Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

An Analytic Framework for Exploring Sampling and Observation Process Biases in Genome and Phenome-wide Association Studies using Electronic Health Records

Lauren J. Beesley* | Lars G. Fritsche | Bhramar Mukherjee

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

Correspondence
*Corresponding author: Email: lbeesley@umich.edu

Present Address

Summary

Large-scale association analyses based on observational health care databases such as electronic health records have been a topic of increasing interest in the scientific community. However, challenges due to non-probability sampling and phenotype misclassification associated with the use of these data sources are often ignored in standard analyses. The extent of the bias introduced by ignoring these factors is not well-characterized. In this paper, we develop an analytic framework for characterizing the bias expected in disease-gene association studies based on electronic health records when disease status misclassification and the sampling mechanism are ignored. Through a sensitivity analysis approach, this framework can be used to obtain plausible values for parameters of interest given *summary results* from standard analysis. We develop an online tool for performing this sensitivity analysis. Simulations demonstrate promising properties of the proposed method. We apply our approach to study bias in disease-gene association studies using electronic health record data from the Michigan Genomics Initiative, a longitudinal biorepository effort within The University Michigan health system.

KEYWORDS:
Keywords: electronic health records, GWAS, non-probability sampling, outcome misclassification, PheWAS

1 | INTRODUCTION

Genome-wide genotype data linked with electronic health records (EHR) are becoming increasingly available through biorepository efforts at academic medical centers, health care organizations, and population-based biobanks.[1] A common use of these linked data is to explore the association between a phenotype, D, with a risk factor of interest, G, after adjusting for potential confounders, Z. Analysis using a regression model for $D|G, Z$ may be repeated for millions of risk factors or genetic variants with a given D of interest (as in genome-wide association studies, or GWAS) or for thousands of phenotypes derived from the content of the EHR with a given variant G of interest (as in phenome-wide association studies, or PheWAS). Association analyses embedded within large observational databases have gained popularity in recent years, and the use of and interest in such analyses continues to increase.[1,2,3] However, unlike curated and well-designed population-based studies, large observational databases are often not originally intended for research purposes, and additional thought is needed to understand potential sources of bias and conduct principled inference. In this paper, we focus on the particular association setting with D being a single EHR-derived phenotype and G being a single genetic marker or a polygenic risk score, but the methods and conceptual framework developed in this paper can be applied quite broadly to general estimation problems using observational databases. This framework is scalable and can be applied to explore the phoneme by genome landscape, where the basic unit of analysis is still of the form of $D|G, Z$.

One potential source of bias in EHR-based studies is misclassification of derived disease status variables. Large-scale agnostic studies using EHR often define patient disease status (phenotypes) based on international classification of disease (ICD) codes or aggregates thereof called “PheWAS codes" or “phecodes" to define phenotypes in an automated and reproducible way.[4] In practice, EHR-derived phenotype definitions are used to represent the underlying ‘true’ disease status. However, these ICD-based phenotype classifications can be erroneous in capturing the ‘true’ disease status for a variety of reasons. For example, psychiatric diseases can be difficult to diagnose, and diagnosis can often be subjective.[5] ICD code-based diagnoses may be an incomplete representation of a patient’s health status, which may also be recorded in doctor’s notes and elsewhere in the EHR. In response to this problem, there exists an extensive literature on using other structured and unstructured content of the EHR to define phenotypes more accurately.[6,7,8,9,10] Additionally, human validation can be used to evaluate phenotyping algorithms.[11] These existing phenotyping approaches can be effective in reducing misclassification given the information available in the EHR, but even the most sophisticated phenotyping algorithms cannot capture diagnoses that were *never recorded* in any form in the EHR. Secondary conditions may not always be entered into the EHR, and symptoms occurring between visits may not always be reported. The EHR cannot adequately capture diseases that a patient had prior to entry into the EHR (outside the observation window). Unlike classical misclassification settings, the chance of correctly capturing a disease is inherently dependent on the length of stay in the EHR or the observation/encounter process for a given patient. We often have a systematic source of misclassification (that we will call “observation window bias") due to a lack of comprehensiveness of the EHR in capturing diagnoses or medical care obtained from outside sources (e.g. at another health care center). Together, these various factors can lead to a potentially large degree of misclassification, particularly due to underreporting of disease.

Several authors have proposed statistical methods for addressing misclassification of binary phenotypes in EHR-based studies. The extent of misclassification can be described using quantities such as sensitivity and specificity, but these quantities can vary from population to population and from phenotype to phenotype.[12] Huang et al. (2018) proposes a likelihood-based approach that integrates over unknown sensitivity and specificity values but requires some limited prior information about the sensitivity and specificity.[13] Wang et al. (2017) proposes an approach for incorporating both human-validated labels and error-prone phenotypes into the estimation, but this approach will not account for observation window bias.[14] Duffy et al. (2004) and Sinnott et al. (2014) expand on results in the measurement error literature to relate parameters in the model for the true outcome with the model for the misclassified outcome, but the former focuses on outcome misclassification with *binary* risk factors, and the latter focuses on the setting in which the probability of having observed disease is explicitly modeled using a variety of information in the EHR.[15,16] Additionally, all of these methods do not directly address the sampling mechanism.

In addition to bias due to phenotype misclassification, the mechanism by which units in the population are included in the EHR dataset can sometimes result in biased inference when not handled appropriately. Complex sampling designs in an epidemiologic study can be addressed using survey design techniques if the sampling strategy is known. However, the probability mechanism for inclusion of a person into a biorepository is not *a priori* fixed or defined. For convenience, we will use terms such as “sampling" and "selection mechanism" to describe this patient inclusion process, but it should be understood that this process is complicated and not well-characterized. Interactions with the health care system are generated by the patient, and it can be difficult to understand the mechanism driving sampling as well as self-selection for donating biosamples, which may be related to a broad spectrum of patient factors including overall health and demographic characteristics. Several authors recommend adjusting for factors such as number of health care visits or referral patterns to better account for the sampling mechanism.[17,18] Moreover, there is a belief in the literature that gene-related association study results may be less susceptible to bias resulting from patient selection.[19] This belief stems from the assumption that an individual genetic locus is not usually appreciably related to selection. However, it has been shown that bias due to genotype relationships with selection can still arise in certain settings.[20] Additionally, a popular topic in genetics-related research right now is the use of polygenic risk scores, which combine information from many genetic loci to quantify a patient’s overall inherited genetic risk for developing a particular disease.[21,22] While it may be reasonable to assume that a specific genetic locus may have little association with selection, this assumption becomes more tenuous for an aggregate polygenic risk score with stronger association with the underlying disease and other factors related to selection.

As we will demonstrate, patient sampling can create substantial bias in estimating genetic associations using EHR data in the presence of disease status misclassification. Existing statistical methods for dealing with phenotype misclassification do not directly take into account the mechanism by which patients are sampled and *vice versa*. Additionally, standard association studies often do not account for either source of potential bias. It is important to understand the implication of these sampling and observation processes on results from standard association analyses.

In this paper, we develop an analytic framework incorporating both disease misclassification and the patient selection mechanism. We use this framework to characterize the amount of bias expected in EHR-based association study results when misclassification and the sampling mechanism are ignored (as is the common strategy). We focus on the particular setting in which phenotypes are underreported (naturally occurring for EHR data due to limited observation window), but we also provide an extension allowing for bi-directional misclassification. The analytical expressions enhance our understanding of study design and phenotype characteristics this bias may depend on. Through a sensitivity analysis approach, this framework can also be used to obtain plausible values for parameters of interest given *summary results* from simpler analysis. Simulations demonstrate promising properties of the proposed approach in capturing the true bias, and we provide an interactive online tool for performing these sensitivity analyses.

We then apply our proposed methods to data from The Michigan Genomics Initiative (MGI), a longitudinal biorepository effort within the University of Michigan health system with linked genotype and EHR information for over 40,000 patients. The patients were recruited in the Anesthesiology clinic while waiting for a surgery/diagnostic procedure. Using these data, we are often interested in studying the association between disease and genetic factors, adjusting for demographics and other patient characteristics. However, our EHR-derived disease status may have substantial misclassification relative to patients' true disease status. Additionally, we are often interested in making statements about external target populations such as the US population, and the MGI patient pool may be poorly representative of this target population. When ignored, these factors can create a large degree of bias in our association analyses of interest. We apply our proposed approach to explore the potential degree of bias in two example genetic association studies in MGI, which can serve as a tutorial for how the proposed methods can inform bias exploration after standard association analysis.

2 | MODEL STRUCTURE

Let the binary variable D represent a person's true disease status. Suppose we are interested in the relationship between D and an individual's inherited genetic information, G, adjusting for additional person-level information, Z. We will call this relationship the *disease mechanism* as seen in **Figure 1**. In genetic association studies, G may represent a single SNP (single nucleotide polymorphism) or a polygenic risk score.[21] In principle, however, G can be any risk factor of interest. Z often contains information such as age, gender, and several principal components derived from the patient's genome-wide data. Principal components are often included as adjustment factors as a way to adjust for patients' genetic ancestry. In practice, we may have many diseases or genotypes of interest in an association study, but for now we will consider the simple setting with a specific (D, G) pair.

In studying the D/G association using a large health care system-based database, we may have the goal of making inference regarding the D/G association in some pre-defined target population. For example, we may want to generalize results to the US adult population during the time period covered by our data. Let S indicate whether a particular person in the population is selected into our dataset (for example, by going to a particular hospital and consenting to share a biosample for research), where the probability of a person in the population being included in the current dataset may depend on the underlying disease status, D, along with additional covariates, W, and perhaps even G. We may often expect the sampled and non-sampled subjects to have different rates of the disease, and other factors such as patient age, residence, access to care and general health status may also impact whether subjects are sampled into the study dataset or not. We will call this the *sampling mechanism.*

Instances of the disease are recorded in hospital or administrative records. We might expect factors X such as the patient age, the length of follow-up, and the number of hospital visits to impact whether we actually *observe/record* the disease of interest for person with the disease. Whether we incorrectly record a person as having the disease might be related to patient factors Y (e.g. age). Define the *observed* disease status as D^*. D^* is a potentially misclassified version of D. We will call the mechanism generating D^* the *observation mechanism.*

The diagram in **Figure 1** shows the conceptual model structure, expressed as follows:

$$
\begin{aligned}
&\textbf{Conceptual Model} && \qquad (1)\\
&\text{Disease Mechanism}: && \text{logit}(P(D=1|Z,G;\theta)) = \theta_0 + \theta_G G + \theta_Z Z\\
&\text{Sampling Mechanism}: && P(S=1|D,W,G,Z;\phi)\\
&\text{Observation Mechanisms}: && P(D^*=1|D=1,S=1,X,G,Z;\beta) \quad (Sensitivity)\\
& && P(D^*=1|D=0,S=1,Y,G,Z;\alpha) \quad (1-Specificity)
\end{aligned}
$$

This framework allows for misclassification of the true disease status in either direction (patients with the disease may be missed and patients may be listed as diseased who aren't). Moving forward, however, we will restrict our focus to the particular setting where disease status misclassification comes through underreporting. In other words, **we will assume perfect specificity** with $P(D^* = 1|D = 0, S = 1, Y, G, Z; \beta) = 0$ for all patients. This assumption may be reasonable for some EHR-derived phenotypes, particularly cancer, where we expect the rate of overdiagnosis of disease to be generally low. We consider the more general setting with imperfect specificity in detail in **Supplementary Section S3**.

Crucially, X and W may not always be fully known or measured for a given data analysis. For example, proximity to the hospital may likely drive whether a patient is included in a given EHR, but patients' home addresses may not be available for privacy reasons. Standard association analyses typically do not adjust for factors that may influence selection of participants into EHR. In our subsequent analysis, we will leverage the model structure in (1) along with assumptions about X and W to explore bias without requiring X and W to be measured or known. In particular, we assume that the following hold:

Assumption 1: $S \perp G|D, W, Z$ and $D^* \perp G|D, S = 1, X, Z$
Assumption 2: $X \perp W|D, G, Z, S = 1$
Assumption 3: $X \perp G|D, Z, S = 1$
Assumption 4: $W \perp G|D, Z$

where W and X correspond to the (theoretical) predictors driving sampling and sensitivity that are *not* included in Z, the set of adjustment factors in the disease model.

The first assumption implies that **G is not an *independent* predictor of S or D^***. This seems reasonable, especially when G represents a single SNP. This is an important assumption, and our proposed methods may often under-estimate the bias of standard analysis when this first assumption is violated. The second assumption states that the factors related to sampling and related to sensitivity that are not adjusted for in the disease model are independent given D, G, and Z in the sampled subjects. Importantly, this independence assumption conditions on D, G, and Z, and Z often includes common information such as patient age, gender, and the first few genetic principal components. Conditionally, we may expect this assumption to be reasonable for many EHR settings. The third assumption implies that G is unassociated with the factors related to misclassification given D and Z on the sampled subjects. X is expected to contain information relating to a patient's observation process (e.g. length of follow-up), and this will generally not be driven by a patient's genetic information given Z and D. These first three conditional independence assumptions, therefore, may often be reasonable in typical EHR data analysis settings.

We view the fourth assumption as the strongest, since it implies that **G is independent of factors related to sampling not included in the disease model given D and Z**. Suppose, however, that sampling is related to a secondary disease, D'. This may often be the case for EHR data. If D' is independent of G, then this dependence will not create a problem, and the fourth assumption will be satisfied. However, if D' is *independently related* to G (given Z and D), the fourth assumption is violated. An example of this would be the setting of pleiotropy, where a particular SNP may be related to multiple phenotypes, which could each contribute to the sampling mechanism. This setting has been explored extensively in the literature on secondary analyses of case-control sampled data.[23,24] Importantly, the fourth assumption will be trivially satisfied if we include D' and other factors related to sampling and G as adjustment factors in the simple analysis model. Therefore, the proposed methods can be applied to characterize bias of standard analysis even in the setting with sampling related to secondary diseases if we adjust for these secondary diseases in the simple analysis model. We discuss this issue in more detail in **Supplementary Section S5**.

3 | APPROXIMATING BIAS RESULTING FROM STANDARD ANALYSIS

3.1 | Approximating the Bias

Suppose that G, Z, X, and W are observed if and only if $S = 1$. We further require that assumptions 1-4 hold. In standard analyses, we often fit a simple logistic regression for $D^*|G, Z, S = 1$ (*analysis model*) with the goal of making inference about θ from the (assumed) *true model* summarized as follows:

$$\begin{aligned} \text{Analysis Model}: &\quad \text{logit}(P(D^* = 1|Z, G, S = 1)) = \theta_0^{(simple)} + \theta_G^{(simple)} G + \theta_Z^{(simple)} Z \\ \text{True Model}: &\quad \text{logit}(P(D = 1|Z, G)) = \theta_0 + \theta_G G + \theta_Z Z \end{aligned} \tag{2}$$

In general, θ_G and $\theta_G^{(simple)}$ may be unequal. As shown in the **Supplementary Section S2**, we can relate the analysis model to the conceptual model in (1) as follows:

$$P(D^* = 1|G, Z, S = 1; \theta^{(simple)}) = \frac{P(D = 1|G, Z; \theta)c_1(Z; \beta)r(Z; \phi)}{P(D = 1|G, Z; \theta)[1 - c_1(Z; \beta)]r(Z; \phi) + P(D = 0|G, Z; \theta)}$$

Interestingly, the impact of selection and misclassification on the relationship between the target and analysis models boils down to contributions of two functions of Z as follows:

$$\begin{aligned} c_1(Z; \beta) &= \int f(D^* = 1|D = 1, X, Z, S = 1; \beta) f(X|D = 1, Z, S = 1) dX = \text{Sensitivity} \\ r(Z; \phi) &= \frac{\int P(S = 1|D = 1, W, Z; \phi) f(W|D = 1, Z) dW}{\int P(S = 1|D = 0, W, Z; \phi) f(W|D = 0, Z) dW} = \text{Sampling Ratio} \end{aligned}$$

We can view D^* as a noisy diagnostic for the true value of D with potentially imperfect sensitivity, where $c_1(Z; \beta)$ represents the sensitivity of D^* for D in the sampled patients, averaged over the distribution of X. We consider the more general setting with imperfect specificity in **Supplementary Section S3.1**. $r(Z; \phi)$ represents the ratio of sampling probabilities comparing subjects with $D = 1$ and $D = 0$, where each probability is averaged over the distribution of W. These expressions may both depend on Z. However, $c_1(Z; \beta)$ and $r(Z; \phi)$ depend on distinct parameters, so they can vary independently conditional on Z. These functions theoretically depend on various covariate distributions, but we will *not* need to specify these distributions in practice.

We assume that covariates Z are centered so that they have mean zero and that the data are modeled as in (1). Let $\bar{c}_1$ and $\bar{r}$ be $c_1(Z; \beta)$ and $r(Z; \phi)$ evaluated at $\bar{Z} = 0$. In **Supplementary Section S2.1**, we use first order Taylor series approximation to obtain expressions for the intercept and coefficient for G in the *analysis model* as functions of parameters in the *true model*. We consider two different cases with two types of genetic variables G.

Case 1: G is a SNP Suppose that G represents a single genetic locus or SNP, coded 0/1/2 to represent the patient's number of copies of the minor allele at a bi-allelic locus. Let MAF denote known minor allele frequency in the *population*, which we assume is at least 0.05 (common variant). We can express the *analysis model* parameters as a function of the *true model* parameters as follows:

$$\begin{aligned} \theta_0^{(simple)} &\approx \theta_0 + \log(\bar{c}_1) + \log(\bar{r}) - \log\left(e^{\theta_0 + 2\theta_G MAF}[1 - \bar{c}_1]\bar{r} + 1\right) + 2\theta_G MAF \left[\frac{e^{\theta_0 + 2\theta_G MAF}(1 - \bar{c}_1)\bar{r}}{e^{\theta_0 + 2\theta_G MAF}(1 - \bar{c}_1)\bar{r} + 1}\right] \\ \theta_G^{(simple)} &\approx \left[\frac{1}{e^{\theta_0 + 2\theta_G MAF}(1 - \bar{c}_1)\bar{r} + 1}\right]\theta_G \end{aligned} \tag{3}$$

Here, we approximate the sample MAF with the population MAF as discussed in **Supplementary Section S3.1**.

Case 2: G is a Polygenic Risk Score Suppose instead that G is a continuous predictor such as a polygenic risk score (as discussed in Dudbridge et al. (2013)) and suppose G has been centered to have mean zero.[25] In this setting, we can express the *analysis model* parameters as follows:

$$\begin{aligned} \theta_0^{(simple)} &\approx \theta_0 + \log(\bar{c}_1) + \log(\bar{r}) - \log\left(e^{\theta_0}[1 - \bar{c}_1]\bar{r} + 1\right) \\ \theta_G^{(simple)} &\approx \left[\frac{1}{e^{\theta_0}(1 - \bar{c}_1)\bar{r} + 1}\right]\theta_G \end{aligned} \tag{4}$$

Simulations exploring the correspondence between the approximations in (3) and (4) and standard estimated parameters can be found in **Figure S2**. These simulations indicate an excellent ability of these expressions to capture bias in simulated data.

3.2 | Understanding the Structure of the Bias

We can use expressions (3) and (4) to develop some intuition for settings in which we expect greater or less relative bias when performing the "simple" routine analysis. **Table S1** describes the general impact of the various model parameters on the bias of $\theta_G^{(simple)}$, the log-odds ratio parameter associated with G in the simple analysis model. As expected under assumptions 1-4, there is no bias in estimating $\theta_G^{(simple)}$ when we have perfect sensitivity. This may not be the case if one or more of these assumptions are violated. Suppose instead that $\bar{c}_1 < 1$. In this case, we expect bias in $\theta_G^{(simple)}$, and the absolute bias will depend on the sensitivity, the disease sampling ratio $\bar{r}$, the population prevalence of the disease, the magnitude of θ_G, and (in case 1) the MAF of the genetic locus of interest.

Suppose we assume that the sampling ratio is 1 or 5, so diseased and non-diseased people are sampled at a 5:1 or 1:1 ratio on average respectively. **Figure 2** shows the value for $\theta_G^{(simple)}$ we may obtain if we fit a model for $D^*|G, Z$ using the sampled data with a minor allele frequency of 0.2 and true θ_G either 0.5 or 0.1. As the sensitivity decreases, the "simple" estimate of θ_G becomes increasingly biased (relative and absolute bias) toward the null. The level of bias depends strongly on the population prevalence of the disease. For less common diseases (e.g. less than 10% of the population), we may expect to observe relatively low relative bias in estimating θ_G even with low sensitivity for observing D in the sampled patients. Additionally, we expect the absolute bias to be generally small when we have lower true values for θ_G. For moderate to large values of θ_G, misclassification and sampling can have a substantial impact on the absolute bias for $\theta_G^{(simple)}$. This provides some support to the notion that absolute bias in GWAS studies (where the θ_G values are expected to be generally small) may be less of a concern, particularly when we are studying a less common disease. However, if we are studying a disease that has higher population prevalence or larger values for θ_G (for example, a highly penetrant SNP), we may be more concerned about the potential for absolute bias induced by ignoring disease misclassification and/or sampling. As shown in **Table S2**, the bias has a more complicated relationship with disease prevalence and sensitivity in the presence of disease overreporting. We created an online RShiny tool called SAMBA-EHR (**S**ampling **A**nd **M**isclassification **B**ias **A**nalysis for EHR-based association studies) for exploring the impact of the various parameters on the bias for both the intercept and association parameters, available at http://shiny.sph.umich.edu/SAMBA-EHR/.

3.3 | Sensitivity analysis of potential bias in genetic association estimates

An alternative use of (3) and (4) is in a sensitivity analysis in a reverse direction, where we can obtain reasonable values of θ_G across plausible values of $\bar{r}$ and $\bar{c}_1$ based on the parameter estimates from the "simple" standard analysis. Suppose we have an estimate for $\theta_G^{(simple)}$ and that G is a single genetic locus. This estimate can be based on direct data analysis or can be obtained from published summary statistics. Using information from the population to obtain reasonable values for θ_0, we can explore plausible values for θ_G by solving (3) for θ_G. In practice, θ_0 itself may not be known. In this setting, we propose performing the sensitivity analysis for a plausible window of θ_0 using a rough estimate for the population prevalence, $P(D = 1)$, or treating θ_0 as an input value. Alternatively, suppose that $\theta_0^{(simple)}$ is also available. In this case, we can obtain θ_G by solving

$$\theta_G^{(simple)} \approx \frac{\bar{c}_1 - e^{\theta_0^{(simple)}}(1 - \bar{c}_1)}{\bar{c}_1 - e^{\theta_0^{(simple)}}(1 - \bar{c}_1)\left[1 - e^{2\theta_G MAF}\right]}\theta_G \tag{5}$$

for θ_G. In this case, predicted θ_G will be a function of $\bar{c}_1$ and $\theta^{(simple)}$ but not $\bar{r}$. Similar expressions for predicting θ_G when G is a PRS can be found in **Supplementary Section S2.2**, and the setting with imperfect specificity is considered in **Supplementary Section S3.2**. In some settings, these expressions may have no solution for given values of $\bar{r}$ and $\bar{c}_1$. When this happens, it usually corresponds to a setting that is highly incompatible with the observed data.

4 | SIMULATIONS EVALUATING PROPOSED SENSITIVITY ANALYSIS

In this section, we present results from a simulation study evaluating the relationship between true θ_G and values predicted using the sensitivity analysis approach proposed in **Section 3.3**. Since the method used to derive the predicted θ_G depends on Taylor series approximations and relies on Assumptions 1-4 in **Section 2** holding, we are particularly interested in evaluating settings in which our sensitivity analysis approach does and does not do a good job of predicting θ_G.

The simulation study is broken up into two parts. In part 1, we consider a simple setting where G is a genotype on a single marker with MAF = 0.2 and where Z, W and X are empty. We simulate 100 datasets with 5000 patients each following (1) and under different population disease rates (1, 5, and 10%). We then impose sub-sampling and disease status misclassification with different values of $\bar{c}_1$ and $\bar{r}$. In order to evaluate the role of relationships between G, Z, X, and W in our sensitivity analysis predictions, we perform a second set of simulations. In simulation part 2, we consider a more complicated simulation scenario in which G is a PRS and has different relationships with non-empty Z, W, and $X = (X_1, X_2)$. In each setting, we again simulate 100 datasets with 5000 patients each, all with an average sensitivity of roughly 0.4 and a population disease rate of roughly 5%. We consider 8 different relationships. In the first setting, G and Z are related but independent of W and X. In settings 2-4, Z is associated with X. In settings 5-6, Z is associated with W. In settings 1-6, key assumptions 1-4 are satisfied. Settings 7 and 8 consider violations of assumptions 3 and 4 respectively. True disease status was generated assuming a 5% population disease rate, and misclassification and sub-sampling were imposed given $\bar{c} \approx 0.4$ and $\bar{r}$ between 1 and 2. Additional details are available in **Supplementary Section S6**. For each simulated dataset, we estimate $\theta^{(simple)}$ by fitting a logistic regression model for the misclassified outcome D^* given G and possibly Z on the sampled patients.

Part 1a: predicting θ_G with available $\theta_0^{(simple)}$: First, we suppose that both $\theta_0^{(simple)}$ and $\theta_G^{(simple)}$ are available for each simulated dataset. We then apply methods in (5) to obtain a predicted value of θ_G assuming a working value for $\bar{c}$. **Figure 3** shows predictions for θ_G across 100 simulated datasets, where each subplot corresponds to different true values for $\bar{c}_1$ and θ_0. In each subplot, the boxplot corresponding to the true sensitivity is bolded. In all settings considered, the median predicted θ_G is near the true value of 0.5 when $\bar{c}_1$ is correctly specified. The sensitivity of predicted θ_G to values of $\bar{c}_1$ strongly depends on the population disease rate. For prevalence $< 1\%$, only very low sensitivities far from the true $\bar{c}_1$ produced predicted θ_G far from the true value of 0.5. Additionally, assuming sensitivities of 1 did not result in much bias in estimated θ_G on average (<0.05). In contrast, predicted sensitivities strongly varied across assumed $\bar{c}_1$ when the population disease rate was larger. Also, assuming sensitivities of 1 produced strong bias up to roughly 0.2 for true $\bar{c}_1 = 0.4$.

Part 1b: predicting θ_G with available θ_0: Suppose instead that $\theta_0^{(simple)}$ is not known and instead we have some sense of θ_0. In practice, we can treat θ_0 as an input value informed by some known population disease rate. **Figures S3-S6** show predictions for θ_G obtained by inverting (3) with respect to θ_G. These predictions depend on assumed values for $\bar{c}_1$ along with $\bar{r}$. These simulations again demonstrate an ability of our sensitivity analysis approach to recover the true θ_G on average when evaluated at the true values for $\bar{r}$ and $\bar{c}_1$.

Part 2: predicting θ_G with available $\theta_0^{(simple)}$: In this section, we obtain predictions for θ_G for each simulated data using (SuppEq. 2.7). **Figure 4** presents boxplots of the predicted θ_G across 500 simulated datasets in each simulation setting and across different assumed values for $\bar{c}_1$. We focus our attention on the boxplot of predicted θ_G values corresponding to the true $\bar{c}_1$ (denoted in bold) in each setting. We assume Z is not a direct driver of sampling or misclassification for all simulations except setting 6. Median predicted values of θ_G are 0.47, 0.47, 0.44, 0.47, 0.50, 0.49, 0.63, and 0.36 for the 8 simulation settings. In settings 7 and 8 where assumptions 1-4 are violated, the predicted θ_G values clearly miss the mark, with median biases of 26% and 28% relative to the truth. We are more concerned with relationships between Z and W and/or X, which are not directly addressed by assumptions 1-4. When Z is related to W (even strongly) or a direct driver of selection, the proposed methods do a good job of recovering θ_G in these simulations (settings 5 and 6). A more challenging setting occurs when Z is related to X. In setting 3, we see median biases of 11% relative to the truth, and this bias tends to increase as the association between Z and X gets stronger. When G and Z are independent, however, the relationship between Z and misclassification does not adversely impact the performance of our proposed methods (e.g. setting 4). These results demonstrate that, unsurprisingly, our proposed methods perform poorly when assumptions 3 and 4 are violated. Additionally, we can have some residual bias when Z is at least moderately related to misclassification in the setting where Z and G are related. Although not seen in our simulations, there may still be a potential for residual bias when Z is related to both selection and G.

5 | BIAS EXPLORATION IN THE MICHIGAN GENOMICS INITIATIVE

The Michigan Genomics Initiative (MGI) is a longitudinal biorepository effort linked to EHR within the University of Michigan health system, referred to as Michigan Medicine. The present analysis contains over 40,000 unrelated patients of recent European ancestry with matched genotype and phenotype information. Using these data, we are often interested in studying the association between disease and genetic factors G, adjusting for demographic factors such as age and gender along with principal components of the genotype data as a whole. Together, these factors compose Z.

Suppose we define our hypothetical "target population" as the general US population. **Figure 5** provides a rough visualization of the sampling/selection stages a patient goes through to be included in MGI. We assume that there is some unknown mechanism associated with each of these selection stages, which together form the composite "sampling mechanism" in Eq. 1. Overall, patients in MGI tend to be sicker and have a greater number of diagnoses than the general Michigan population and even the Michigan Medicine population.[1] Patient selection may likely be related to disease status, creating a large potential for bias due to patient selection in association models. EHR-derived disease status may also be misclassified, resulting in information bias.

Previous analyses have explored associations between single genetic markers and polygenic risk scores with cancer phenotypes for patients in MGI using simple analysis methods involving fitting a model for $D^*|G, Z, S = 1$.[1,21] These analyses ignore the potential bias induced by the sampling and phenotype misclassification mechanisms. We are interested in evaluating the potential impact of the sampling and misclassification mechanisms on inference. Below, we perform the proposed sensitivity analyses to study to robustness of SNP-phenotype and PRS-phenotype associations in MGI for different assumed sampling and phenotype misclassification mechanisms. Since we are interested in cancer phenotypes, which may only rarely be incorrectly diagnosed, we assume perfect specificity for these analyses.

For our proposed sensitivity analysis approach, X and W do not need to be specified explicitly; rather, we consider effects of misclassification and patient selection integrated over predictor components in X and W. Still, our analytic derivations hinge on our ability to satisfy assumptions 1-4, so hypothetical consideration of X and W is necessary. First, we clarify our notation. In our analyses, we will consider six cancer phenotypes for breast, prostate, bladder, and colorectal cancers along with melanoma and non-Hodgkins lymphoma. In each exploration, EHR-derived phenotypes are denoted by D^*, and we let D represent the true disease status. G either represents a single SNP or a polygenic risk score, and our analyses adjusted for patient age, gender, and the principal components of the genotype data (Z). For these data, we believe X primarily consists of factors related to each patient's observation process such as the number of doctor's visits and the length of the observation window. Other factors such as age and gender may also be related to misclassification. It is difficult to assess factors related to patient selection, but we expect patient selection into MGI to be broadly related to the patient's overall health status and demographic factors. We explore differences between MGI patients and Michigan Medicine and the US adult population in **Table S4**. We believe assumptions 1-3 are reasonably well satisfied for these analyses, and assumption 4 is reasonable when G represents a single genetic locus. When G represents a PRS, the independence between G and other diseases included in W may hold conditional on true disease status and demographic variables in Z.

5.1 | Association analysis with Polygenic Risk Scores

Previous work in Fritsche et al. (2018) explored associations between polygenic risk scores (PRS) and their corresponding EHR-derived phenotypes D^* in MGI for several cancer phenotypes.[21] They fit the following analysis model: $\text{logit}(P(D^*|G, Z, S = 1)) = \theta_0^{(simple)} + \theta_G^{(simple)} PRS + \theta_Z^{(simple)} Z$ where Z contained age, gender (when relevant), 4 principal components of the genotype data, and genotype batch information. D^* indicators were defined as 1 if the patient ever had a given disease diagnosis record in the EHR via ICD codes. We are interested in exploring the robustness of these *published summary results* to different sampling and misclassification mechanisms for six different cancer diagnoses. We treat the published PRS association summary statistics as our $\theta_G^{(simple)}$ values, and we use the corresponding Michigan prevalence rates reported in Beesley et al. (2018) to approximate θ_0 for each of the six cancers of interest.[1] These published PRS summary statistics and prevalences values are listed in **Table S3**. We use the upper and lower confidence interval limits for $\theta_G^{(simple)}$ to create an interval for θ_G.

Figure 6 shows the predicted interval for θ_G assuming different sampling ratios and sensitivities for each of the six selected cancers, where the vertical bars correspond to transformations of the 95% confidence intervals for standard analysis under different assumed values for $\bar{r}$ and $\bar{c}_1$. Estimates for cancer of the bladder, non-Hodgkins lymphoma, and colorectal cancer are all very robust to different values for the sampling ratio and sensitivity, where the predicted PRS log-odds ratios were never beyond the 95% confidence intervals for the standard analysis. There are two primary reasons for this phenomenon. Firstly, the population prevalences of these three cancers are all low (less than 5%). As shown previously, we expect to see less relative bias in this setting. Secondly, the estimated association between the PRS and the disease is small in all of these cases (log-odds ratio values less than 0.35). While the relative bias will not appreciably change as θ_G changes, the absolute bias will be small for small θ_G. For evaluated settings in these three cancers, absolute bias was always less than 0.10. In contrast, breast and prostate cancer have high prevalences (>10%), and the corresponding PRS associations are strong (log-odds ratios of 0.83 and 1.19 respectively). For these two PRS associations, our results suggest we may have appreciable relative bias if the sampling ratio is moderate to large and the disease outcome is at least moderately misclassified. In an extreme setting where the sensitivity is very low (e.g. 0.1) and the sampling fraction is fairly high (e.g. 10), we see predicted θ_G nearly doubles the observed $\theta_G^{(simple)}$ for both breast and prostate cancers. Notably, this corresponds to absolute biases of up to 1.35 on the log-odds ratio scale. In reality, we do not know the sampling ratio, and we explore plausible values for the sampling ratio for these cancers in MGI in **Figure S10**.

In all cases in **Figure 6**, the confidence interval for the PRS-phenotype association is far from zero, and adjustment for misclassification and sampling under our model would move estimated θ_G even farther from zero. Therefore, our general conclusions of a moderate or strong association between the PRS and the phenotype of interest remain robust across scenarios. However, we expect decreases in power for a test for $\theta_G^{(simple)}$ being nonzero when we have bias toward the null. We will evaluate properties of resulting tests and decision rules in future work exploring corrected estimation and inference techniques for θ_G.

5.2 | Genetic association analysis using individual markers

We perform a breast cancer GWAS using a cohort of over 40,000 unrelated patients in MGI of recent European descent. We fit the following logistic mixed model using the method in Zhou et al. (2018)[26]: $\text{logit}(P(D^*|G, Z, S = 1, \omega)) = \theta_0^{(simple)} + \theta_G^{(simple)} SNP + \theta_Z^{(simple)} Z + \omega$ where Z contained age, 4 principal components of the genotype data, and genotype batch information. Here, ω is a random effect term accounting for potential residual sample relatedness. Although not strictly the analysis model in 3.1, we will treat the resulting $\theta_G^{(simple)}$ as if it were from a standard logistic regression. This analysis was performed on a 10:1 matched subset of patients based on age and the first four principal components of the genome-wide data. The breast cancer phenotype D^* indicated whether the patient ever had a ICD-based breast cancer diagnosis recorded at Michigan Medicine, defined using phecodes based on ICD codes using the R package PheWAS.[4]

We first compare GWAS results in MGI ($\theta_G^{(simple)}$) with 563 reported associations from the NHGRI-EBI GWAS catalog (θ_G), which combines results from meta-analysis of the largest and highest quality studies (https://www.ebi.ac.uk/gwas/). We treat NHGRI-EBI GWAS catalog as a comparative gold standard. As shown in **Figure S7**, the association results using MGI data are generally similar to GWAS catalog results, but there are some specific SNPs for which the results differ. Overall, MGI results appear attenuated relative to the GWAS catalog results, with Lin's concordance correlation coefficient of 0.59 (95% CI: 0.52, 0.63). We focus on 6 individual loci for which the GWAS catalog estimate (θ_G) differs appreciably from the MGI estimate ($\theta_G^{(simple)}$). For this exploration, we fix $\theta_0 = \text{logit}(0.124)$, using the US female population lifetime rate of breast cancer.[1]

In **Figure S8**, we explore plausible values of θ_G in MGI across different potential sampling ratios and sensitivities. We obtain rough intervals for θ_G for fixed values of $\bar{r}$ and $\bar{c}_1$ by transforming the upper and lower 95% confidence interval limits of the standard MGI GWAS estimate for each locus. In **Supplementary Section S8**, we describe one approach for getting a sense of plausible values for the sampling ratio. As either the sensitivity or sampling ratio goes to 1, the interval shown in the figure gets closer to the 95% confidence interval for the corresponding GWAS estimate in MGI. We note that, in principle, we expect some bias in $\theta_G^{(simple)}$ when we have imperfect sensitivity, but we see very little absolute bias (e.g. < 0.03) when $\bar{r}$ is small (e.g. < 2). This is primarily due to the small estimates for $\theta_G^{(simple)}$. Bias in $\theta_G^{(simple)}$ is expected to be proportional to the true value of θ_G, and for small θ_G, we may not see much absolute bias (e.g. < 0.02) even for very low sensitivity when the sampling ratio is near 1. As the sampling ratio increases, however, the relative and absolute biases increase, and the predicted θ_G becomes more extreme. For example, $\bar{r} = 20$ and $\bar{c}_1 = 0.1$ results in a predicted θ_G of 0.66 from a standard analysis value of 0.13 (95% CI: 0.07, 0.20). For the sake of comparison, we present an example for which the association in MGI and the GWAS catalog are nearly identical in the **Figure S9**. These results suggest that even GWAS results can potentially be impacted by sampling and misclassification.

6 | DISCUSSION

The proposed conceptual and sensitivity analysis framework allows us to explore the amount of bias we might expect in large association study results when we ignore issues of disease status misclassification and sampling related to disease and patient characteristics, as is often done in association studies using EHR-derived outcomes. Previous literature generally suggests that we may usually expect little bias in genotype-phenotype associations, and our statistical results lend credence to this belief when the disease of interest has low prevalence, say less than 10%, and sampling does not depend on the underlying disease status. When the disease of interest has higher prevalence in the population or the phenotype misclassification rates are higher, however, the sampling mechanism can have a substantial impact in biasing association results. Without any misclassification of the outcome, we do not expect much bias in log-odds ratio association parameters unless sampling is associated with both the underlying disease status and the covariate of interest given the adjustment factors. This is a property of the logistic regression modeling assumptions as has been explored in detail in the literature on secondary analysis of case-control sampled data and in **Supplementary Section S5**, and modeling disease using different link functions may produce slightly different results.[23,24]

We consider settings in which misclassification has perfect or imperfect specificity. Assuming perfect specificity (no overreporting of disease), we observe a higher potential for bias toward the null in the simple analysis for diseases with higher population disease prevalence. This assumption may be reasonable for many EHR-derived phenotypes, where the large part of misclassification is expected to be a result of underreporting of disease. In the setting of imperfect specificity, which may occasionally arise for diseases that are difficult to diagnose such as psychiatric disorders, we may have bias either toward or away from the null, and the general relationship between disease prevalence and bias is much less clear. The current exploration focuses on bias in effect estimates, but we may also be interested in p-values and hypothesis testing. We expect effect estimates biased toward the null to result in a loss of statistical power. We will explore the issues of Type I error and power in detail in a follow-up paper focused on estimation under the proposed model.

One advantage of the proposed modeling framework is that it does not require parametric modeling assumptions to be made for the sampling mechanism or the observation mechanisms (related to sensitivity and specificity). Additionally, we do not require factors driving misclassification and patient selection to be entirely understood or even observed. Instead, our results are guided by independence assumptions made on the relationships between drivers of these various mechanisms, and our proposed analysis approach involves terms integrating over these unknown factors driving selection and misclassification. A strong assumption made in the course of this paper is that the predictors related to sampling that are not included in the simple analysis model are independent of the genetic information of interest, conditional on the true disease status and adjustment factors Z. Since Z often contains age, gender, and several principal components of the genotype information, this may often be a reasonable assumption for EHR data. A challenging setting, however, is one in which sampling is related to a secondary disease $\boldsymbol{D'}$ that is *independently* related to $\boldsymbol{G}$ even adjusting for $\boldsymbol{D}$ and Z (perhaps due to pleiotropy). Our results can only be applied in this setting when simple analysis adjusts for any secondary diseases that are independently related to $\boldsymbol{G}$. Secondary diseases independent of $\boldsymbol{G}$, however, do not need to be adjusted for.

When our model assumptions are satisfied, the proposed methods can often recover the true relationship between genetic factors $\boldsymbol{G}$ and disease status $\boldsymbol{D}$ for known sensitivity and sampling mechanisms. One setting in which the proposed methods may struggle is when genetic factors $\boldsymbol{G}$ and adjustment factors Z are associated and Z is associated with factors driving phenotype misclassification, X. In this case, the first order Taylor series approximations used to derive the bias expressions may be inadequate, resulting in residual bias in the predicted parameter of interest. For example, suppose that body mass index (BMI) is included as an adjustment factor in Z and is also at least moderately related to whether the disease is observed. If BMI is also related to the genetic factor $\boldsymbol{G}$, the proposed methods can run into trouble. However, the relationship between Z and X only causes complications when $\boldsymbol{G}$ and Z are associated. When Z contains age, gender, and the principal components of the genotype data, it may be reasonable to assume that $\boldsymbol{G}$ is independent or only very weakly related to Z, particularly when analysis focuses on patients of recent European descent as in the MGI example. For many EHR-based data analyses, therefore, we do not expect this potential for residual bias to be of much concern. In the course of our statistical development, we also assume that Z is mean-centered, but we could equivalently assume Z is median-centered for highly skewed data. A challenging setting occurs when Z is highly variable. In this case, we hypothesize that the the first order Taylor series approximations used in our statistical development may be insufficient, and additional orders of approximation may be needed.

In this paper, we focus on exploring the potential impact of selection and information biases on a single genotype-phenotype association, but we are often interested in studying many genotype-phenotype associations. A natural question is the extent at which this bias impacts comparison across naíve parameter estimates in a large association study. In the case of GWAS, we perform association tests across many genotypes for a single phenotype. Here, the sampling ratio, sensitivity, and specificity are primarily properties of the particular disease we are interested in, and we do not expect these values to change much across the various association tests. In contrast, association tests in a PheWAS consider many different disease phenotypes. In this setting, we expect the sensitivity, specificity, and sampling ratios to differ across phenotypes, and accounting for differential bias toward the null across the various association tests may be of particular importance. There may be an opportunity to incorporate additional information such as the genetic architecture and disease heritability into the assessment of comparative bias.

The proposed analytic framework can be useful for guiding analyses exploring sensitivity to violations of the common implicit assumptions of no outcome misclassification and ignorable sampling in EHR-based association studies *using summary results*. Individual-level data are not required for our method to be applied. Our proposed approach is not intended for parameter estimation or to correct for bias; rather, it is meant to be used as a tool for evaluating model robustness to these various biases. In future work, we will develop statistical methodology to perform parameter *estimation* and characterize uncertainty under the proposed conceptual model. As part of the current work, we have developed an online tool called SAMBA-EHR (**S**ampling **A**nd **M**isclassification **B**ias **A**nalysis for EHR-based association studies) available at http://shiny.sph.umich.edu/SAMBA-EHR/. This will allow the proposed methods to be easily implemented in practice as a part of routine sensitivity explorations for association studies using EHR-derived outcomes.

7 | ACKNOWLEDGMENTS

The authors would like to thank Chad Brummet, Goncalo Abecasis, and Sachin Kheterpal along with the large group of collaborators at Michigan Genomics Initiative and also the MGI participants for generously donating their biosamples for research. This work was supported by NSF grant DMS1712933, The University of Michigan Comprehensive Cancer Center core grant supplement 5P30-CA-046592, and The University of Michigan precision health award U063790. The authors acknowledge the University of Michigan Medical School Central Biorepository for providing biospecimen storage, management, and distribution services in support of the research reported in this publication. *Dr. Lauren J. Beesley* wrote the manuscript and developed the methodology and the online SAMBA-EHR application. *Dr. Lars G. Fritsche* helped edit the paper and facilitated data acquisition and interpretation. *Dr. Bhramar Mukherjee* wrote the manuscript and provided key guidance on the methodology and manuscript. *Conflict of Interest*: None declared. *Data Availability*: The Michigan Genomics Initiative data used in this paper may be available for research purposes after IRB approval. More information available at https://precisionhealth.umich.edu/michigangenomics/.

References

1. Beesley LJ, Salvatore M, Fritsche LG, et al. The Emerging Landscape of Epidemiological Research Based on Biobanks Linked to Electronic Health Records. *Preprints.org* 2018: 1–35.

2. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010; 26(9): 1205–1210.

3. Wolford BN, Willer CJ, Surakka I. Electronic health records: The next wave of complex disease genetics. *Human Molecular Genetics* 2018; 27(R1): R14–R21.

4. Carroll RJ, Bastarache L, Denny JC. R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 2014; 30(16): 2375–2376.

5. Castro VM, Minnier J, Murphy SN, et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry* 2015; 172(4): 363–372.

6. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nature Reviews Genetics* 2016; 17(3): 129–145.

7. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* 2015; 350: 1–5.

8. Castro V, Shen Y, Yu S, et al. Identification of subjects with polycystic ovary syndrome using electronic health records. *Reproductive Biology and Endocrinology* 2015; 13(116): 1–8.

9. Yu C, Guo Y, Bian Z, et al. Association of low-activity ALDH2 and alcohol consumption with risk of esophageal cancer in Chinese adults: A population-based cohort study. *International Journal of Cancer* 2018: 1–28.

10. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* 2018; 172(5): 1122–1131.

11. Liao KP, Sparks JA, Hejblum BP, et al. Phenome-Wide Association Study of Autoantibodies to Citrullinated and Noncitrullinated Epitopes in Rheumatoid Arthritis. *Arthritis and Rheumatology* 2017; 69(4): 742–749.

12. Hubbard RA, Benjamin-Johnson R, Onega T, Smith-Bindman R, Zhu W, Fenton JJ. Classification accuracy of claims-based methods for identifying providers failing to meet performance targets. *Statistics in Medicine* 2015; 34(1): 93–105.

13. 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. *Journal of the American Medical Informatics Association* 2018; 25(3): 345–352.

14. Wang L, Damrauer SM, Zhang H, et al. Phenotype validation in electronic health records based genetic association studies. *Genet Epidemiol.* 2017; 41(8): 790–800.

15. Duffy SW, Warwick J, Williams AR, et al. A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology and Community Health* 2004; 58(8): 712–717.

16. Sinnott JA, Dai W, Liao KP, et al. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics* 2014; 133(11): 1369–1382.

17. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology* 2016; 184(11): 847–855.

18. Phelan M, Bhavsar N, Goldstein BA. Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 2017; 5(1): 22.

19. Smith GD, Ebrahim S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease?. *International Journal of Epidemiology* 2003; 32(1): 1–22.

20. Avery CL, Monda KL, North KE. Genetic association studies and the effect of misclassification and selection bias in putative confounders. *BMC Proceedings* 2009; 3(Suppl 7): S48.

21. 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. *The American Journal of Human Genetics* 2018; 102(6): 1–14.

22. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 2018; 19: 581–590.

23. Jiang Y, Scott AJ, Wild CJ. Secondary analysis of case-control data. *Statistics in Medicine* 2006; 25(8): 1323–1339.

24. Tchetgen Tchetgen EJ. A general regression framework for a secondary outcome in case – control studies. *Biostatistics* 2014; 15(1): 117–128.

25. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013; 9(3): 1003348.

26. Zhou X, Douglas IJ, Shen R, Bate A. Signal Detection for Recently Approved Products: Adapting and Evaluating Self-Controlled Case Series Method Using a US Claims and UK Electronic Medical Records Database. *Drug Safety* 2018; 41(5): 523–536. doi: 10.1007/s40264-017-0626-y

Figure 1 *Diagram of the assumed data structure*

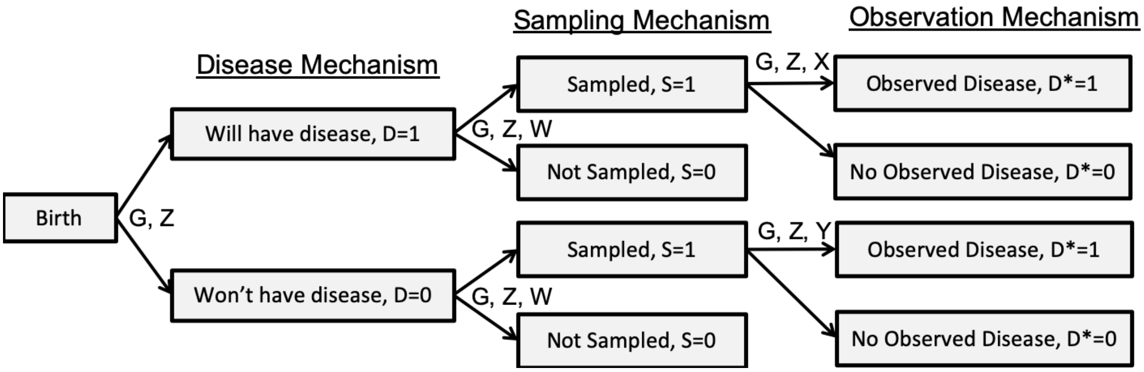
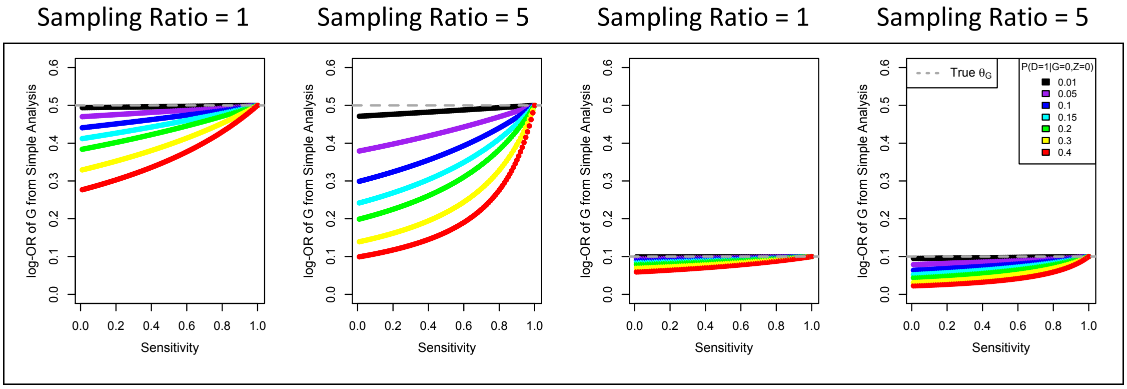
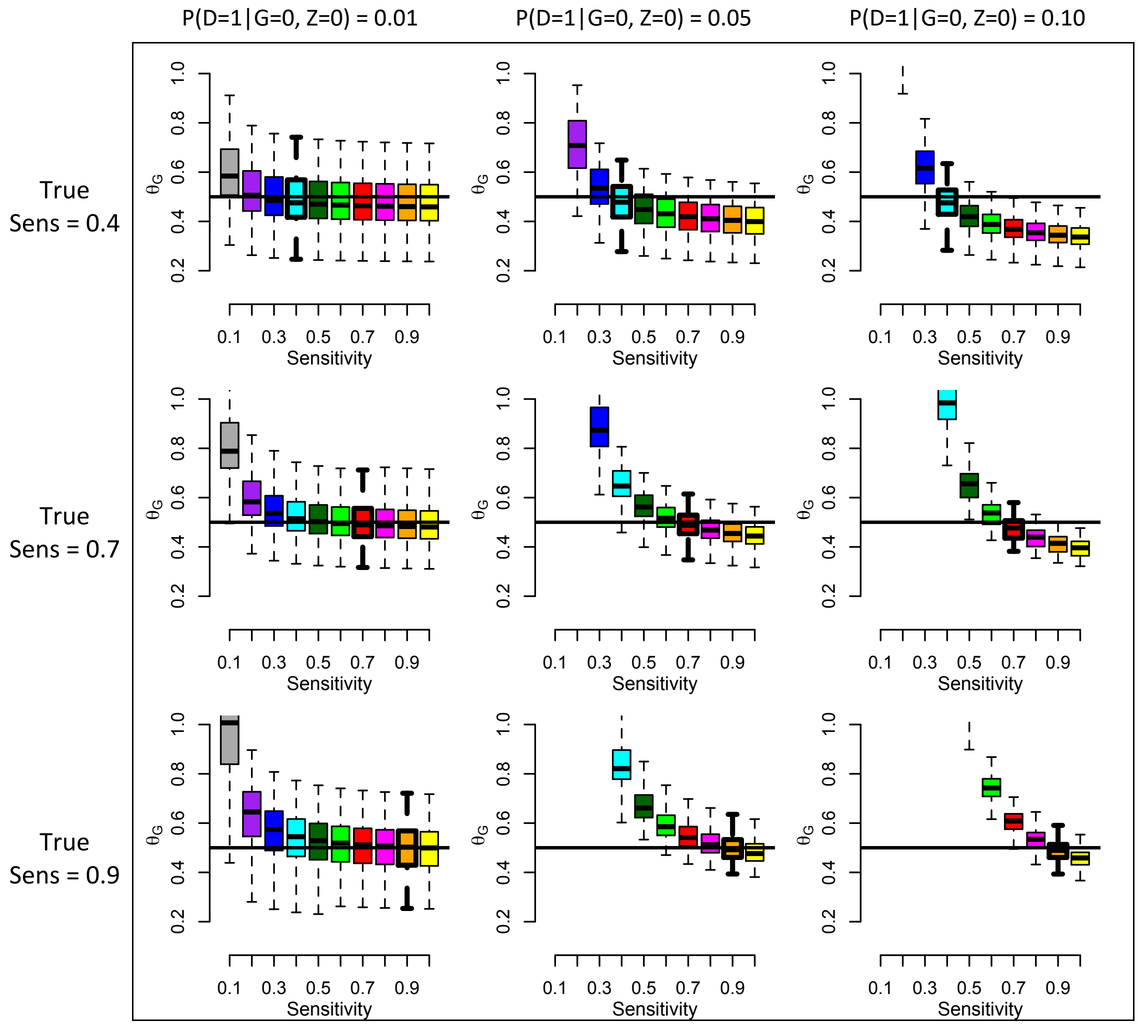


Figure 2 *Standard analysis log-odds ratio for G when we model $D^*|G, Z, S = 1$**



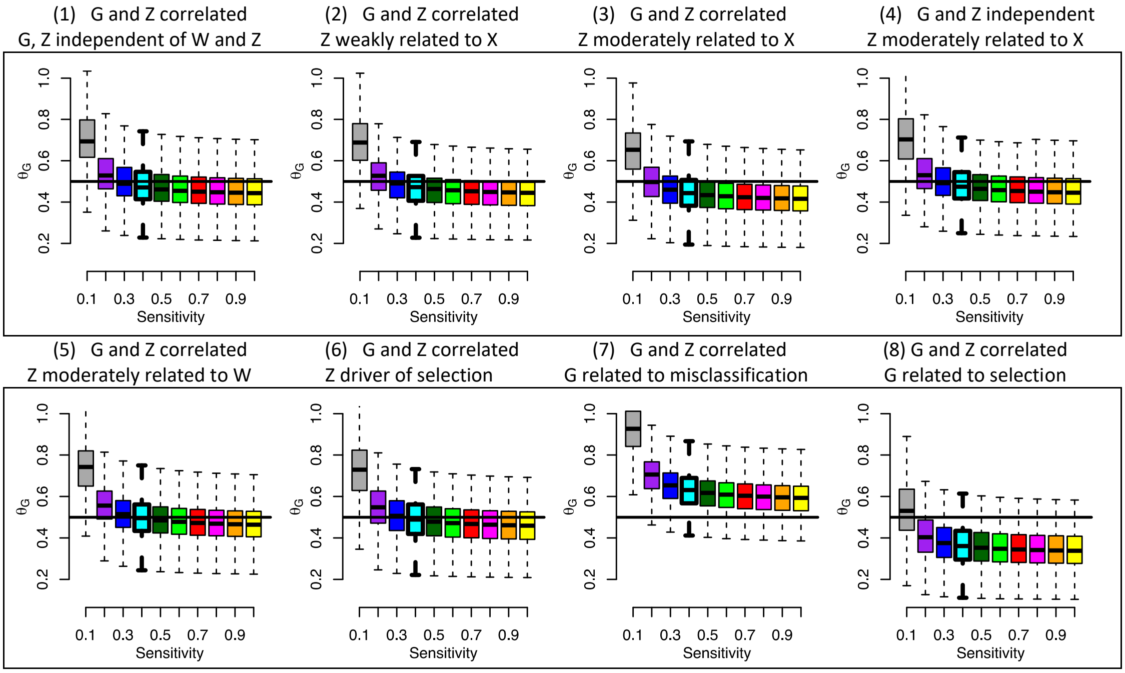
* This figure presents the analytical prediction for $\theta_G^{(simple)}$ (y-axis) from (3) for a SNP with population MAF of 0.2 as a function of true sensitivity $\bar{c}$ (x-axis). The analytical expression is evaluated across multiple population disease rates, calculated as expit(θ_0). Figure panels correspond to different values for the true sampling ratio, $\bar{r}$, and the true association between G and D, denoted θ_G and plotted with the horizontal dotted line.

Figure 3 *Predicted θ_G for SNP across different assumed sensitivities using available $\theta_0^{(simple)}$* *



* Each panel of this figure summarizes the predicted θ_G values for a SNP obtained using expression (5) across 100 different simulated datasets. Each boxplot corresponds to the predicted θ_G values (y-axis) for a different working value of $\bar{c}_1$ (x-axis). Each panel in this figure corresponds to one of nine simulation settings corresponding to different *true* values for $\bar{c}$ (denoted by the bolded boxplot in each panel) and different values for $P(D = 1|G = 0, Z = 0)$ across the columns. In each panel, the horizontal line indicates the true value of θ_G, 0.5. Across all simulation settings, data are simulated to have true $\bar{r} = 5$ and MAF of 0.2. For these simulated datasets, Z, W, and X are empty.

Figure 4 *Predicted θ_G for PRS under more complicated covariate relationships*



* Each panel of this figure summarizes the predicted θ_G values for a PRS obtained using using expression (SuppEq. 2.7) across 100 different simulated datasets. Each boxplot corresponds to the predicted θ_G values (y-axis) for a different working value of $\bar{c}_1$ (x-axis). Each panel in this figure corresponds to one of eight simulation settings corresponding to different associations between simulated G, Z, X_1, X_2, and W. These variables are all multivariate normal with differing pairwise correlations. Unless otherwise specified, pairwise correlations were set to zero. In all but setting 4, G and Z have a correlation of 0.2. We assume pairwise correlations of 0.5 between Z and X_1, Z and W, and G and W for settings 3/4, setting 7, and setting 8 respectively. Setting 6 explores a strong 0.9 correlation between Z and W, and setting 2 explores a weak 0.1 correlation between Z and X_1. The final two simulation settings correspond to settings where assumptions 3 and 4 are violated. In each panel, the horizontal line indicates the true value of θ_G, 0.5. The simulated marginal disease rate was roughly 5%, and the simulated marginal sensitivity was roughly 0.4.

Figure 5 *Schematic representation of the sampling stages and phenotype ascertainment in MGI with respect to a US target population**

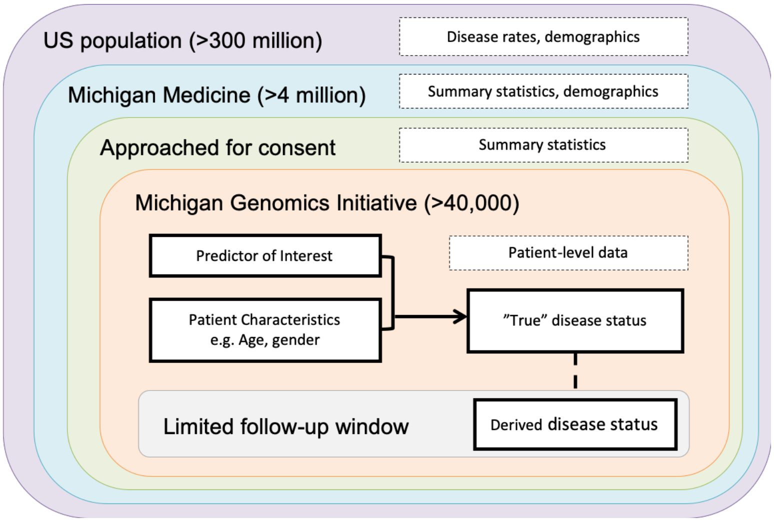
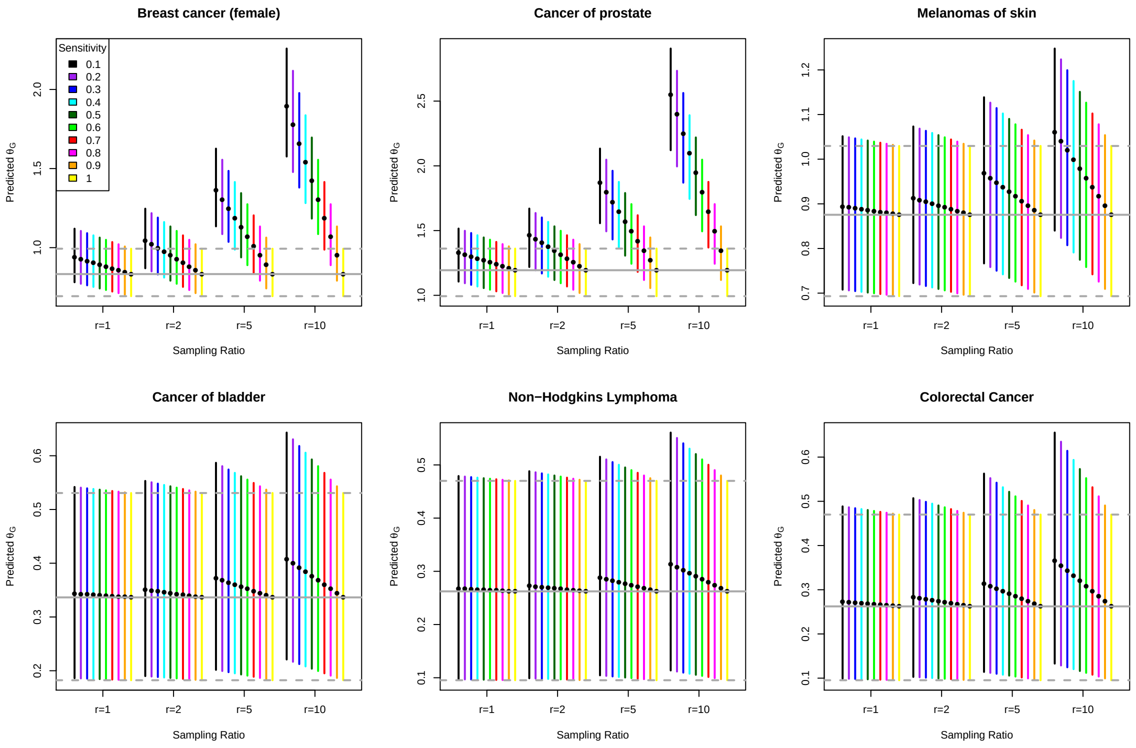


Figure 6 *Sensitivity Analysis for PRS-phenotype associations for six different cancers**



* This figure presents predicted confidence intervals for θ_G (y-axis) obtained by transforming the published standard PRS-disease odds ratio and upper/lower 95% interval limits using $\theta_G \approx \left[\frac{1}{e^{\theta_0}(1-\bar{c}_1)\bar{r}+1}\right]^{-1}\theta_G^{(simple)}$. This expression is a function of $\bar{c}_1$ and $\bar{r}$, and we plot the predicted intervals across different working values for $\bar{c}_1$ and $\bar{r}$ along the x-axis of each panel. Vertical bars correspond to the transformed interval for θ_G assuming the corresponding value for $\bar{r}$ and $\bar{c}_1$. This transformation is a function of θ_0, which was determined by transforming population disease rates as shown in **Table S3**. Each panel corresponds to the PRS-disease association for a different cancer of interest as reported in Fritsche et al. (2018). Horizontal lines correspond to the standard point estimate and 95% confidence intervals.

SCHOOL OF PUBLIC HEALTH
BIOSTATISTICS
UNIVERSITY OF MICHIGAN

February 14th, 2020

Dear Dr. Greenhouse,

We would like to thank the editors and the reviewer for their constructive feedback on our submission of the revised manuscript SIM-19-0350.R1, titled "*An Analytic Framework for Exploring Sampling and Observation Process Biases in Genome and Phenome-wide Association Studies using Electronic Health Records*" by Lauren J Beesley, Lars G Fritsche and Bhramar Mukherjee.

We are delighted to receive a conditional acceptance. In this document, we highlight how we have addressed the remaining concerns of the reviewer and the associate editor.

Thank you for your consideration of our revised manuscript. We hope you find the revised manuscript suitable for publication in *Statistics in Medicine.*

Sincerely,
Dr. Lauren J Beesley
Post-doctoral Research Fellow
Department of Biostatistics, University of Michigan

Response to the Reviewer 1

SPECIFIC POINTS

1. I think it would be easier to follow the derivations and assumptions if you expressed models (1) and later expressions in terms of factors Z, W, X and Y, where W, X and Y are not in Z; that is, they are what you now call $W_{\dagger}$ etc.

   Response: We agree with the reviewer that the previous notation was hard to follow at times. In response, we have updated the notation as suggested by the reviewer. Please see new notation as described on page 4 of the manuscript.

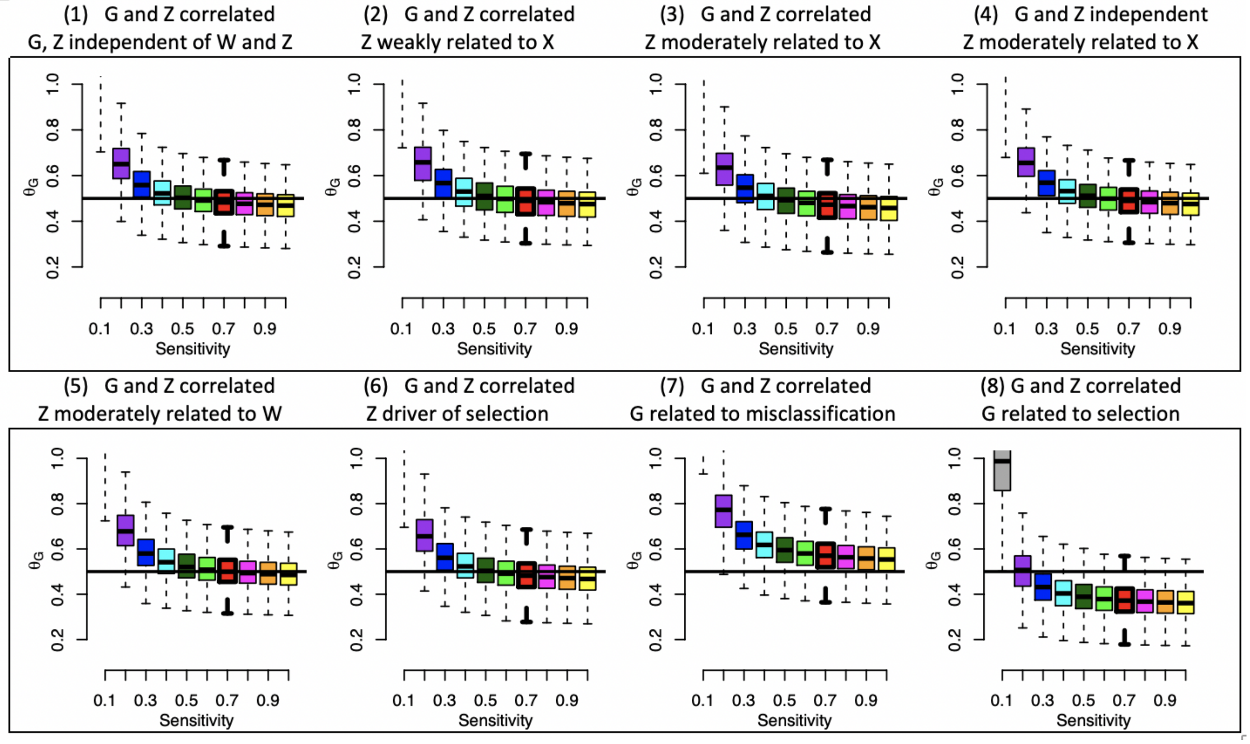
2. Section 4: I find the part 1 simulation scenario to be of little interest; I cannot think of settings where it would be an adequate framework. Part 2 is more reasonable, though still quite simple. In Figure 4, the true “average” sensitivities for each panel are, as I understand it, equal to 0.4, which is the bold value in each panel. This seems low; what is the motivation? It would also help if you gave this value explicitly in section 4.

   Response: The simulations in part 1, while simple, do demonstrate that the proposed sensitivity analysis approach performs well under these settings where assumptions under which we derive the results hold. Since this approach is based on Taylor series approximations, such good performance was not guaranteed, and we wanted to demonstrate the adequacy of the approximation in simple cases prior to moving on to more complicated situations in part 2.

   In reference to the choice of 0.4 as the sensitivity in **Figure 4**, we chose this lower value for two main reasons. First, the impact of ignoring sensitivity is larger for smaller sensitivities, and we felt that a lower sensitivity would demonstrate the properties of interest in **Figure 4** better than a very high sensitivity. Secondly, while this sensitivity may seem low initially, we believe that this could actually be a plausible value in practice with EHR data. In this paper, we are considering D defined not based on the medical record itself but based on a patient’s true disease state up to the current time. Previous work has explored the rates of misclassification for particular ICD code-based phenotypes relative to information in the greater EHR (e.g. doctor’s notes, lab values, etc), but the actual level of misclassification relative to true disease state is comparatively unexplored. For patients with short follow-up in a given EHR (as is the case for many patients in MGI), it seems reasonable that the MGI-based phenotyping would miss a large number of diseases a patient truly has or had. Therefore, a sensitivity of 0.4 on average is within the realm of possibility for certain datasets such as MGI.

   We have added text in **Section 4** to include this 0.4 sensitivity value explicitly. However, we have also repeated the simulations presented in **Figure 4** using a marginal sensitivity of roughly 0.7, resulting in **Revision Figure 1 (see below)**. The results are very similar to those under marginal sensitivity of 0.4. This figure is included in the supplementary material.

Revision Figure 1: Predicted θ_G for PRS under more complicated covariate relationships for marginal sensitivity of 0.7


(1) G and Z correlated
G, Z independent of W and Z

(2) G and Z correlated
Z weakly related to X

(3) G and Z correlated
Z moderately related to X

(4) G and Z independent
Z moderately related to X

(5) G and Z correlated
Z moderately related to W

(6) G and Z correlated
Z driver of selection

(7) G and Z correlated
G related to misclassification

(8) G and Z correlated
G related to selection

3. (in reference to Section 5) (a) What exactly is Michigan Medicine? (b) I am nitpicking a bit, but I would not call what you use “mechanisms” (line 6 of paragraph 2); they are models but I don’t see any underlying mechanism motivating them. Similarly, I would not say Figure 5 shows mechanisms; it just shows the different “populations”. (c) As per point 1, it would be clearer if you first said what you include in Z and then discuss additional factors that might affect selection and misclassification. (d) In Figure 6 I am not sure where the vertical bars come from. Are you “inverting” the naive confidence intervals to give an interval for θ_G? This seems circular. (e) I wonder what your conceptual “target” population is. Is it the general Michigan population? It seems to me that the target population could be almost anything, with selection model parameters differing across target populations. This seems crucial in order to relate results to other published results but at the same time, we know that conceptual “true” θ_G values will vary across populations because of heterogeneity and other factors.

Response: Thank you for these important comments. Please note our response to each part of your comment.

(a) We have updated the text in **Section 5** to clarify that “Michigan Medicine” refers to the University of Michigan health system. The text now reads: “The Michigan Genomics Initiative (MGI) is a longitudinal biorepository effort linked to

EHR within The University of Michigan health system, referred to as Michigan Medicine."

(b) We have updated the text to clarify what we mean by "mechanism" and how it relates to **Figure 5**. We included the following text:

"Suppose we define our hypothetical ``target population'' as the general US population. **Figure 5** provides a rough visualization of sampling/selection stages a patient goes through to be included in MGI. We assume that there is some unknown mechanism associated with each of these selection stages, which together form the composite ``sampling mechanism'' in Eq.1."

(c) We have updated the text as follows:

"The Michigan Genomics Initiative (MGI) is a longitudinal biorepository effort linked to EHR within The University of Michigan health system, referred to as Michigan Medicine. The present analysis contains over 40,000 unrelated patients of recent European ancestry with matched genotype and phenotype information. Using these data, we are often interested in studying the association between disease and genetic factors G, adjusting for demographic factors such as age and gender along with principal components of the genotype data as a whole. Together, these factors compose Z.

Suppose we define our hypothetical ``target population'' as the general US population. **Figure 5** provides a rough visualization of sampling/selection stages a patient goes through to be included in MGI. (continues...)"

(d) We have clarified in the text related to **Figure 6** where the vertical bars come from. They are obtained by applying our sensitivity analysis inversion formula to the MLE and to the 95% confidence interval limits from the naive analysis. The updated text is included here:

"**Figure 6** shows the predicted interval for θ_G assuming different sampling ratios and sensitivities for each of the six selected cancers, where the vertical bars correspond to transformations of the 95% confidence intervals for standard analysis under different assumed values for r-bar and c1-bar"

(e) We agree that the concept of target population is not specific. This was left intentionally vague for two reasons. Firstly, we wanted our method to be easily applicable to many different data analyses, which would have many different potential target populations. For example, it could be the adult population from the catchment area for the health system, the State of Michigan, the midwest region or United States. Generally, we view the target population as the population about which we want to make inferential statements based on our sample and our analysis. Secondly, even with a well-defined target population, the sampling ratio cannot be easily estimated. In the supplementary materials, we do provide a formula for obtaining a rough guess for the sampling ratio based on the observed disease prevalence in the defined target population (SuppEq. 8.11). In general, we view the sampling ratio as an unknown input parameter to be used to conceptualize how different the observed and target populations would need to be to get different levels of bias in our point estimates. Therefore, the target population is a hypothetical concept and the practical application of the proposed sensitivity analysis methods allow for the possibility of our “limited knowledge” regarding the specific characteristics of the target population and how it relates to the sample.

4. (in reference to Table S3) (a) It would be good to give the θ_G values as well as the exp(θ_G) values shown here, since they are prominent in Figure 6. (b) What are the units for prevalence and disease rates? (c) I assume the same covariates Z were used in Beesley et al. (2018) as here.

Response: (a) We have added a column to **Table S3** with the corresponding θ_G values as suggested.

(b) We thank the reviewer for pointing out this omission. Both the prevalence and lifetime disease rates are expressed as percentages for MGI and the target US population. We have updated **Table S3** and text accordingly. Below, we include the updated **Table S3**.

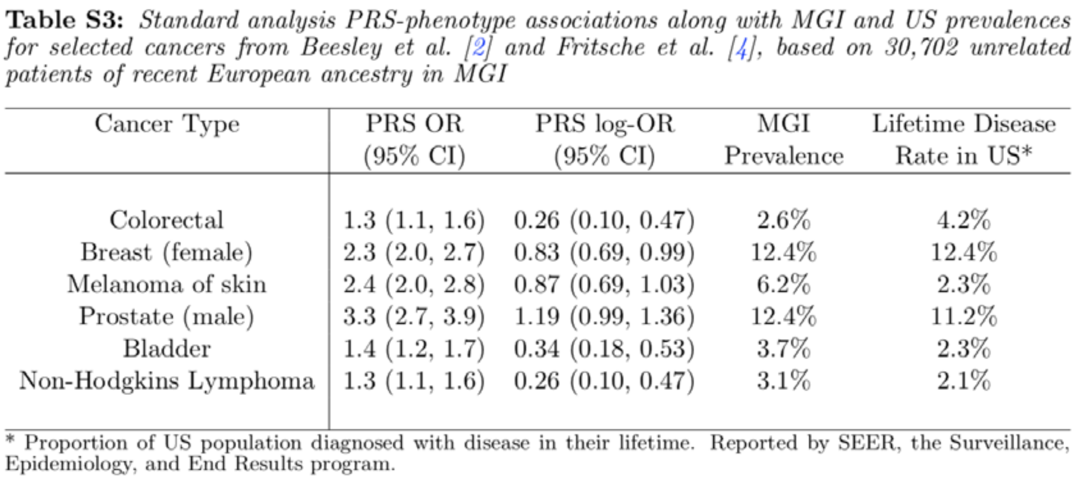


Table S3: *Standard analysis PRS-phenotype associations along with MGI and US prevalences for selected cancers from Beesley et al. [2] and Fritsche et al. [4], based on 30,702 unrelated patients of recent European ancestry in MGI*

Cancer Type	PRS OR (95% CI)	PRS log-OR (95% CI)	MGI Prevalence	Lifetime Disease Rate in US*
Colorectal	1.3 (1.1, 1.6)	0.26 (0.10, 0.47)	2.6%	4.2%
Breast (female)	2.3 (2.0, 2.7)	0.83 (0.69, 0.99)	12.4%	12.4%
Melanoma of skin	2.4 (2.0, 2.8)	0.87 (0.69, 1.03)	6.2%	2.3%
Prostate (male)	3.3 (2.7, 3.9)	1.19 (0.99, 1.36)	12.4%	11.2%
Bladder	1.4 (1.2, 1.7)	0.34 (0.18, 0.53)	3.7%	2.3%
Non-Hodgkins Lymphoma	1.3 (1.1, 1.6)	0.26 (0.10, 0.47)	3.1%	2.1%

* Proportion of US population diagnosed with disease in their lifetime. Reported by SEER, the Surveillance, Epidemiology, and End Results program.

(c) Beesley et al. (2018) provides the MGI prevalence and the lifetime disease rates, not the PRS associations. All PRS associations were originally reported in Fritsche et al. (2017).

Response to Associate Editor

My main remaining concerns are clarity of some terminology, and accessibility, ie. Is the presentation going to be accessible to a potential user of the proposed approach and SHINY APP? The concept and operationalization of “target population” seems particularly challenging.

1. I wonder whether the MGI study could be framed a bit more as a case study to help a potential user understand how to formulate their own application of the methods. At present, the reader may not begin to see this until the illustration and the discussion. My suggestion is to expand the last sentence in the introduction into a paragraph that lets the reader know that practical aspects will be addressed later – this could include brief background on the MGI setting, and highlight the motivation for the analysis that will be presented and discussed later - particularly “robustness of …” (page 24, line45 / page 25, line 16) and considerations such as formulation of the target population.

   Response: We agree that better highlighting MGI as a case study may help the reader understand how to apply our methods. We have updated the introduction section to discuss MGI in more detail. This text has been pasted below:

   “We then apply our proposed methods to data from The Michigan Genomics Initiative (MGI), a longitudinal biorepository effort within The University of Michigan health system with linked genotype and EHR information for over 40,000 patients. The patients were recruited in anesthesiology clinic while waiting for a surgery/diagnostic procedure. Using these data, we are often interested in studying the association between disease and genetic factors, adjusting for demographics and other patient characteristics. However, our EHR-derived disease status may have substantial misclassification relative to patients' true disease status. Additionally, we are often interested in making statements about external target populations such as the US population, and the MGI patient pool may be poorly representative of this target population. When ignored, these factors can create a large degree of bias in our association analyses of interest. We apply our proposed approach to explore the potential degree of bias in two example genetic association studies in MGI, which can serve as a tutorial for how the proposed methods can inform bias exploration after standard association analysis.”

2. Keeping in mind that all SIM readers are not familiar with typical approaches to genetic analysis, comments on use of terminology:

   (a) Practically “sampling mechanism” is “sample selection” since there is no explicit probability sampling – before the end of the intro, I suggest to state this for the reader, before adopting the “sampling” terminology in what follows in the model structure.

   Response: We agree with the associate editor that the term “sample selection” could be confusing in the case of non-probability sampling. We have added text to **Section 2** to clarify what by “sample selection” and “sampling mechanism”, we mean the theoretical unknown probability mechanism governing patient inclusion in the EHR and in turn analytic sample. The added text is included here:

“In addition to bias due to phenotype misclassification, the mechanism by which units in the population are included in the EHR dataset can sometimes result in biased inference when not handled appropriately. Complex sampling designs in an epidemiologic study can be addressed using survey design techniques if the sampling strategy is known. However, the probability mechanism for inclusion of a person into a biorepository is not *a priori* fixed or defined. For convenience, we will use terms such as ``sampling" and "selection mechanism" to describe this patient inclusion process, but it should be understood that this process is complicated and not well-characterized. Interactions with the healthcare system are generated by the patient, and it can be difficult to understand the mechanism driving sampling as well as self-selection for donating biosamples, which may be related to a broad spectrum of patient factors including overall health and demographic characteristics (continues…)”

(b) Terminology of “sensitivity analysis” and “tuning parameter” may be potentially confusing to some readers. At least in this paper, “Sensitivity analysis” is actually “Prediction of bias in genetic association estimates” – so it would be more informative to title section 3.3 in this way. In many other contexts, “tuning parameter” refers to an optimization parameter based on data – I’m not sure what would be better here, descriptively it is an parameter for population disease prevalence

Response: We agree that the “tuning parameter” terminology may be confusing. Although we do provide a strategy for estimating the sampling ratio in a rough sense, we still view these quantities as essentially unknown. We expect users to consider many plausible values for the sampling ratio and sensitivity based on their scientific problem. In this way, they are somehow informed by the data and the problem at hand, but in a loose way. However, to better communicate the role of these parameters, we now refer to them as “input values”, which we hope will clarify that these parameters are not part of an optimization.

Additionally, we have updated the title of **Section 3.3** to read “Sensitivity analysis of potential bias in genetic association estimates”, which we believe better reflects the uncertainty inherent in these data explorations.

(c) In description of genetic association analysis, are “naïve”, “standard”, and “summary” intended to be interchangeable? Does it refer specifically to a case-control design or a cohort design (ie. How does observation period enter in)? It would be helpful to the reader to have one clear definition if possible.

Response: In terms of “naive” and “standard” analysis, we are referring to estimation under the analysis model in (2). In this paper, we refer to existing summary statistics (e.g. point estimates and 95% confidence intervals) from “naive” or “standard” model fits that one can obtain from previous literature. To help clear up this confusion, we have updated this manuscript to use the term “standard” throughout and avoid the term “naive.” We have clarified our use of “standard analysis” in our discussion of the “analysis model” in **Section 3.1**.

In reference to the study design used for the naive analysis (case-control, cohort, etc.), we do not specify a particular design, because our statistical framework incorporates the sampling mechanism of the analytical data in the definition of the sampling ratio. Recall, the sampling ratio is the ratio of the average sampling probabilities in the diseased and non-diseased patients. This ratio will naturally be impacted by the design used for the naive analysis. Other factors such as observation period would also impact this term.

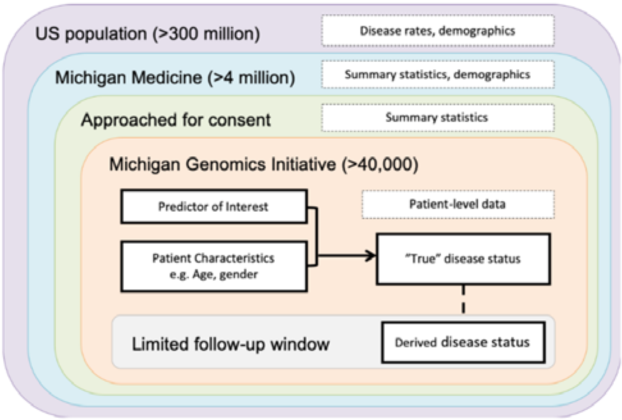
Now, the sampling ratio itself is defined in terms of some “target population.” In this manuscript, the definition of “target population” is not specific. This was left intentionally vague for two reasons. First, we wanted our method to be easily applicable to many different data analyses, which would have many different potential target populations. For example, it could be the adult population from the catchment area for the health system, the State of Michigan, the midwest region or United States. Generally, we view the target population as the population about which we want to make inferential statements based on our sample and our analysis. Second, even with a well-defined target population, the sampling ratio cannot be easily estimated. In the supplementary materials, we do provide a formula for obtaining a rough guess for the sampling ratio based on the observed disease prevalence in the defined target population (SuppEq. 8.11). In general, we view the sampling ratio as an unknown input parameter to be used to conceptualize how different the observed and target populations would need to be to get different levels of bias in our point estimates. Therefore, the target population is a hypothetical concept and the practical application of the proposed sensitivity analysis methods allow for the possibility of our “limited knowledge” regarding the target population and how it relates to the sample.

(d) There is potential for confusion between “target model” and “target population”, especially in formulation of the latter in the context of the MGI example. Figure 5 which has been added in this revision and referred to in section 5, is concise and does illustrate the nested nature of the various populations and sample selection, but doesn’t capture the relative population sizes (which are however evident in Table S4) or the selection on ethnicity.

Response: We agree that these terms are confusing. We have updated the manuscript to refer to the “target” model as the “true” model. By “true” model, we mean the model that we want to fit, and there is also an underlying assumption that this is the data generating model.

In response to the point about population sizes, we have added rough size numbers to **Figure 5** accordingly. The updated **Figure 5** is included here:

Figure 5 *Schematic representation of the sampling stages and phenotype ascertainment in MGI with respect to a US target population**


US population (>300 million)
Disease rates, demographics
Michigan Medicine (>4 million)
Summary statistics, demographics
Approached for consent
Summary statistics
Michigan Genomics Initiative (>40,000)
Predictor of Interest
Patient-level data
Patient Characteristics e.g. Age, gender
"True" disease status
Limited follow-up window
Derived disease status