Mukherjee.[5] These methods are summarized

in Figure 2 and can be implemented in R using package *SAMBA* available at https://cran.r-project.org/web/packages/SAMBA/index.html. Example code is provided as Supplementary Material. These methods rely on an assumed logistic regression structure for the distribution of D^* given $D = 1$ and X, but more general model structures could also be used.[5] When potential X has large dimension, penalization methods could aid in estimation of β. However, we require an estimate of $P(D = 1|X)$, and simulations in Beesley and Mukherjee demonstrating robustness of disease model estimates to misspecification of $P(D = 1|X)$ were performed for the small-dimensional X. Similar strategies could also be used to incorporate a larger number of predictors or more complicated covariate relationships (eg, interactions) in the disease model. Estimation of selection weights (Step 3) presents a harder problem, and several strategies are highlighted in Figure 2. When individual-level data from the target population are available, inclusion in the EHR sample can be directly modeled. In case study (a), we use logistic regression to model this selection probability in the merged internal and external datasets, but more sophisticated modeling strategies, penalization, and so on can also be used. Additional strategies for handling multi-stage sampling, overlapping EHR and external probability samples, and use of probability samples from a different target population can be found in Beesley and Mukherjee.[5]

ACKNOWLEDGEMENTS

We thank Chad Brummett, Goncalo Abecasis, and Sachin 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 award 1712933 and The University of Michigan precision health award U067541. We thank Alexander Rix for his work developing R package *SAMBA*. We thank Lars Fritsche and the International AMD Genomics Consortium (IAMDGC) for providing GWAS summary statistics. This work was partially funded by the Laboratory Directed Research and Development (LDRD) Richard Feynman Postdoctoral Fellowship 20210761PRD1. This work is approved for distribution under LA-UR-22-23875. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of Los Alamos National Laboratory. Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is managed by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract 89233218CNA000001.

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/.

ORCID

Lauren J. Beesley https://orcid.org/0000-0002-3788-5944
Bhramar Mukherjee https://orcid.org/0000-0003-0118-4561

REFERENCES

1. Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs*. 2016;4(1):1-17.
2. Huang J, Duan R, Hubbard RA, et al. PIE: a prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *J Am Med Inform Assoc*. 2018;25(3):345-352.
3. Sinnott JA, Dai W, Liao KP, et al. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Hum Genet*. 2014;133(11):1369-1382.
4. Duffy SW, Warwick J, Williams AR, et al. A simple model for potential use with a misclassified binary outcome in epidemiology. *J Epidemiol Community Health*. 2004;58(8):712-717.
5. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*. 2022;78(1):214-226.
6. Elliot MR. Combining data from probability and non- probability samples using pseudo-weights. *Surv Pract*. 2009;2(3):1-7.
7. Fritsche LG, Gruber SB, Wu Z, et al. Association of Polygenic Risk Scores for multiple cancers in a Phenome-wide study: results from the Michigan genomics initiative. *Am J Hum Genet*. 2018;102(6):1-14.
8. Beesley LJ, Salvatore M, Fritsche LG, et al. The emerging landscape of epidemiological research based on biobanks linked to electronic health records. *Stat Med*. 2019;39(6):773-800.
9. Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. *Eur J Epidemiol*. 2019;34(8):719-722. doi:10.1007/s10654-019-00533-2
10. Degtiar I, Rose SA. Review of generalizability and transportability arXiv preprint, arXiv:2102.11904.

11. Tipton E. How generalizable is your experiment? An index for comparing experimental samples and populations. *J Edu Behav Stat*. 2014;39(6):478-501.
12. Fritsche LG, Igl W, Cooke Bailey JN, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*. 2016;48(2):134-143. doi:10.1038/ng.3448.A
13. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models author. *Biometrika*. 1993;80(4):807-815.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Beesley LJ, Mukherjee B. Case studies in bias reduction and inference for electronic health record data with selection bias and phenotype misclassification. *Statistics in Medicine*. 2022;41(28):5501-5516. doi: 10.1002/sim.9579