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

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

Abstract

Large-scale agnostic association analyses based on existing observational health care databases such as electronic health records have been a topic of increasing interest in the scientific community. However, particular challenges of non-probability sampling and phenotype misclassification associated with the use of these data sources are often ignored in standard analyses. In general, the extent of the bias that may be introduced by ignoring these factors is unknown. In this paper, we develop a statistical framework for characterizing the degree of bias expected in association studies based on electronic health records when disease status misclassification and the sampling mechanism are ignored. Through a sensitivity analysis type approach, this framework can be used to obtain plausible values for parameters of interest given results obtained from standard naive analysis methods under varying degree of misclassification and sampling biases. We develop an online tool for performing this sensitivity analysis in some special cases that occur frequently. Simulations demonstrate promising properties of the proposed way of characterizing biases. We apply our approach to study bias in genetic association studies using data from the Michigan Genomics Initiative, a longitudinal biorepository effort within Michigan Medicine.

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

1 Introduction

Genome-wide genotype information linked with electronic health records (EHR) are becoming increasingly available through biorepository efforts at academic medical centers (e.g. BioVU at Vanderbilt), health care organizations/providers (e.g DiscovEHR from Geisinger Health System), and population-based biobanks (e.g UK Biobank) [1]. Often, the fundamental question of interest is to explore association between a phenotype, say D, with a risk factor of interest, say G, after adjusting for potential confounders, say Z. This 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 a genome-wide association study (GWAS)), or for thousands of phenotypes derived from the structured and unstructured content of the EHR with a given variant G of interest (as in phenome-wide association studies (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, data obtained from the EHR and other existing sources are often not originally intended for research purposes, and additional thought is needed to understand potential sources of bias. In this paper we focus on the particular setting of 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.

One potential source of bias in EHR-based studies is misclassification of the derived disease status. Large-scale agnostic studies using EHR often define patient disease status (phenotypes) based on international classification of diseases codes (ICD codes) or aggregates thereof called "PheWAS codes" or "phecodes" to define phenotypes in an automated and reproducible way [4]. However, there are many reasons for which these ICD-based phenotype classifications can be misclassified. For example, existing or secondary disease conditions may not be entered into the chart, and only the primary reason for the visit and the associated primary diagnoses may be reported for billing purposes. In the study of psychiatric diseases, accurate diagnosis is often challenging and subjective, and we may expect misclassification to occur in both directions [5]. Additionally, symptoms occurring between visits may not always be reported, and the use of diagnostic guidelines and assessment of the phenotype may vary across providers [5]. There may also be a tendency for patients to receive ICD classifications that result in greater reimbursement from the insurance provider.

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]. The goal of such methods is to reduce the amount of misclassification of the disease phenotype, and these methods can be quite successful in this regard. Additionally, human validation can be used on a subset of subjects in order to quantify the degree to which automatic phenotyping algorithms recover disease diagnoses contained in the medical record [11]. These existing phenotyping approaches can be effective in reducing misclassification in reported codes by using text mining and natural language processing techniques on the clinical notes or using supporting information from treatment or lab results, but even the most sophisticated phenotyping algorithms cannot capture diagnoses that were never recorded in any form in the EHR. 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 EHR in capturing all encounters of the patient with a particular health system. In PheWAS and GWAS, we are often interested in identifying the association between genetic information and whether a patient will or will not develop a particular disease in his/her *lifetime.* In practice, EHR-derived phenotype definitions are used to represent the underlying 'true' disease status. These definitions often do not capture diagnoses or medical care obtained from external sources (for example, patients' local health care providers), and they do not adequately capture diseases that the subject will go on to develop in the future or has previously had prior to entrance into the EHR (outside the sampling window). Generally, we are more likely to observe and record a disease in the EHR if we follow a subject more often for a longer period of time (particularly, when they are at the greatest risk of developing the disease). This can lead to a potentially large degree of underreporting of lifetime disease status.

It has been well-established in the literature that outcome misclassification can potentially bias effect estimates in association studies. However, this issue of misclassification of the disease outcome of interest is usually ignored in large-scale agnostic studies. Several authors have proposed statistical methods for addressing misclassification of binary phenotypes for 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. [13] proposes a likelihood-based approach that integrates over unknown sensitivity and specificity values but requires some limited prior information about the sensitivity and specificity. Wang et al. [14] proposes an approach for incorporating both human-validated labels and error-prone phenotypes into the estimation, but this approach will

not account for misclassification due to information not included in the medical record. Duffy et al. [15] and Sinnott et al. [16] expand on results in the measurement error literature to relate parameters in the model for the true unknown outcome with the model for the misclassified outcome, but Duffy et al. [15] focuses on the setting with binary exposures, and Sinnott et al. [16] focuses on the setting in which the probability of having observed disease is explicitly modeled using a variety of information in the EHR. Additionally, all of these methods do not directly address the sampling mechanism.

In addition to potential bias due to misclassification of disease phenotypes, the mechanism by which subjects are sampled into the dataset can sometimes result in biased inference when not handled appropriately. This issue is often ignored in association studies based on EHR data. 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 subject into a biorepository is not *a priori* fixed or defined. Interactions with the health care system are generated by the patient, and it can be difficult to understand the mechanism driving sampling as well self-selection for donating biosamples. Several authors recommend adjusting for factors such as number of health care visits or referral status to better account for the sampling mechanism [17, 18]. Additionally, there is a belief in the literature that gene-related association study results may be less susceptible to bias resulting from the patient sampling mechanism [19]. This belief stems from the assumption that an individual genetic locus is not usually appreciably related to the sampling mechanism and, therefore, resulting associations may be only rarely impacted by the sampling mechanism. However, bias due to genotype relationships with the sampling mechanism can still arise in certain settings [20]. Additionally, a particularly popular topic in genetics-related research right now is the use of polygenic risk scores, which combines information from many genetic loci into a score to quantify a patient's genetic risk for developing a particular disease [21, 22, 23]. While it may be reasonable to assume that a specific genetic locus may have little association with the sampling mechanism, this assumption becomes more tenuous for an aggregate polygenic risk score with stronger association with the underlying disease. There may be cases where sampling may depend on the underlying disease status and thus, indirectly with the polygenic risk score for that disease.

Sampling being related to the disease status may not be a problem when disease status is perfectly observed in the sampled patients. As we will demonstrate, however, this type of disease-dependent sampling mechanism can create substantial problems in the presence of misclassification of the disease status. Existing statistical methods for dealing with phenotype misclassification such as those presented in Huang et al. [13], Duffy et al. [15] and Sinnott et al. [16], do not directly take into account the mechanism by which subjects are sampled into the dataset 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 selection and observation processes on results from a naive association analyses.

While many sources of bias in the study design and measurement processes are likely to exist, the degree of the bias for model parameters of direct interest often remains unclear. Additional characterization of the mechanisms by which we can expect bias and their impact on resultant inference may help guide improved study design and analysis in the future. In this paper, we develop a statistical framework incorporating models for both disease misclassification and disease-dependent sampling. We use this framework to characterize the degree of bias expected in EHR-based association study results when misclassification and the sampling mechanism are ignored in the naive and commonly-used analysis approach. We focus on the particular setting in which phenotypes are underreported (naturally occurring due to limited observation

window bias), but the approaches used can be extended to the more general setting of bi-directional misclassification. By providing analytic expressions for the bias in a naive analysis, the proposed framework can provide a more statistically-rigorous motivation for determining when we do and do not expect bias in our study results. The analytical expressions also enhance our understanding of which study design and phenotype characteristics this bias may depend on. Through a sensitivity-analysis type approach, this framework can also be used to obtain plausible values for parameters of interest given results from simpler analysis methods. Simulations demonstrate promising properties of the proposed approach in capturing the true bias. We provide an interactive online tool for performing these sensitivity analyses. We apply our approach to study bias in genetic association studies using EHR-derived phenotypes in the Michigan Genomics Initiative, a longitudinal biorepository effort within Michigan Medicine.

2 Model Structure

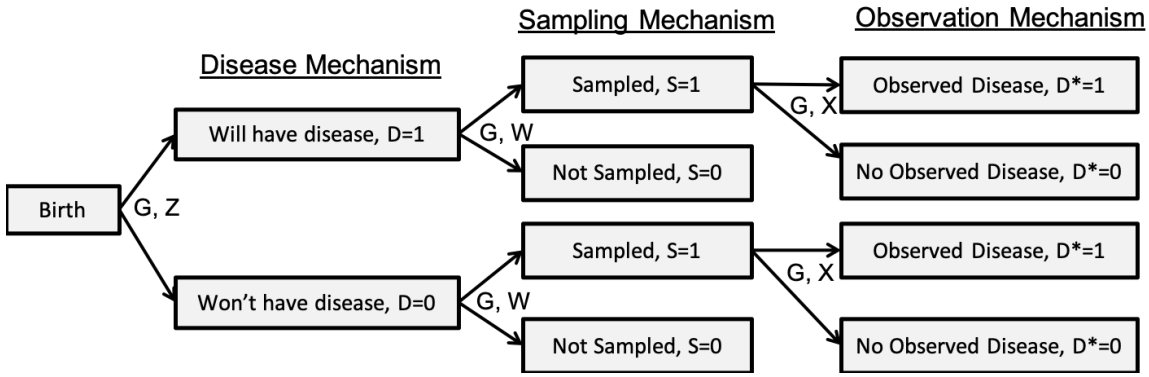
Let binary variable D represent a subject's lifetime disease status, which is unknown at birth. Suppose we are interested in the relationship between D and an individual's inherited genetic information, G, adjusting for additional person-level information, Z. We will call this relationship the *disease 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. 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 single specified D and G.

In order to make inference about D, suppose we consider a large health care system-based database with the goal of making inference about the general population. Let S indicate whether a particular subject in the population is sampled into our dataset (for example, by going to a particular hospital and consenting to share biosample for research), where the probability of an individual in the population being included in the current dataset may depend on the underlying lifetime 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 state may also impact whether subjects are sampled into the study base or not. We will call this mechanism the *sampling mechanism.*

Subjects sampled into our dataset are then followed for the disease of interest, D. Instances of the disease are recorded in hospital or administrative records. We might expect factors 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 in a given subject. 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 data structure. Our primary interest is in making inference about D given G and Z, which has corresponding parameter θ. Due to the sampling and phenotyping practices often used in analyses, however, inference is often being made instead using a model for D^* given G and Z among the sampled subjects ($S = 1$). This standard modeling strategy ignores potential misclassification and disease/covariate dependent sampling. In this paper, we are interested in characterizing which factors in the sampling and disease observation process may generate bias when the analysis is based on $D^*|G, Z, S = 1$ and interpreted as inference about $D|G, Z$. In order to characterize the bias, we define statistical models for the various sampling and observation mechanisms as follows:

Figure 1: *Diagram of the assumed data structure*



Conceptual Model *(Eq. 1)*

$$\begin{aligned}
\text{Disease Mechanism}: &\quad \text{logit}(P(D=1|Z,G;\theta)) = \theta_0 + \theta_G G + \theta_Z Z \\
\text{Sampling Mechanism}: &\quad \text{logit}(P(S=1|D,W,G;\phi)) = \phi_0 + \phi_W W + \phi_D D + \phi_G G \\
\text{Observation Mechanism}: &\quad \text{logit}(P(D^*=1|D=1,S=1,X,G;\beta)) = \beta_0 + \beta_X X + \beta_G G \\
&\quad \text{logit}(P(D^*=1|D=0,S=1,X,G;\alpha)) = \alpha_0 + \alpha_X X + \alpha_G G
\end{aligned}$$

The above modeling framework allows for misclassification of the true disease status in either direction (subjects with the disease may be missed and subjects may be listed as diseased who aren't), and we allow both the sampling and observation mechanisms to depend on G. Moving forward, however, we will restrict our focus to a particular setting in which we assume (1) *misclassification occurs only through underreporting* (so essentially $\alpha_0 = -\infty$ and $\alpha_X = \alpha_G = 0$) and (2) *G is independent of both the sampling and misclassification mechanisms given D*. This first assumption implies that subjects without the disease are never diagnosed with the disease. Viewing D^* as a test for D, this is equivalent to assuming perfect specificity in the sampled subjects. When D^* represents disease status derived from the EHR and D represents lifetime disease status, through assumption (1) we are trying to capture limited follow-up and, possibly, underdiagnosis. The second assumption restricts $\phi_G = \beta_G = 0$ and implies that X and W are independent of G given D. The second restriction may often be reasonable when G represents a single genetic locus; we do not expect genotype status at a single marker to be related to the sampling or observation mechanisms, especially given D. A possible exception to this is in the case where G represents a rare, highly penetrant mutation. Suppose, however, G is another variable such as a polygenic risk score (PRS). The PRS, which represents aggregate effects of many genetic loci, may often be marginally related to sampling if our sampled dataset is enriched for the disease of interest. However, conditioning on the true disease status, D, it may be more reasonable to assume that the PRS is independent of the sampling and observation mechanisms. This assumption may not be tenable when sampling is based on whether the subject has a disease $\tilde{D}$ *related to* the disease of interest, D. Conditional on D but not $\tilde{D}$, we may still have some association between the PRS and the sampling mechanism.

Under assumptions (1) and (2), this model is an adaptation of the zero-inflated logistic regression model developed in Diop et al. (2011), where the Diop et al. (2011) model does not directly incorporate the sampling mechanism [24]. The model in Diop et al. (2011) can also be viewed as a reformulation of a mixture cure model, where the failure time distribution is replaced with a logistic regression model for whether or not the event of interest will ever occur [25]. Our model is a further reformulation to incorporate sampling dependent on the underlying

disease status (D) and/or other subject-level variables.

3 Approximating Bias Resulting from Standard Analysis

3.1 Approximating the Bias

We suppose that G, Z, X, and W are observed if and only if $S = 1$. We further assume that (G, Z), X, and W are all independent, so different factors impact the outcome generation, the response mechanism, and the reporting/observation mechanism in the $D = 1$ subjects. However, G may be related to Z. Future explorations can consider the setting where the covariate sets are not assumed to be independent.

In practice, 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 *target model* as follows:

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

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

$$P(D^* = 1|G, Z, S = 1; \theta^{(simple)}) = \frac{P(D = 1|G, Z; \theta)c(\beta)r(\phi)}{1 + (r(\phi) - 1)P(D = 1|G, Z; \theta)} \qquad \text{where}$$

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

We can view D^* as a noisy test for the true value of D with perfect specificity and potentially imperfect sensitivity, and constant c represents the sensitivity of D^* for D in the sampled subjects averaged across the distribution of X. Constant r represents the sampling ratio comparing subjects with $D = 1$ and $D = 0$. We note that since c and r depend on distinct parameters, they can vary independently. Additionally, c and r depend on distributions $f(X)$ and $f(W)$. In practice, however, we will *not* need to specify these distributions.

We assume that covariates Z are centered so that they have mean zero and that the data are modeled as in (*Eq. 1*). As shown in **Supplementary Section 1**, we use Taylor series approximation to obtain expressions for the intercept and coefficient for G in the *analysis model* as functions of parameters in the *target model*. We consider two different cases with two types of genetic variables G.

<u>Case 1: G is a SNP</u>

Suppose that G represents a single genetic locus or SNP (single nucleotide polymorphism), coded 0/1/2 to represent the number of copies of the minor allele for each subject. Let MAF denote the minor allele frequency in the *population*. In this case, we can express the *analysis model* parameters as a function of the *target model* parameters as follows:

$$\theta_0^{(simple)} \approx \theta_0 + \log(r) + \log(c) - \log\left(1 + r(1-c)e^{\theta_0 + 2\theta_G MAF}\right) + 2\theta_G MAF \frac{r(1-c)e^{\theta_0 + 2\theta_G MAF}}{1 + r(1-c)e^{\theta_0 + 2\theta_G MAF}}$$

$$\theta_G^{(simple)} \approx \theta_G - \theta_G \frac{r(1-c)e^{\theta_0 + 2\theta_G MAF}}{1 + r(1-c)e^{\theta_0 + 2\theta_G MAF}} \qquad (Eq.\ 2)$$

Here, we are approximating the sample MAF with the population MAF. For a discussion of this issue, see **Supplementary Section 1.1**. Additionally, we implicitly assume there are only two non-neglible alleles for the genetic locus of interest.

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 [26]) and suppose G has been centered to have mean zero. In this setting, we can express the *analysis model* parameters as follows:

$$\theta_0^{(simple)} \approx \theta_0 + \log(r) + \log(c) - \log\left(1 + r(1-c)e^{\theta_0}\right)$$

$$\theta_G^{(simple)} \approx \theta_G - \theta_G \frac{r(1-c)e^{\theta_0}}{1 + r(1-c)e^{\theta_0}} \qquad (Eq.\ 3)$$

Simulations exploring the correspondence between the approximations in (*Eq. 2*) and (*Eq. 3*) with estimated parameters can be found in **Figure S2**.

3.2 Understanding the Structure of the Bias

We can use expressions (*Eq. 2*) and (*Eq. 3*) to develop some intuition for settings in which we expect greater or less bias when performing the "simple" routine analysis. **Table 1** describes the general impact of the various model parameters on the bias of $\theta^{(simple)}$, the parameter in the simple analysis model.

Table 1: *Bias in target model parameters in the simple analysis*

Variable	Value	G **is a PRS*** Bias of $\theta_0^{(simple)}$	G **is a PRS or SNP** Bias of $\theta_G^{(simple)}$
Sensitivity (c)	Decreases toward 0	Bias increases	Bias increases
	• No Underreporting ($c = 1$)	Biased if $r \neq 1$ or $\tilde{r} \neq 1^\dagger$	No Bias
	• Underreporting ($c < 1$)	Biased	Biased
Sampling Ratio (r)	Increases toward ∞	Bias depends on θ_0, c	Bias increases if $c < 1$
	• Sampling Independent of D ($r = 1$)	Biased if $c < 1$	Biased if $c < 1$
	• Sampling Dependent on D ($r \neq 1$)†	Biased	Biased if $c < 1$
Disease Prevalence	Increases toward 1	Bias depends on r, c	Bias increases if $c < 1$
Absolute Effect of G	Increases toward ∞	Little or no added bias	Bias increases if $c < 1$

*Direction of Bias of $\theta_0^{(simple)}$ under Case 1 (G is a SNP) depends on other parameters

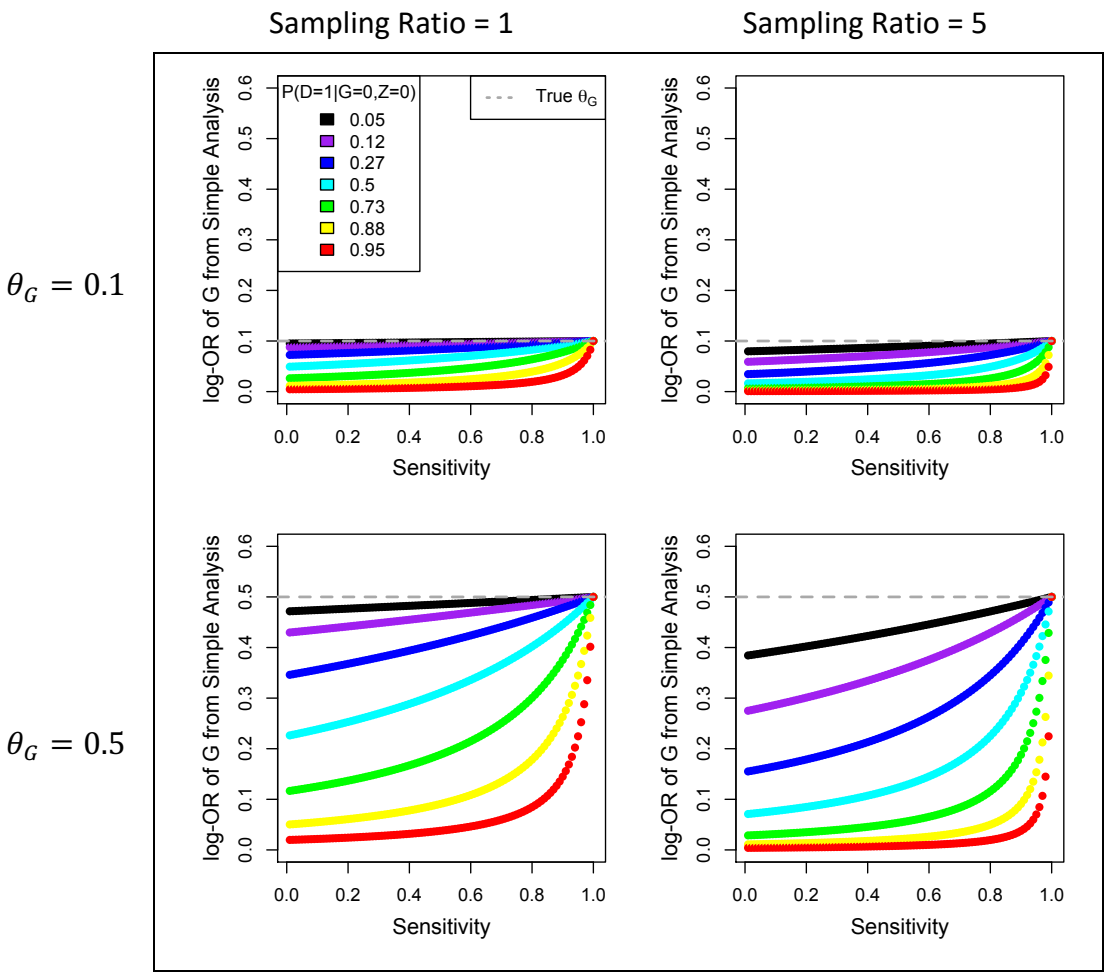
† We consider the case of an additional sampling step dependent on observed case-control status, D^*, in **Supplementary Section 2** and **Figure S1**.

The first thing we note is, as expected, there is no bias in estimating $\theta_G^{(simple)}$ when we have perfect sensitivity (so $c = 1$). However, suppose $c < 1$. In this case, we expect bias in the estimated effect of G, and the degree of the bias will depend on the value of the sensitivity, the disease sampling ratio r, the population prevalence of the disease, 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 subjects 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. As the sensitivity decreases, the "simple" estimate of θ_G becomes increasingly biased toward the null. However, we note that the degree 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 observed relatively low bias in estimating θ_G even with low sensitivity for observing D in the sampled subjects. This is a result of the assumption that we have perfect specificity, and bidirectional misclassification of the disease phenotype will change the impact of the population prevalence on the bias. 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 disease-dependent sampling can have a substantial impact on the absolute bias for $\theta_G^{(simple)}$. This provides some support to the notion that 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 lifetime population prevalence or larger values for θ_G (for example, a highly penetrant SNP), we may be more concerned about the potential for bias induced by ignoring disease misclassification and/or disease-dependent sampling. 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/`.

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



3.3 Sensitivity Analysis

An alternative use of the proposed model is in a sensitivity analysis in a reverse direction, where we can obtain reasonable values of θ_G across plausible values of r and c based on the parameter estimates from the "simple analysis." In **Supplementary Section 6**, we describe one approach

for getting a sense of plausible values for the sampling ratio for a given data 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 (*Eq. 2*) 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, $P(D = 1)$. Alternatively, suppose that $\theta_0^{(simple)}$ is also available. In this case, we can obtain θ_G by solving

$$\theta_G^{(simple)} \approx \theta_G \frac{c - (1-c)e^{\theta_0^{(simple)}}}{c - (1-c)e^{\theta_0^{(simple)}}\left[1 - e^{2\theta_G MAF}\right]} \qquad (Eq.\ 4)$$

for θ_G as shown in **Supplementary Section 3.1**. In this case, predicted θ_G will be a function of c and $\theta^{(simple)}$ but not r. Similar expressions for predicting θ_G when G is a PRS can be found in **Supplementary Section 3.2**.

We note that in some fraction of cases these expressions may have no solution for given values of r and c. For example, through our analytic work we observed that we might have no solutions for smaller values of c if we have a particularly large population prevalence and a large value for $\theta_G^{(simple)}$. This follows our intuition about compatibility of parameter configurations as we have large bias toward the null when c is small and the prevalence of the disease is moderate to large, so we would not often see very large values of $\theta_G^{(simple)}$ in that setting. For reasonable combinations of variables, however, these expressions can be used in a sensitivity analysis after the primary analysis has been completed to explore the potential impact of ignoring misclassification and disease-dependent selection.

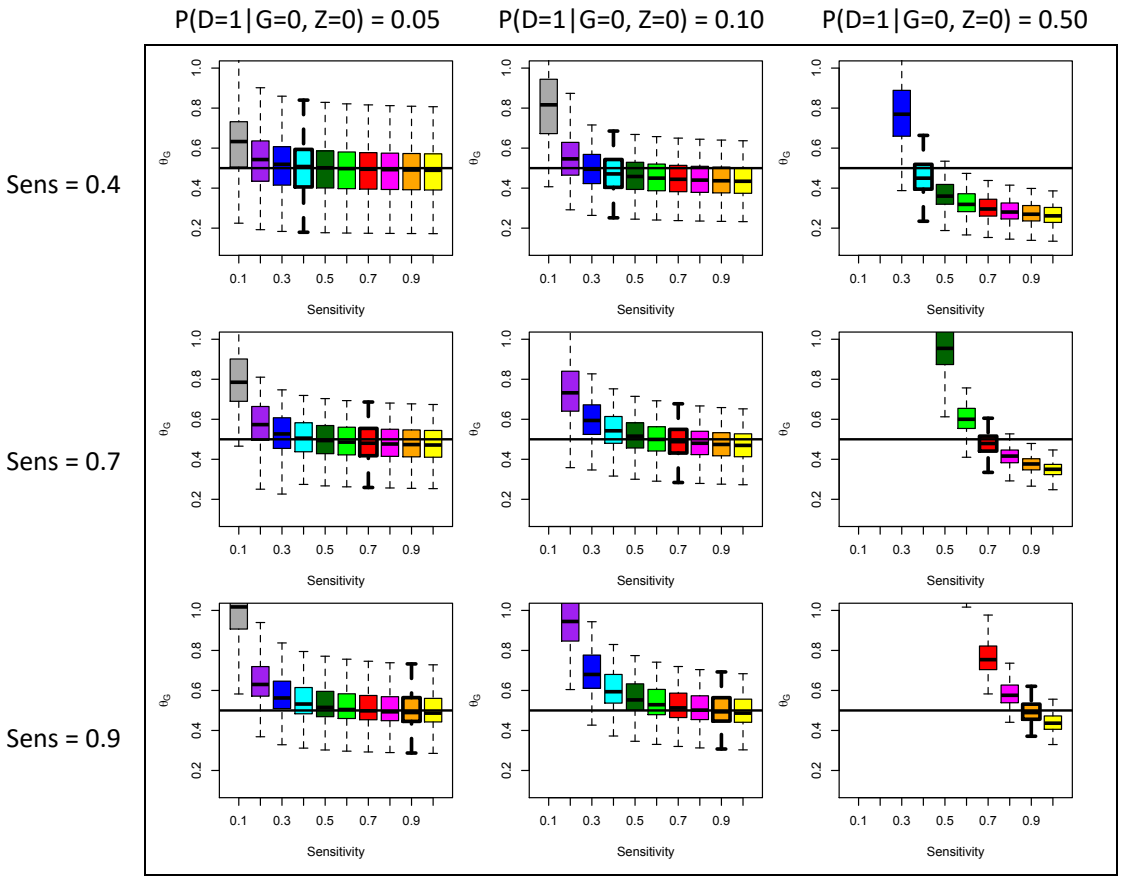
4 Simulations Evaluating Proposed Sensitivity Analysis

The usefulness of the expressions in (*Eq. 2*) and (*Eq. 3*) relies on the accuracy of the Taylor series approximations used to derive them. In **Supplementary Section 4.1**, we show strong concordance between our approximation of predicted estimates for $\theta^{(simple)}$ from (*Eq. 2*) and (*Eq. 3*) with estimates from simulated data.

One use of the approximations in (*Eq. 2*) and (*Eq. 3*) is to guide a sensitivity analysis after $\theta^{(simple)}$ has already been estimated, and we describe how this can be done in **Section 3.3**. Here, we present results from a small set of simulations in which we compare the correspondence between true values of θ_G and values predicted by the sensitivity analysis in different settings. We will focus on the setting where G is a single genotype as this setting is more complicated, and we will simulate assuming a sampling ratio of 1. Other sampling ratios produce similar results.

A detailed description of the simulation procedure can be found in **Supplementary Section 4.1**. Briefly, we simulate data under different combinations of c and $P(D = 1|G = 0, Z = 0)$ (100 datasets each). For each simulated dataset, we fit the simple analysis model and obtain a corresponding prediction for θ_G by solving either (*Eq. 2*) or (*Eq. 4*) for θ_G. **Figure 3** and **Figure 4** show the predictions resulting when $\theta_0^{(simple)}$ is known and unknown respectively. Each subplot corresponds to a different combination of true c (denoted by the bolded boxplot) and $P(D = 1|G = 0, Z = 0)$. $\theta_0^{(simple)}$ may often be unknown if we obtain $\theta_G^{(simple)}$ using published summary statistics or if the data analysis pipeline does not routinely store $\theta_0^{(simple)}$.

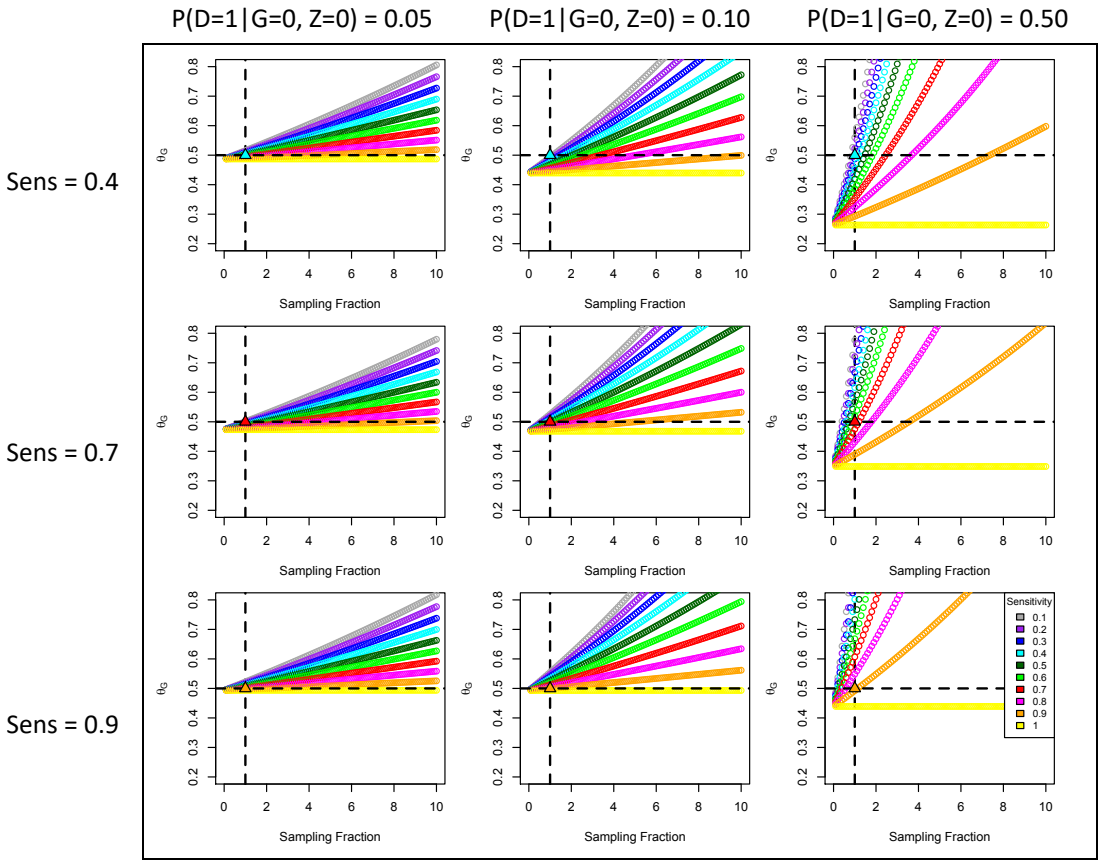
Figure 3: *Predicted θ_G across different sensitivities when simple analysis intercept is known.* The black horizontal line indicates the true value of $\theta_G = 0.5$. The bolded boxplot indicates the true sensitivity. Each boxplot contains estimates from 100 simulations.



In **Figure 3**, we can see that the boxplot for predicted θ_G corresponding to the true sensitivity (bolded) always includes the true value of θ_G, and the corresponding median for the predicted θ_G is generally close to the true value of 0.5. These boxplots illustrate how a smaller population prevalence for the disease of interest is associated with greater robustness of inference (or greater similarity between θ_G and $\theta_G^{(simple)}$) across sensitivity values under our model.

Figure 4 shows predicted θ_G when $\theta_0^{(simple)}$ is unknown but θ_0 is known. In practice, θ_0 may not be known and we can perform sensitivity analyses over a range of θ_0. In each subplot in **Figure 4**, we provide predictions using the average value of $\theta_G^{(simple)}$ across the 100 simulated datasets. We might expect some additional variability if we use $\theta_G^{(simple)}$ from a single model fit. The plotted triangle in this figure indicates the true sampling fraction and θ_G, and the color of the triangle indicates the true sensitivity value. We can see generally good concordance with predicted values of θ_G and the true value of θ_G for the true value of c. We emphasize, however, that we cannot use these explorations to determine or estimate the "correct" values of r and c, as they cannot be identified independently. However, we can explore around reasonable values of r and c based on our understanding of the data to determine the range of plausible values for the parameter of interest, θ_G. In **Figure S3**, we provide additional explorations assuming the sampling ratio is known to be 1.

Figure 4: *Predicted θ_G across different sensitivities and sampling ratios when simple analysis intercept is unknown.* Predictions are obtained assuming true $P(D = 1|G = 0, Z = 0)$ is known and using the *average* estimated $\theta_G^{(simple)}$ across 100 simulated datasets. The triangle's location and color denotes the true values of θ_G, the sensitivity, and the sampling ratio.



5 Bias Exploration in the Michigan Genomics Initiative

The Michigan Genomics Initiative (MGI) is a longitudinal biorepository effort within Michigan Medicine linked to EHR containing over 30,000 patients with matched genotype and phenotype information. We have strong reason to believe that sampling in MGI is somehow related to patient characteristics and, in particular, underlying disease status. Firstly, all subjects in MGI are people who visited Michigan Medicine. Subjects may go to the doctor when they feel sick or for other personal reasons, and these reasons are often unknown. Additionally, subjects are recruited into MGI from the pool of subjects having surgery at Michigan Medicine. Subjects 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]. Due to its sampling mechanism, MGI is strongly enriched for cancer cases, as shown in Beesley et al. [1] and reproduced in **Table S1**. The large number of cancer cases makes MGI an attractive dataset for exploring associations between cancer diagnosis and genetics.

We are primarily interested in associations between *lifetime* cancer diagnosis and genetic factors. When we have perfectly classified phenotypes (so there is no observation window bias and no under- or over-reporting of disease status), simple association analyses may produce reasonable results. However, when disease status is potentially misclassified, disease-dependent sampling can result in biased estimates from the target model. We have several reasons to

believe that underreporting of disease status may be of concern for these data. Diagnosis information is only available for MGI subjects during the time in which they are being treated at Michigan Medicine. Many patients seek treatment at Michigan Medicine for specialized care and then return to primary care providers, and these primary care interactions are not available in the Michigan Medicine EHR. EHR of past disease diagnoses may be incomplete, and we cannot know future disease diagnoses. When we are interested in lifetime disease rates, this results in underreporting of the true lifetime disease status (i.e. some subjects who have had or will have the disease are listed as non-diseased).

Previous analyses in Beesley et al. [1] and Fritsche et al. [21] have explored associations between single genetic markers and polygenic risk scores with 1,711 phenotypes for 30,702 patients in MGI using simple analysis methods involving fitting a model for $D^*|G, Z, S = 1$. These analyses ignore the potential bias induced by disease-related sampling and phenotype misclassification. Based on the results of our bias explorations, which indicate that disease-dependent sampling can potentially strongly impact bias in estimating phenotype-genotype associations in the presence of phenotype misclassification, we are interested in evaluating the potential impact of this disease-dependent sampling on inference. Below, we will perform the sensitivity analyses proposed previously to study SNP-phenotype and PRS-phenotype associations in MGI.

5.1 Genetic association analysis using individual markers

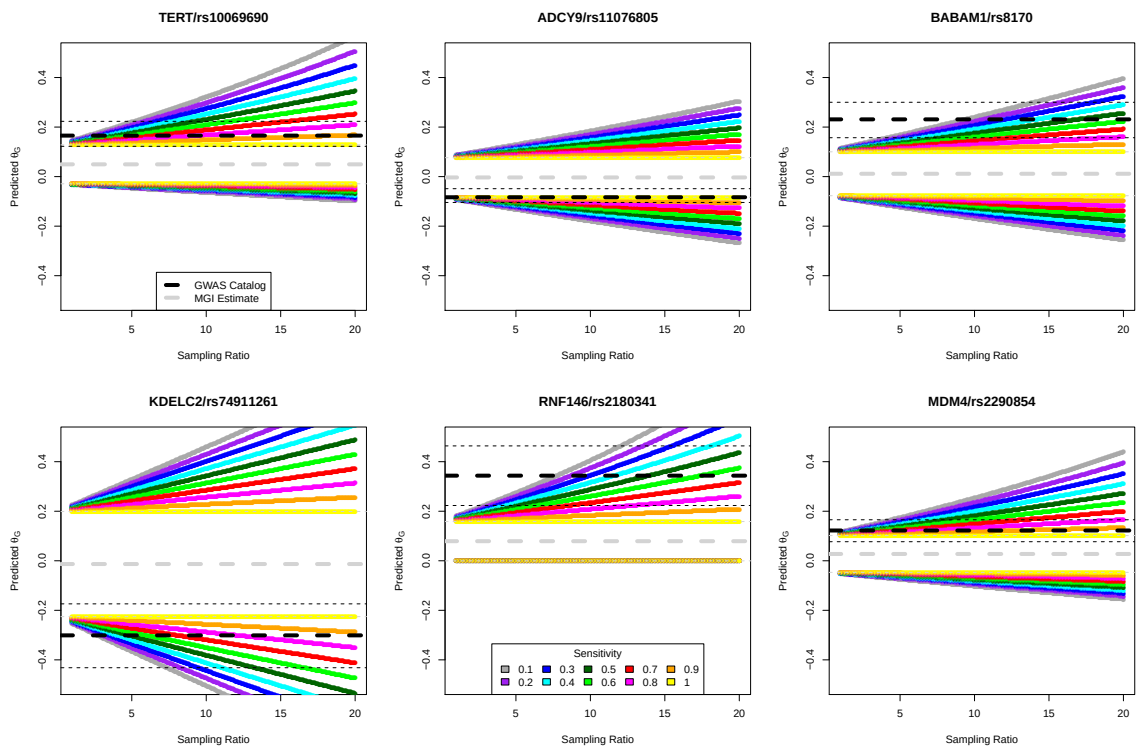
In Beesley et al. [1], we compared genome-wide association study results for an EHR-derived breast cancer phenotype in MGI with reported associations from the NHGRI GWAS catalog, which combines results from meta-analysis of the largest and high quality studies. Here, we will treat NHGRI GWAS catalog as a comparative gold standard). We demonstrated that the association results using MGI data are generally similar to results in the GWAS catalog, but there are some specific SNPs for which the results differ, as shown in the **Figure S4**. For a detailed description of the phenotype generation, locus pruning, and GWAS analysis, see Beesley et al. [1]. Briefly, phenotypes were defined using phecodes based on ICD codes using the R package PheWAS [4].

We focus our attention on 6 individual loci for which the GWAS catalog estimate differs appreciably from the MGI estimate. For this exploration, we fix $\theta_0 = \text{logit}(0.124)$, using the US female population lifetime rate of breast cancer [1]. $\theta_0^{(simple)}$ is unknown based on the currently available GWAS results, and we explore plausible values of θ_G across different potential sampling ratios and sensitivities. The simple analysis in MGI was performed on a matched subset of MGI based on age and the first four principal components of the genotype data, and our results also adjusted for age and the principal components. We obtain rough intervals for θ_G for fixed values of r and c using the upper and lower 95% confidence interval limits of the MGI GWAS estimate for each locus. In **Supplementary Section 6**, we describe one approach for getting a sense of plausible values for the sampling ratio.

Figure 5 shows the results. 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 bias when r is small. 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 even for very low sensitivity when the sampling ratio is near 1. We would expect to see a bigger difference across sensitivity for low values of r when the estimated $\theta_G^{(simple)}$ is larger, as is seen in the next

section. As the sampling ratio increases, however, the degree of bias increases, and the predicted θ_G becomes more extreme. 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 S5**. These results suggest that even GWAS results can potentially be impacted by disease-dependent sampling and misclassification. However, we note that comparison across multiple association tests within a GWAS study should have similar sampling ratios and misclassification rates. Therefore, we may have greater confidence in comparisons across association tests rather than in the absolute estimated effect for a particular association. We will explore this issue further in future work.

Figure 5: *Plausible values for θ_G in MGI for six SNPs known to be associated with breast cancer.* The NHGRI GWAS catalog (Michigan Genomics Initiative) estimate and corresponding 95% confidence interval are shown in black (gray)



5.2 Association analysis with Polygenic Risk Scores

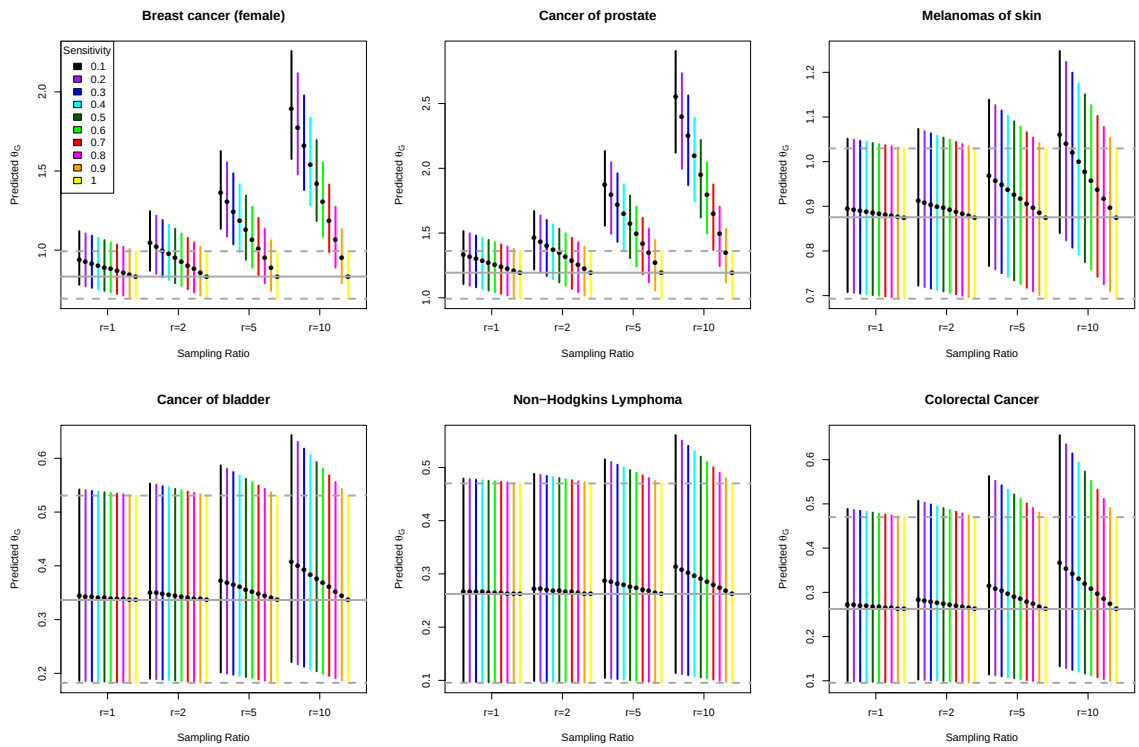
Previous work in Fritsche et al. [21] explored associations between polygenic risk scores (PRS) and their corresponding EHR-derived phenotypes in MGI for several cancer phenotypes. We are interested in exploring associations we might have seen accounting for different levels of disease-dependent sampling and misclassification for six different cancer diagnoses. We treat the published PRS association summary statistics (values listed in **Table S1**) as our $\theta_G^{(simple)}$ values, and we use the corresponding Michigan lifetime prevalence rates reported in Beesley et al. [1] to approximate θ_0 for each of the six cancers of interest. 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. 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. 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 bias in this setting. Secondly, the simple estimated association between the PRS and the disease is small in all of these cases. While the relative bias will not appreciably change as θ_G changes, the absolute bias will be small for small θ_G in this setting. In contrast, breast and prostate cancer have high prevalences, and the corresponding PRS associations are strong. For these two PRS associations, our results suggest we may have appreciable 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. We explore plausible values for the sampling ratio for these cancers in MGI in **Figure S6**.

In all cases in **Figure 6**, the confidence interval for the PRS-phenotype association is far from zero, and adjustment for misclassification and disease-dependent 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.

Figure 6: *PRS-phenotype associations for six different cancers*



6 Discussion

The proposed modeling and sensitivity analysis framework allows us to explore the degree of bias we might expect in large association study results when we ignore issues of disease-dependent sampling (or sampling dependent on other patient characteristics) and disease underreporting,

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 rate is high, however, disease-dependent sampling can have a substantial impact in biasing association results. Without any misclassification of the outcome, we do not expect much bias in these association study results even when the sampling depends on the underlying disease status under our model. This is a property of the logistic regression modeling assumptions, and modeling using different link functions may produce slightly different results. We focus our attention on the dependence between underlying disease status and sampling, but sampling may be also related to other patient characteristics. As long as the sampling mechanism is independent of the disease and observation mechanisms given the true disease status, we do not expect much additional bias beyond the bias induced by the disease-dependent sampling and the disease misclassification.

One assumption made throughout this paper is that the primary driver of misclassification is underreporting of disease status, either through observation window bias or through underdiagnosis. This assumption drives many of the bias properties seen in this paper. In particular, we find that higher disease prevalence in the population is associated with a higher potential for bias in the simple analysis. This property is a result of our unidirectional misclassification assumption. Additionally, the unidirectional misclassification mechanism paired with disease-dependent sampling is associated with bias toward the null for genotype-phenotype association parameters. Bias expressions incorporating bidirectional misclassification will allow for bias in either direction. The current exploration focuses on the impact of disease-dependent sampling and underreporting on effect estimates. However, we are often interested in resulting 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.

In this paper, we focus on estimation of a single genotype-phenotype association, but we are often interested in performing such estimation for many genotype-phenotype associations. A natural question is the extent at which this bias impacts comparison across estimated parameters 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 and sensitivity 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 sensitvity 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. Additionally, there may be an opportunity to incorporate additional information into the assessment of comparative bias including the genetic architecture and the disease heritability.

The approximations derived in this paper are similar to the relationships between models for true and misclassified binary outcomes applied to EHR data in Duffy et al. [15] and Sinnott et al. [16], which both make use of existing results in the measurement error literature for misclassified binary outcomes. Our derived expressions can be viewed as an extension of this existing work to also incorporate the sampling mechanism and to allow for the predictor of interest (here, G) to take different forms.

The proposed framework can be useful for guiding analyses exploring sensitivity to vio-

lations of the common implicit assumptions of no outcome misclassification and no disease-dependent sampling in EHR-based association studies. In future work, we will develop statistical methodology to perform parameter *estimation* 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 and staff at Michigan Genomics Initiative along with MGI participants for generously donating their biosamples for research. This work was supported by National Science Foundation 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.

8 Contributions

Dr. Lauren J. Beesley wrote the manuscript, developed the statistical methodology, developed the online SAMBA-EHR application.
Dr. Lars G. Fritsche helped edit the paper, facilitated data acquisition and interpretation.
Dr. Bhramar Mukherjee wrote the manuscript and provided key guidance in the development of the methodology and manuscript throughout the entire process.

References

[1] Lauren J Beesley, Maxwell Salvatore, Lars G Fritsche, Anita Pandit, Arvind Rao, Cristen J Willer, Lynda D Lisabeth, and Bhramar Mukherjee. The Emerging Landscape of Epidemiological Research Based on Biobanks Linked to Electronic Health Records. *Preprints.org*, pages 1–35, 2018. doi: 10.20944/preprints201809.0388.v1.

[2] Joshua C. Denny, Marylyn D. Ritchie, Melissa A. Basford, Jill M. Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R. Masys, Dan M. Roden, and Dana C. Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 26(9):1205–1210, 2010. doi: 10.1093/bioinformatics/btq126.

[3] Brooke N. Wolford, Cristen J. Willer, and Ida Surakka. Electronic health records: The next wave of complex disease genetics. *Human Molecular Genetics*, 27(R1):R14–R21, 2018. doi: 10.1093/hmg/ddy081.

[4] Robert J. Carroll, Lisa Bastarache, and Joshua C. Denny. R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, 30(16):2375–2376, 2014. doi: 10.1093/bioinformatics/btu197.

[5] Victor M. Castro, Jessica Minnier, Shawn N. Murphy, Isaac Kohane, Susanne E. Churchill, Vivian Gainer, Tianxi Cai, Alison G. Hoffnagle, Yael Dai, Stefanie Block, Sydney R. Weill, Mireya Nadal-Vicens, Alisha R. Pollastri, J. Niels Rosenquist, Sergey Goryachev, Dost

Ongur, Pamela Sklar, Roy H. Perlis, Jordan W. Smoller, Phil Hyoun Lee, Eli A. Stahl, Shaun M. Purcell, Douglas M. Ruderfer, Alexander W. Charney, Panos Roussos, Carlos Pato, Michele Pato, Helen Medeiros, Janet Sobel, Nick Craddock, Ian Jones, Liz Forty, Arianna DiFlorio, Elaine Green, Lisa Jones, Katherine Dunjewski, Mikael Landén, Christina Hultman, Anders Juréus, Sarah Bergen, Oscar Svantesson, Steven McCarroll, Jennifer Moran, Kimberly Chambert, and Richard A. Belliveau. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry*, 172(4):363–372, 2015. doi: 10.1176/appi.ajp.2014.14030423.

[6] William S. Bush, Matthew T. Oetjens, and Dana C. Crawford. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, 17(3):129–145, 2016. doi: 10.1038/nrg.2015.36.

[7] K. P. Liao, T. Cai, G. K. Savova, S. N. Murphy, E. W. Karlson, A. N. Ananthakrishnan, V. S. Gainer, S. Y. Shaw, Z. Xia, P. Szolovits, S. Churchill, and I. Kohane. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*, 350:1–5, 2015. doi: 10.1136/bmj.h1885.

[8] Victor Castro, Yuanyuan Shen, Sheng Yu, Sean Finan, Cindy Ta Pau, Vivian Gainer, Candace C. Keefe, Guergana Savova, Shawn N. Murphy, Tianxi Cai, and Corrine K. Welt. Identification of subjects with polycystic ovary syndrome using electronic health records. *Reproductive Biology and Endocrinology*, 13(116):1–8, 2015. doi: 10.1186/s12958-015-0115-z.

[9] Canqing Yu, Yu Guo, Zheng Bian, Ling Yang, Iona Y. Millwood, Robin G. Walters, Yiping Chen, Yan Chen, Xi Zhang, Yulong Lei, Junshi Chen, Zhengming Chen, Jun Lv, and Liming Li. 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. doi: 10.1002/ijc.31566.

[10] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131., 2018. doi: 10.1016/j.cell.2018.02.010.

[11] Katherine P. Liao, Jeffrey A. Sparks, Boris P. Hejblum, I. Hsin Kuo, Jing Cui, Lauren J. Lahey, Andrew Cagan, Vivian S. Gainer, Weidong Liu, T. Tony Cai, Jeremy Sokolove, and Tianxi Cai. Phenome-Wide Association Study of Autoantibodies to Citrullinated and Noncitrullinated Epitopes in Rheumatoid Arthritis. *Arthritis and Rheumatology*, 69(4): 742–749, 2017. doi: 10.1002/art.39974.

[12] Rebecca A. Hubbard, Rhondee Benjamin-Johnson, Tracy Onega, Rebecca Smith-Bindman, Weiwei Zhu, and Joshua J. Fenton. Classification accuracy of claims-based methods for identifying providers failing to meet performance targets. *Statistics in Medicine*, 34(1): 93–105, 2015. doi: 10.1002/sim.6318.

[13] Jing Huang, Rui Duan, Rebecca A. Hubbard, Yonghui Wu, Jason H. Moore, Hua Xu, and Yong Chen. PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*, 2018. ISSN 1527974X. doi: 10.1093/jamia/ocx137.

[14] Lu Wang, Scott M Damrauer, Hong Zhang, Alex X Zhang, Rui Xiao, Jason H Moore, and Jinbo Chen. Phenotype validation in electronic health records based genetic association studies. *Genet Epidemiol.*, 41(8):790–800, 2017.

[15] S. W. Duffy, J. Warwick, A. R.W. Williams, H. Keshavarz, F. Kaffashian, T. E. Rohan, F. Nili, and A. Sadeghi-Hassanabadi. A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology and Community Health*, 58(8): 712–717, 2004. ISSN 0143005X. doi: 10.1136/jech.2003.010546.

[16] Jennifer A. Sinnott, Wei Dai, Katherine P. Liao, Stanley Y. Shaw, Ashwin N. Ananthakrishnan, Vivian S. Gainer, Elizabeth W. Karlson, Susanne Churchill, Peter Szolovits, Shawn Murphy, Isaac Kohane, Robert Plenge, and Tianxi Cai. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics*, 133(11):1369–1382, 2014. ISSN 14321203. doi: 10.1007/s00439-014-1466-9.

[17] Benjamin A. Goldstein, Nrupen A. Bhavsar, Matthew Phelan, and Michael J. Pencina. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology*, 184(11):847–855, 2016. ISSN 14766256. doi: 10.1093/aje/kww112.

[18] Matthew Phelan, Nrupen Bhavsar, and Benjamin A Goldstein. 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)*, 5(1):22, 2017. ISSN 2327-9214. doi: 10.5334/egems.243.

[19] George D. Smith and Shah Ebrahim. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003. ISSN 03005771. doi: 10.1093/ije/dyg070.

[20] Christy L Avery, Keri L Monda, and Kari E North. Genetic association studies and the effect of misclassification and selection bias in putative confounders. *BMC Proceedings*, 3 (Suppl 7):S48, 2009.

[21] Lars G. Fritsche, Stephen B. Gruber, Zhenke Wu, Ellen M. Schmidt, Matthew Zawistowski, Stephanie E. Moser, Victoria M. Blanc, Chad M. Brummett, Sachin Kheterpal, Gonçalo R. Abecasis, and Bhramar Mukherjee. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *The American Journal of Human Genetics*, 102(6):1–14, may 2018.

[22] Lars G. Fritsche, Lauren J. Beesley, Peter VandeHaar, Robert B. Peng, Maxwell Salvatore, Matthew Zawistowski, Sarah Gagliano, Sayantan Das, Jonathon LeFaive, Erin O. Kaleba, Thomas T. Klumpner, Stephanie E. Moser, Victoria Blanc, Chad Brummett, Sachin Kheterpal, Goncalo Abecasis, Stephen B. Gruber, and Bhramar Mukherjee. Exploring Various Polygenic Risk Scores for Basal Cell Carcinoma, Cutaneous Squamous Cell Carcinoma and Melanoma in the Phenomes of the Michigan Genomics Initiative and the UK Biobank. *bioRxiv*, 2018. doi: 10.1101/384909.

[23] A Torkamani, N E Wineinger, and E J Topol. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*, 19:581–590, 2018.

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

[25] Judy P Sy and Jeremy M G Taylor. Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[26] F Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genet*, 9(3): e1003348, 2013. doi: 10.1371/journal.pgen.1003348.