A Vignette with Example Usage of SAMBA-EHR

Dr. Lauren J Beesley
Department of Biostatistics, University of Michigan
Contact: lbeesley@umich.edu

March 27, 2019

We consider the setting in which we want to study the relationship between D^*, a potentially misclassified electronic health record-derived binary phenotype, and G, genetic information. Common practice is to study this relationship by fitting a logistic regression (or in some cases, linear regression, which we will not consider here) for D^* given G and possibly other covariates Z on the sampled dataset. This approach implicitly assumes that $D* = D$, where D is the true disease status. Additionally, in order to generalize to the general population, we make implicit assumptions on the sampling mechanism.

When we have misclassification of the phenotype with respect to the true disease status and/or when sampling depends on the underlying disease status, this standard approach may produce biased estimates for the quantity of interest. The goal of SAMBA-EHR is to enable users to understand the degree of bias that may be introduced by the standard analysis approach. SAMBA-EHR allows users to perform sensitivity analysis after the primary analysis to explore the robustness of modeling results to different underlying sampling and misclassification mechanisms. In this document, we provide a short example of how this tool can be used.

1 Modeling Framework

The home page of the app provides information about the modeling framework and assumptions. Using notation on this page, θ_G is the quantity we are really interested in, and it represents the log-odds ratio relationship between D, the true lifetime disease status, and G. G can be a continuous variable such as a polygenic risk score (assumed to be mean-centered in the sample), or G can represent a single genetic locus or SNP coded 0/1/2 for the number of minor alleles.

Z represents the covariates adjusted for in the target model. We assume these covariates are mean-centered. We also make some assumptions regarding the relationships between X, Y, W, and G. These assumptions can be strong in some cases, and they can be weakened by including predictors contained in W, and X, and/or Y in Z. In other words, we only

need these independence assumptions when our simple regression model does not adjust for some of these factors.

2 Characterizing the Bias

While we are really interested in the Z-adjusted relationship between D and G in the general population, in practice we fit a model for misclassified phenotype D^* given G and Z on the sampled subjects. We then interpret the corresponding log-OR for G as θ_G. However, the resulting estimate of θ_G may be biased.

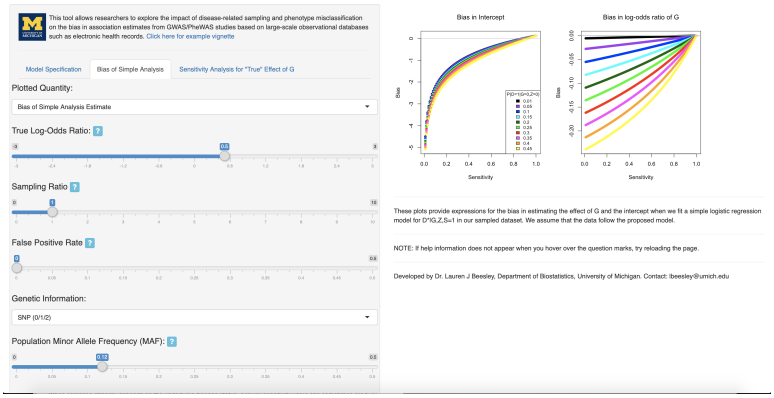
We can explore the potential degree of bias by clicking on the **"Bias of Simple Analysis"** tab. We can use the various sliders and drop-down menus on the left-hand side of the page to select scenarios of interest. For example, suppose G is a SNP with a population minor allele frequency (MAF) about about 12%, and we believe the true log-odds ratio of G is roughly 0.5. The sampling ratio represents the ratio of the average probability of being sampled in the diseased and non-diseased subjects. A sampling ratio of 1 indicates that sampling is unrelated to the true disease status. **Figure 1** shows the expected bias for different values of the false positive rate (1-specificity) and the sampling ratio.

Suppose the disease of interest is breast cancer, which has a lifetime prevalence rate of about 12% in the US. The blue lines present the degree of bias we expect in the intercept and log-odds ratio of G if we perform the simple analysis in this setting. Suppose we assume that we have a false positive rate of zero. This may be reasonable for a cancer phenotype, since the consequences of breast cancer diagnosis in terms of treatment and patient health can be severe. Suppose the sampling ratio equals 1. In this setting, the bias for the log-odds ratio of G increases as sensitivity decreases, but even extremely low sensitivity produces what might be viewed as a small absolute bias in estimating θ_G. Instead, suppose the sampling ratio is 5, so diseased subjects are 5 times more likely to be sampled on average than non-diseased subjects. Even with perfect sensitivity, the intercept estimate from the simple analysis is expected to be biased. As before, we see increased bias in estimating the log-odds ratio of G for decreased sensitivity, but the absolute bias is larger when the sampling ratio is 5 than when the true sampling ratio is 1. When sampling strongly depends on the underlying disease status, a lower sensitivity value can sometimes produce substantial bias in estimating θ_G.

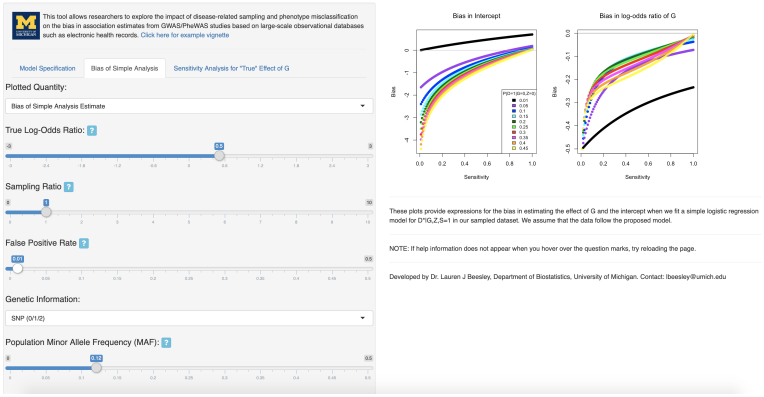
We can use this tool to explore how the degree of this bias depends on various model factors. We encourage you to play around and develop an intuition for when we expect large biases and when biases are expected to be much smaller. In particular, when we vary the false positive rates, the biases can have some very interesting properties. When we expect less bias, we may be less concerned about biases in the standard analysis approach.

Figure 1: Bias with sampling ratio of 1

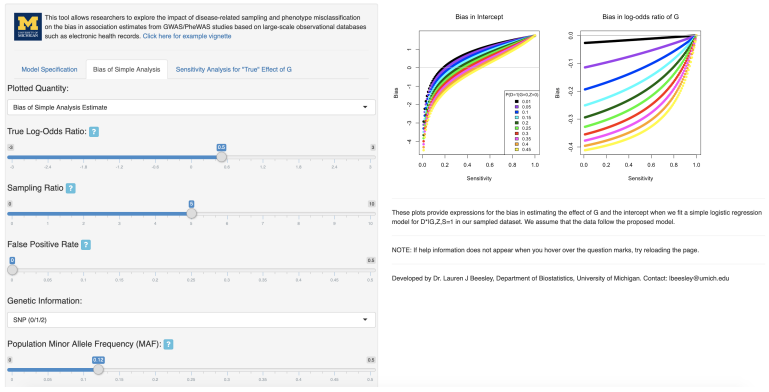
(a) Sampling Ratio = 1, Perfect Specificity



(b) Sampling Ratio = 1, Specificity = 0.99



(c) Sampling Ratio = 5, Perfect Specificity



3 Sensitivity Exploration after Primary Analysis

One question of interest is whether we need to "worry about" the misclassification and disease-dependent sampling in a given analysis. Restated, would accounting for these mechanisms give different results? If so, how different? The third tab of this online tool was developed to guide these explorations.

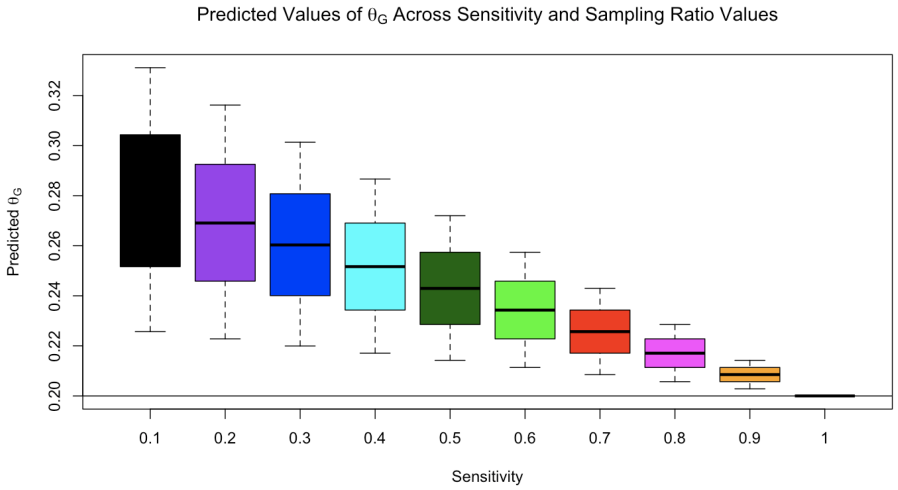
Suppose we believe that the sampling ratio might be between 1 and 5 based on our prior knowledge. Suppose we are again studying breast cancer, which has a population lifetime disease prevalence in the US around 12%. Suppose we are perform the "simple" standard analysis for a particular SNP with a minor allele frequency of 10%, and suppose we had an estimated log-odds ratio of 0.20 from this standard analysis. What are plausible values for θ_G? We can plug in the point estimate and both ends of the confidence intervals to see what types of values of θ_G could produce the observed estimate under different sensitivity levels. **Figure 2** shows predicted values of θ_G for the standard analysis point estimate of 0.20. Each boxplot corresponds to predictions for sampling ratio values between 1 and 5.

If we believe our sensitivity values are pretty high (e.g. greater than 0.7) and we have perfect specificity, we may not have too much cause to worry about bias induced by ignoring misclassification and disease-dependent sampling since the absolute bias is predicted to be fairly small. We may worry more if the sensitivity could be very low or if we have a lot of false positives.

The degree to which we worry about bias in a particular analysis setting will depend on the goals of the analysis, but this tool can help provide a first step in understanding the scale of the bias and what factors might impact this bias. We hope that this type of sensitivity analysis can be incorporated in standard EHR-based association analyses as a way to explore the robustness of results to violations of the common implicit assumptions of no disease-dependent sampling and no phenotype misclassification.

Figure 2: Predicted θ_G given simple standard analysis results

(a) Perfect Specificity



(b) Specificity = 0.95