DOMAIN CONSTRAINTS IMPROVE RISK PREDICTION WHEN OUTCOME DATA IS MISSING

Sidhika Balachandar *
Cornell Tech

Nikhil Garg
Cornell Tech

Emma Pierson
Cornell Tech

ABSTRACT

Machine learning models are often trained to predict the outcome resulting from a human decision. For example, if a doctor decides to test a patient for disease, will the patient test positive? A challenge is that historical decision-making determines whether the outcome is observed: we only observe test outcomes for patients doctors historically tested. Untested patients, for whom outcomes are unobserved, may differ from tested patients along observed and unobserved dimensions. We propose a Bayesian model class which captures this setting. The purpose of the model is to accurately estimate risk for both tested and untested patients. Estimating this model is challenging due to the wide range of possibilities for untested patients. To address this, we propose two domain constraints which are plausible in health settings: a *prevalence constraint*, where the overall disease prevalence is known, and an *expertise constraint*, where the human decision-maker deviates from purely risk-based decision-making only along a constrained feature set. We show theoretically and on synthetic data that domain constraints improve parameter inference. We apply our model to a case study of cancer risk prediction, showing that the model's inferred risk predicts cancer diagnoses, its inferred testing policy captures known public health policies, and it can identify suboptimalities in test allocation. Though our case study is in healthcare, our analysis reveals a general class of domain constraints which can improve model estimation in many settings.

1 INTRODUCTION

Machine learning models are often trained to predict outcomes in settings where a human makes a high-stakes decision. In criminal justice, a judge decides whether to release a defendant prior to trial, and models are trained to predict whether the defendant will fail to appear or commit a crime if released (Lakkaraju et al., 2017; Jung et al., 2020a; Kleinberg et al., 2018). In lending, a creditor decides whether to grant an applicant a loan, and models are trained to predict whether the applicant will repay (Björkegren & Grissen, 2020; Crook & Banasik, 2004). In healthcare—the setting we focus on in this paper—a doctor decides whether to test a patient for disease, and models are trained to predict whether the patient will test positive (Jehi et al., 2020; McDonald et al., 2021; Mullainathan & Obermeyer, 2022). Machine learning predictions help guide decision-making in all these settings. A model which predicts a patient's risk of disease can help allocate tests to the highest-risk patients, and also identify suboptimalities in human decision-making: for example, testing patients at low risk of disease, or failing to test high risk patients (Mullainathan & Obermeyer, 2022).

A fundamental challenge in all these settings is that historical decision-making determines whether the outcome is observed. In criminal justice, release outcomes are only observed for defendants judges have historically released. In lending, loan repayments are only observed for applicants historically granted loans. In healthcare, test outcomes are only observed for patients doctors have historically tested. This is problematic because the model must make accurate predictions for the entire population, not just the historically tested population. Learning only from the tested population also risks introducing bias against underserved populations who are less likely to get medical tests partly due to worse healthcare access (Chen et al., 2021; Pierson, 2020; Servik, 2020; Jain et al., 2023). Thus, there is a challenging distribution shift between the tested and untested populations. The

*Correspondence to: sidhikab@cs.cornell.edu

two populations may differ both along *observables* recorded in the data and *unobservables* known to the human decision-maker but unrecorded in the data. For example, tested patients may have more symptoms recorded than untested patients—but they may also differ on unobservables, like how much pain they are in or how sick they look, which are known to the doctor but are not available for the model. This setting, referred to as the *selective labels* setting (Lakkaraju et al., 2017), occurs in high-stakes domains including healthcare, hiring, insurance, lending, education, welfare services, government inspections, tax auditing, recommender systems, wildlife protection, and criminal justice and has been the subject of substantial academic interest (see §6 for related work).

Without further constraints on the data generating process, there is a wide range of possibilities for the untested patients. They could all have the disease or never have the disease. However, selective labels settings often have *domain-specific constraints* which would allow us to limit the range of possibilities. For example, in medical settings, we might know the prevalence of a disease in the population. Recent distribution shift literature has shown that generic methods generally do not perform well across all distribution shifts and that domain-specific constraints can improve generalization (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021; Sagawa et al., 2022; Gao et al.; Kaur et al., 2022; Tellez et al., 2019; Wiles et al., 2022). This suggests the utility of domain constraints in improving generalization from the tested to untested population.

Motivated by this reasoning, we make the following contributions:

1. We propose a Bayesian model class which captures the selective labels setting and nests classic econometric models. We model a patient's risk of disease as a function of observables and unobservables. The probability of testing a patient increases with disease risk and other factors (e.g., bias). The purpose of the model is to accurately estimate risk for both the tested and untested patients and to quantify deviations from purely risk-based test allocation.
2. We propose two constraints informed by the medical domain to improve model estimation: a *prevalence constraint*, where disease prevalence is known, and an *expertise constraint*, where the decision-maker deviates from risk-based decision-making along a constrained feature set. We show theoretically and on synthetic data that the constraints improve inference.
3. We apply our model to a breast cancer risk prediction case study. We conduct a suite of validations, showing that the model's (i) inferred risks predict cancer diagnoses, (ii) inferred unobservables correlate with known unobservables, (iii) inferred predictors of cancer risk correlate with known predictors, and (iv) inferred testing policy correlates with public health policies. We also show that our model identifies deviations from risk-based test allocation and that the prevalence constraint increases the plausibility of inferences.

Though our case study is in healthcare, our analysis reveals a general class of domain constraints which can improve model estimation in many selective labels settings.

2 MODEL

We now describe our Bayesian model class. Following previous work (Mullainathan & Obermeyer, 2022), our underlying assumption is that whether a patient is tested for a disease should be determined primarily by their risk of disease. Thus, the purpose of the model is to accurately estimate risk for both the tested and untested patients and to quantify deviations from purely risk-based test allocation. The latter task relates to literature on diagnosing factors affecting human decision-making (Mullainathan & Obermeyer, 2022; Zamfirescu-Pereira et al., 2022; Jung et al., 2018).

Consider a set of people indexed by i. For each person, we see observed features $X_i \in \mathbb{R}^D$ (e.g., demographics and symptoms in an electronic health record). We observe a *testing decision* $T_i \in \{0, 1\}$, where $T_i = 1$ indicates that the ith person was tested. If the person was tested ($T_i = 1$), we observe an outcome Y_i. Y_i might be a binary indicator (e.g. $Y_i = 1$ means that the person tests positive), or Y_i might be a numeric outcome of a medical test (e.g. T cell count or oxygen saturation levels). Throughout, we generally refer to Y_i as a binary indicator, but our framework extends to non-binary Y_i, and we derive our theoretical results in this setting. If $T_i = 0$ we do not observe Y_i.

There are *unobservables* (Angrist & Pischke, 2009; Rambachan et al., 2022), denoted by $Z_i \in \mathbb{R}$, that affect *both* T_i and Y_i but are not recorded in the dataset – e.g., whether the doctor observes that the

person is in pain. Consequently, the risk of the tested population differs from the untested population even conditional on observables X_i: i.e. $p(Y_i|T_i = 1, X_i) \neq p(Y_i|T_i = 0, X_i)$.

A person's risk of disease is captured by their *risk score* $r_i \in \mathbb{R}$, which is a function of X_i and Z_i. Whether the person is tested ($T_i = 1$) depends on their risk score r_i, but also factors like screening policies or socioeconomic disparities. More formally, our data generating process is

$$\begin{aligned} \text{Unobservables:} \quad & Z_i \sim f(\cdot|\sigma^2) \\ \text{Risk score:} \quad & r_i = X_i^T \boldsymbol{\beta_Y} + Z_i \\ \text{Test outcome:} \quad & Y_i \sim h_Y(\cdot|r_i) \\ \text{Testing decision:} \quad & T_i \sim h_T(\cdot|\alpha r_i + X_i^T \boldsymbol{\beta_\Delta}) \,. \end{aligned} \tag{1}$$

In words, Z_i is drawn from a distribution f with parameter σ^2, which captures the relative importance of the unobserved versus observed features. The disease risk score $r_i \in \mathbb{R}$ is modeled as a linear function of observed features (with unknown coefficients $\boldsymbol{\beta_Y} \in \mathbb{R}^D$) and the unobserved Z_i. Y_i is drawn from a distribution h_Y parameterized by r_i – e.g., $Y_i \sim \text{Bernoulli}(\text{sigmoid}(r_i))$. Analogously, the testing decision T_i is drawn from a distribution h_T parameterized by a linear function of the true disease risk score and other factors, with unknown coefficients $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta_\Delta} \in \mathbb{R}^D$. Because T_i depends on r_i, and r_i is a function of Z_i, T_i depends on Z_i. Figure 1 illustrates the effect of α and $\boldsymbol{\beta_\Delta}$. A larger α indicates that testing probability increases more steeply in risk. $\boldsymbol{\beta_\Delta}$ captures human or policy factors which affect a patient's probability of being tested beyond disease risk. In other words, $\boldsymbol{\beta_\Delta}$ captures deviations from purely risk-based test allocation. Putting things together, the model parameters are $\theta \triangleq (\alpha, \sigma^2, \boldsymbol{\beta_\Delta}, \boldsymbol{\beta_Y})$.

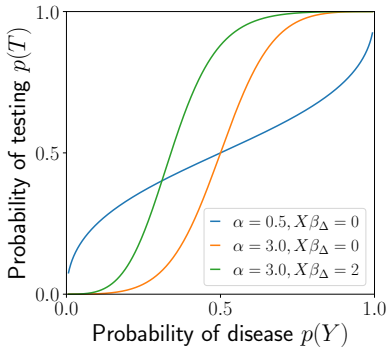


Figure 1: Effect of α and $X\boldsymbol{\beta_\Delta}$: α controls how steeply testing probability $p(T_i)$ increases in disease risk $p(Y_i)$, while $X\boldsymbol{\beta_\Delta}$ captures factors which affect $p(T_i)$ when controlling for $p(Y_i)$.

Medical domain knowledge: Besides the observed data, in medical settings we often have constraints to aid model estimation. We consider two constraints.

- **Prevalence constraint:** The average value of Y across the entire population is known ($\mathbb{E}[Y]$). When Y is a binary indicator of whether a patient has a disease, this corresponds to assuming that the *disease prevalence* is known. This assumption is plausible because estimating prevalence has been the focus of substantial public health research, and estimates thus exist in many medical settings; for more details see appendix A. For example, this information is available for cancer (Cancer Research UK), COVID-19 (NIH National Cancer Institute, 2023), and heart disease (CDC, 2007). In some cases, the prevalence is only *approximately* known (Manski & Molinari, 2021; Manski, 2020; Mullahy et al., 2021); our Bayesian formulation can incorporate such soft constraints as well.
- **Expertise constraint:** Because doctors and patients are informed decision-makers, we can assume that tests are allocated *mostly* based on disease risk. Specifically, we assume that there are some features which do not affect a patient's probability of receiving a test when controlling for their risk: i.e., that $\boldsymbol{\beta_\Delta}_d = 0$, for at least one dimension d. For example, we may assume that when controlling for disease risk, a patient's height does not affect their probability of being tested for cancer, and thus $\boldsymbol{\beta_\Delta}_{\text{height}} = 0$.

3 Theoretical Analysis

In this section, we prove why our proposed constraints improve parameter inference by analyzing a special case of our general model in equation 1. In Proposition 3.1, we show that this special case is equivalent to the Heckman model (Heckman, 1976; 1979), which is used to correct bias from non-randomly selected samples. In Proposition 3.2, we analyze this model to show that constraints can improve the precision of parameter inference. The full proofs are in Appendix B. In Sections 4 and 5 we empirically generalize our theoretical results beyond the special Heckman case.

3.1 DOMAIN CONSTRAINTS CAN IMPROVE THE PRECISION OF PARAMETER INFERENCE

We start by defining the Heckman model and showing it is a special case of our general model.

Definition 1 (Heckman correction model)**.** *The Heckman model can be written in the following form (Hicks, 2021):*

$$\begin{aligned} T_i &= \mathbb{1}[X_i^T \tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} + u_i > 0] \\ Y_i &= X_i^T \tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}} + Z_i \\ \begin{bmatrix} u_i \\ Z_i \end{bmatrix} &\sim \textit{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \tilde{\rho} \\ \tilde{\rho} & \tilde{\sigma}^2 \end{bmatrix} \right). \end{aligned} \tag{2}$$

Proposition 3.1. *The Heckman model (Definition 1) is equivalent to the following special case of the general model in equation 1:*

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, \sigma^2) \\ r_i &= X_i^T \boldsymbol{\beta}_{\boldsymbol{Y}} + Z_i \\ Y_i &= r_i \\ T_i &\sim \textit{Bernoulli}(\Phi(\alpha r_i + X_i^T \boldsymbol{\beta}_{\boldsymbol{\Delta}})). \end{aligned} \tag{3}$$

It is known that the Heckman model is identifiable (Lewbel, 2019), and thus the special case of our model is identifiable (i.e., distinct parameter sets correspond to distinct observed expectations) without further constraints. However, past work has often placed constraints on the Heckman model (though different constraints from those we propose) to improve parameter inference. Without constraints, the model is only weakly identified by functional form assumptions (Lewbel, 2019). This suggests that our proposed constraints could also improve model estimation. In Proposition 3.2, we make this intuition precise by showing that our proposed constraints improve the *precision* of the parameter estimates as measured by the *variance* of the parameter posteriors.

In our Bayesian formulation, we estimate a posterior distribution for parameter θ given the observed data: $g(\theta) \triangleq p(\theta|X, T, Y)$. Let $\text{Var}(\theta)$ denote the variance of $g(\theta)$. We show that constraining the value of any one parameter *will not worsen* the precision with which other parameters are inferred. In particular, constraining a parameter θ_{con} to a value drawn from its posterior distribution will not in expectation increase the posterior variance of any other unconstrained parameters θ_{unc}. To formalize this, we define the *expected conditional variance*:

Definition 2 (Expected conditional variance)**.** *Let the distribution over model parameters $g(\theta) \triangleq p(\theta|X, T, Y)$ be the posterior distribution of the parameters θ given the observed data $\{X, T, Y\}$. We define the expected conditional variance of an unconstrained parameter θ_{unc}, conditioned on the value of a constrained parameter θ_{con}, to be $\mathbb{E}[Var(\theta_{unc}|\theta_{con})] \triangleq \mathbb{E}_{\theta^*_{con} \sim g}[Var(\theta_{unc}|\theta_{con} = \theta^*_{con})]$.*

Proposition 3.2. *In expectation, constraining the parameter θ_{con} does not increase the variance of any other parameter θ_{unc}. In other words, $\mathbb{E}[Var(\theta_{unc}|\theta_{con})] \leq Var(\theta_{unc})$. Moreover, the inequality is strict as long as $\mathbb{E}[\theta_{unc}|\theta_{con}]$ is non-constant in θ_{con} (i.e., $Var(\mathbb{E}[\theta_{unc}|\theta_{con}]) > 0$).*

In other words, we reason about the effects of fixing a parameter θ_{con} to its true value θ^*_{con}. That value θ^*_{con} is distributed according to the posterior distribution g, and so we reason about expectations over g. In expectation, fixing the value of θ_{con} does not increase the variance of any other parameter θ_{unc}, and strictly reduces it as long as the expectation of θ_{unc} is non-constant in θ_{con}.

Both the expertise and prevalence constraints fix the value of at least one parameter. The expertise constraint fixes the value of $\boldsymbol{\beta}_{\boldsymbol{\Delta}\,d}$ for some d. For the Heckman model, the prevalence constraint fixes the value of the intercept $\boldsymbol{\beta}_{\boldsymbol{Y}\,0}$ (assuming the standard condition that columns of X are zero-mean except for an intercept column of ones). Thus, Proposition 3.2 implies that both constraints will not increase the variance of other model parameters, and will strictly reduce it as long as the posterior expectations of the unconstrained parameters are non-constant in the constrained parameters. In Appendix B we prove Proposition 3.2 and provide conditions under which the constraints strictly reduce the variance of other model parameters. We also verify and extend these theoretical results on synthetic data (Appendix D.1 Figure S1).

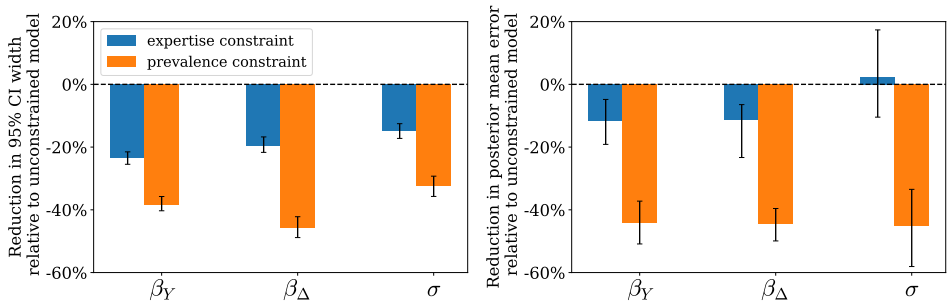


Figure 2: The prevalence and expertise constraints each produce more precise and accurate inferences on synthetic data drawn from the Bernoulli-sigmoid model with uniform noise (equation 4). To quantify precision (left), we report the percent reduction in 95% confidence interval width as compared to the unconstrained model. To quantify accuracy (right), we report the percent reduction in posterior mean error — i.e., the absolute difference between the posterior mean and the true parameter value — as compared to the unconstrained model. We plot the median across 200 synthetic datasets. Error bars denote the bootstrapped 95% confidence interval on the median.

3.2 Empirical extension beyond the Heckman special case

While we derive our theoretical results for a special case of our general model class, in our experiments (§4 and §5) we validate they hold beyond this special case by using a Bernoulli-sigmoid model:

$$\begin{aligned} Z_i &\sim \text{Uniform}(0, \sigma^2) \\ r_i &= X_i^T \boldsymbol{\beta_Y} + Z_i \\ Y_i &\sim \text{Bernoulli}(\text{sigmoid}(r_i)) \\ T_i &\sim \text{Bernoulli}(\text{sigmoid}(\alpha r_i + X_i^T \boldsymbol{\beta_\Delta}))\,. \end{aligned} \tag{4}$$

We note two ways in which this model differs from the Heckman model. First, it uses a *binary* disease outcome Y because this is an appropriate choice for our breast cancer case study (§5). With a binary outcome, models are known to be more challenging to fit: one cannot simultaneously estimate α and σ, and models fit without constraints may fail to recover the correct parameters (StataCorp, 2023; Van de Ven & Van Praag, 1981; Toomet & Henningsen, 2008). Even in this more challenging case, we show that our proposed constraints improve model estimation. Second, this model uses a uniform distribution of unobservables instead of a normal distribution of unobservables. As we show in Appendix C, this choice allows us to marginalize out Z_i, greatly accelerating model-fitting.

4 Synthetic experiments

We now validate our proposed approach on synthetic data. Our theoretical results imply that our proposed constraints should reduce the variance of parameter posteriors (improving precision). We verify that this is the case. We also show empirically that the proposed constraints produce posterior mean estimates which lie closer to the true parameter values (improving accuracy).

In Appendix D.1, we show experimentally that these results hold for the Heckman special case of our general model. Here we show that our theoretical results apply beyond the Heckman special case by conducting experiments on models with binary outcomes and multiple noise distributions. For all experiments, we use the Bayesian inference package Stan (Carpenter et al., 2017), which uses the Hamiltonian Monte Carlo algorithm (Betancourt, 2017). We report results across 200 trials. For each trial, we generate a new dataset from the data generating process the model assumes; fit the model to that dataset; and evaluate model fit using two metrics: *precision* (width of the 95% confidence interval) and *accuracy* (difference between the posterior mean and the true parameter value). We wish to assess the effect of the constraints on model inferences. Thus, we compare inferences from models with: (i) no constraints (unconstrained); (ii) a prevalence constraint; and (iii) an expertise constraint on a subset of the features. Details are in Appendix D and the code is at `https://github.com/sidhikabalachandar/domain_constraints`.

Figure 2 shows results for the Bernoulli-sigmoid model with uniform unobservables (equation 4). Both constraints generally produce more precise and accurate inferences for all parameters relative to

the unconstrained model. The one exception is that the expertise constraint does not improve accuracy for σ^2. Overall, the synthetic experiments corroborate and extend the theoretical analysis, showing that the proposed constraints improve precision and accuracy of parameter estimates for several variants of our general model. (In Appendix D, we also provide results for other variants of our general model, including alternate distributions of unobservables (Figures S2 and S3); higher-dimensional features (Figure S4); and non-linear interactions between features (Figure S5).)

5 Real-world case study: Breast cancer testing

To demonstrate our model's applicability to healthcare settings, we apply it to a breast cancer testing dataset. In this setting, X_i consists of features capturing the person's demographics, genetics, and medical history; $T_i \in \{0, 1\}$ denotes whether a person has been tested for breast cancer; and $Y_i \in \{0, 1\}$ denotes whether the person is diagnosed with breast cancer. Our goal is to learn each person's risk of cancer—i.e., $p(Y_i = 1|X_i)$. We focus on a younger population (age ≤ 45) because it creates a challenging distribution shift between the tested and untested populations. Younger people are generally not tested for cancer (Cancer Research UK, 2023), so the tested population $(T_i = 1)$ may differ from the untested population, including on unobservables.

In the following sections, we describe our experimental set up and the model we fit (§5.1), we conduct four validations on the fitted model (§5.2), we use the model to assess historical testing decisions (§5.3), and we compare to a model fit without a prevalence constraint (§5.4).

5.1 Experimental setup

Our data comes from the UK Biobank (Sudlow et al., 2015), which contains information on health, demographics, and genetics for the UK (see Appendix E for details). We analyze 54,746 people by filtering for women under the age of 45 (there is no data on breast cancer tests for men). For each person, X_i consists of 7 health, demographic, and genetic features found to be predictive of breast cancer (NIH National Cancer Institute, 2017; Komen, 2023; Yanes et al., 2020). $T_i \in \{0, 1\}$ denotes whether the person receives a mammogram (the most common breast cancer test) in the 10 years following measurement of features. $Y_i \in \{0, 1\}$ denotes whether the person is diagnosed with breast cancer in the 10 year period. $p(T = 1) = 0.51$ and $p(Y = 1|T = 1) = 0.03$.[1]

As in the synthetic experiments, we fit the Bernoulli-sigmoid model with uniform unobservables (equation 4). We include a prevalence constraint $\mathbb{E}[Y] = 0.02$, based on previously reported breast cancer incidence statistics (Cancer Research UK). We also include an expertise constraint by allowing $\boldsymbol{\beta}_{\boldsymbol{\Delta}}$ to deviate from 0 only for features which plausibly influence a person's probability of being tested beyond disease risk. We do not place the expertise constraint on (i) racial/socioeconomic features, due to disparities in healthcare access (Chen et al., 2021; Pierson, 2020; Shanmugam & Pierson, 2021); (ii) genetic features, since genetic information may be unknown or underused (Samphao et al., 2009); and (iii) age, due to age-based breast cancer testing policies (Cancer Research UK, 2023). In Appendix F.2 Figures S7, S8, and S9, we run robustness experiments.

In Figure 3, we plot the inferred coefficients for the fitted model. The model infers a large $\sigma^2 = 5.1$ (95% CI, 3.7-6.8), highlighting the importance of unobservables. In Appendix F.1 Figure S6, we also compare our model's performance to a suite of additional baselines, including (i) baselines trained solely on the tested population, (ii) baselines which treat the untested population as negative, and (iii) additional baselines commonly used in selective labels settings (Rastogi et al., 2023). Collectively, these baselines all suffer from various issues our model does not, including learning implausible age trends inconsistent with prior literature or worse predictive performance.

5.2 Validating the model

Validating models in real-world selective labels settings is difficult because outcomes are not observed for the untested. Still, we leverage the rich data in the UK Biobank to validate our model in four ways.

[1]We verify that very few people in the dataset have $T = 0$ and $Y = 1$ (i.e., are diagnosed with no record of a test): $p(Y = 1|T = 0) = 0.0005$. We group these people with the untested $T = 0$ population, since they did not receive a breast cancer test.

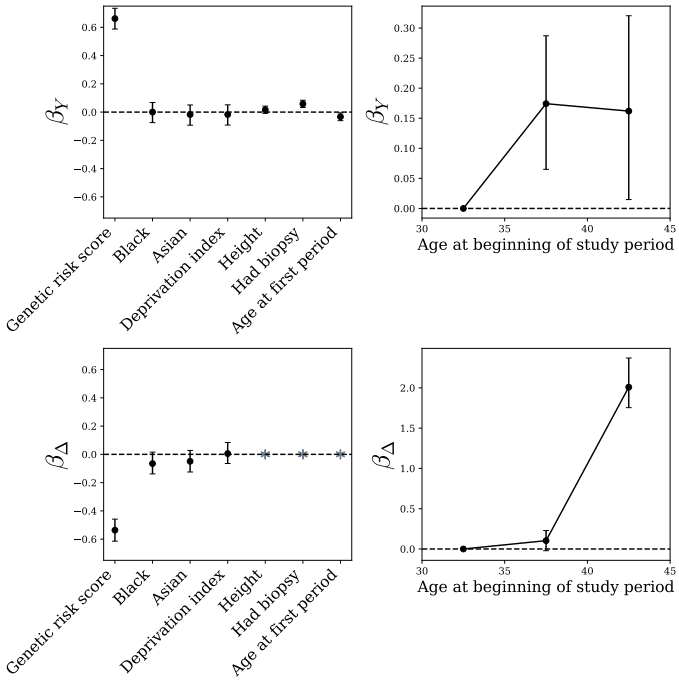


Figure 3: Estimated $\boldsymbol{\beta_Y}$ (top) capture known cancer risk factors: genetic risk, previous biopsy, age at first period (menarche), and age (NIH National Cancer Institute, 2017; Yanes et al., 2020). Estimated $\boldsymbol{\beta_\Delta}$ (bottom) capture the underuse of genetic information (left) and known age-based testing policies (right). Points indicate posterior means and vertical lines indicate 95% confidence intervals. Gray asterisks indicate coefficients set to 0 by the expertise constraint.

Inferred risk predicts breast cancer diagnoses: Verifying that inferred risk predicts diagnoses among the *tested* population is straightforward. Since Y is observed for the tested population, we check (on a test set) whether people with higher inferred risk ($p(Y_i = 1|X_i)$) are more likely to be diagnosed with cancer ($Y_i = 1$). People in the highest inferred risk quintile[2] have $3.3\times$ higher true risk of cancer than people in the lowest quintile (6.0% vs 1.8%). Verifying that inferred risk predicts diagnoses among the *untested* population is less straightforward because Y_i is not observed. We leverage that a subset have a *follow-up* visit (i.e., an observation after the initial 10-year study period) to show that inferred risk predicts cancer diagnosis at the follow-up. For the subset of the untested population who attend a follow-up visit, people in the highest inferred risk quintile have $2.5\times$ higher true risk of cancer during the follow-up period than people in the lowest quintile (4.1% vs 1.6%).[3]

Inferred unobservables correlate with known unobservables: For each person, our model infers a posterior over unobservables $p(Z_i|X_i, T_i, Y_i)$. We confirm that the inferred posterior mean of unobservables correlates with a true unobservable—whether the person has a family history of breast cancer. This is an unobservable because it influences both T_i and Y_i but is not included in the data given to the model.[4] People in the highest inferred unobservables quintile are $2.1\times$ likelier to have a family history of cancer than people in the lowest quintile (15.6% vs 7.5%).

$\boldsymbol{\beta_Y}$ captures known cancer risk factors: $\boldsymbol{\beta_Y}$ measures each feature's contribution to risk. The top left plot in Figure 3 shows that the inferred $\boldsymbol{\beta_Y}$ captures known cancer risk factors. Cancer risk is strongly correlated with genetic risk, and is also correlated with previous breast biopsy, age, and younger age at first period (menarche) (NIH National Cancer Institute, 2017; Yanes et al., 2020).

[2]Reporting outcome rates by inferred risk quintile or decile is a common metric in health risk prediction settings (Mullainathan & Obermeyer, 2022; Einav et al., 2018; Obermeyer et al., 2019).

[3]AUC amongst the tested population is 0.63 and amongst the untested population that attended a followup is 0.63. These AUCs are similar to past predictions which use similar feature sets (Yala et al., 2021). For instance, the Tyrer-Cuzick (Tyrer et al., 2004) and Gail (Gail et al., 1989) models achieved AUCs of 0.62 and 0.59.

[4]Although UKBB has family history data, we do not include it as a feature both so we can use it as validation and because we do not have information on *when* family members are diagnosed. So we cannot be sure that the measurement of family history precedes the measurement of T_i and Y_i, as is desirable for features in X_i.

β_Δ captures known public health policies: In the UK, all women aged 50-70 are invited for breast cancer testing every 3 years (Cancer Research UK, 2023). Our study period spans 10 years, so we expect women who are 40 or older at the start of the study period (50 or older at the end) to have an increased probability of testing when controlling for true cancer risk. The bottom right plot in Figure 3 shows this is the case, since the β_Δ indicator for ages 40-45 is greater than the indicators for ages <35 and 35-39.

5.3 Assessing historical testing decisions

Non-zero components of β_Δ indicate features that affect a person's probability of being tested even when controlling for their disease risk. The bottom left plot in Figure 3 plots the inferred β_Δ, revealing that genetic information is underused. While genetic risk is strongly predictive of Y_i, its negative β_Δ indicates that people at high genetic risk are tested less than expected given their risk. This is plausible, given that their genetic information may not have been available to guide decision-making. The model also infers negative point estimates for β_Δ for Black and Asian women, consistent with known racial disparities in breast cancer testing (Makurumidze et al., 2022) as well as broader racial inequality in healthcare and other domains (Nazroo et al., 2007; Zink et al., 2023; Movva et al., 2023; Obermeyer et al., 2019; Franchi et al., 2023; Otu et al., 2020; Devonport et al., 2023). However, both confidence intervals overlap zero (due to the small size of these groups in our dataset).

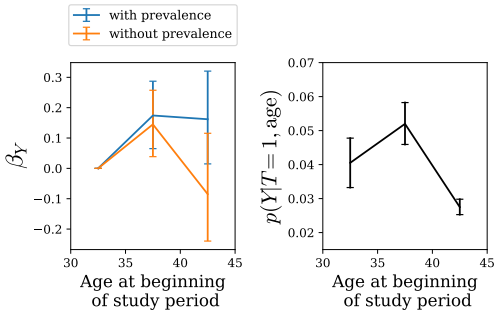


Figure 4: Without the prevalence constraint, the model learns that cancer risk first increases and then decreases with age (left orange), contradicting prior literature (Komen, 2023; Cancer Research UK; US Cancer Statistics Working Group et al., 2013; Campisi, 2013). This incorrect inference occurs because the tested population has the same misleading age trend (right). In contrast, the prevalence constraint encodes that the (younger) untested population has lower risk, allowing the model to learn a more accurate age trend (left blue).

5.4 Comparison to model without prevalence constraint

The prevalence constraint also guides the model to more plausible inferences. We compare the model fit with and without a prevalence constraint. As shown in the left plot in Figure 4, without the prevalence constraint, the model learns that cancer risk first increases with age and then falls, contradicting prior epidemiological and physiological evidence (Komen, 2023; Cancer Research UK; US Cancer Statistics Working Group et al., 2013; Campisi, 2013). This is because, due to the age-based testing policy in the UK (Cancer Research UK, 2023), being tested for breast cancer before age 50 is unusual. Thus, the tested population under age 50 is non-representative because their risk is much higher than the corresponding untested population. The prevalence constraint guides the model to more plausible inferences by preventing the model from predicting that a large fraction of the untested (younger) population has the disease.

6 Related work

Selective labels problems occur in many domains, including hiring, insurance, government inspections, tax auditing, recommender systems, lending, healthcare, education, welfare services, wildlife protection, and criminal justice (Lakkaraju et al., 2017; Jung et al., 2020a; Kleinberg et al., 2018; Björkegren & Grissen, 2020; Jung et al., 2018; Jehi et al., 2020; McDonald et al., 2021; Laufer et al.; McWilliams et al., 2019; Crook & Banasik, 2004; Hong et al., 2018; Parker et al., 2019; Sun et al., 2011; Kansagara et al., 2011; Waters & Miikkulainen, 2014; Bogen, 2019; Jawaheer et al., 2010; Wu et al., 2017; Coston et al., 2020; De-Arteaga et al., 2021; Pierson, 2020; Pierson et al., 2020; Simoiu et al., 2017; Mullainathan & Obermeyer, 2022; Henderson et al., 2022; Gholami et al., 2019; Farahani et al., 2020; Liu & Garg, 2022; Cai et al., 2020; Daysal et al., 2022; Guerdan et al., 2023; Chan et al., 2022; Jiang et al., 2021; Chien et al., 2023; Jia et al., 2019). As such, there are related literatures

in machine learning and causal inference (Coston et al., 2020; Schulam & Saria, 2017; Lakkaraju et al., 2017; Kleinberg et al., 2018; Shimodaira, 2000; De-Arteaga et al., 2021; Levine et al., 2020; Koh et al., 2021; Sagawa et al., 2022; Kaur et al., 2022; Sahoo et al., 2022; Cortes-Gomez et al., 2023), econometrics (Mullainathan & Obermeyer, 2022; Rambachan et al., 2022; Heckman, 1976; Hull, 2021; Künzel et al., 2019; Shalit et al., 2017; Wager & Athey, 2018; Alaa & Schaar, 2018), statistics and Bayesian models (Ilyas et al., 2020; Daskalakis et al., 2021; Mishler & Kennedy, 2022; Jung et al., 2020b), and epidemiology (Groenwold et al., 2012; Perkins et al., 2018). We extend this literature by providing constraints which both theoretically and empirically improve parameter inference. We now describe the three lines of work most closely related to our modeling approach.

Generalized linear mixed models (GLMMs): Our model is closely related to GLMMs (Gelman et al., 2013; Stroup, 2012; Lum et al., 2022), which model observations as a function of both observed features X_i and unobserved "random effects" Z_i. We extend this literature by (i) proposing and analyzing a novel model to capture our selective labels setting; (ii) incorporating the uniform distribution of unobservables, as opposed to the normal distribution typically used in GLMMs, to yield more tractable inference; and most importantly (iii) incorporating healthcare domain constraints into GLMMs to improve model estimation.

Improving robustness to distribution shift using domain information: The selective labels setting represents a specific type of distribution shift from the tested to untested population. Previous work shows that generic methods often fail to perform well across all types of distribution shifts (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021; Sagawa et al., 2022; Wiles et al., 2022; Kaur et al., 2022) and that incorporating domain information can improve performance. Gao et al. proposes *targeted augmentations*, which augment the data by randomizing known spurious features while preserving robust ones. Tellez et al. (2019) presents an example of this strategy for histopathology slide analysis. Kaur et al. (2022) shows that modeling the data generating process is necessary for generalizing across distribution shifts. Motivated by this, we propose a data generating process suitable for selective labels settings and show that using domain information improves performance.

Breast cancer risk estimation: There are many related works on estimating breast cancer risk (Daysal et al., 2022; Yala et al., 2019; 2021; 2022; Shen et al., 2021). Our work complements this literature by proposing a Bayesian model which captures the selective labels setting and incorporating domain constraints to improve model estimation. While a linear model suffices for the low-dimensional features used in our case study, our approach naturally extends to more complex inputs (e.g., medical images) and deep learning models sometimes used in breast cancer risk prediction (Yala et al., 2019; 2021; 2022).

7 Discussion

We propose a Bayesian model class to infer risk and assess historical human decision-making in selective labels settings, which commonly occur in healthcare and other domains. We propose the prevalence and expertise constraints which we show both theoretically and empirically improve parameter inference. We apply our model to cancer risk prediction, validate its inferences, show it can identify suboptimalities in test allocation, and show the prevalence constraint prevents misleading inferences.

A natural future direction is applying our model to other healthcare settings, where a frequent practice is to train risk-prediction models only on the tested population (Jehi et al., 2020; McDonald et al., 2021; Farahani et al., 2020). This is far from optimal both because only a small fraction of the population is tested, increasing variance, and because the tested population is highly non-representative, increasing bias. The paradigm we propose offers a solution to both problems. Using data from the entire population reduces variance, and modeling the distribution shift and constraining inferences on the untested population reduces bias. Beyond healthcare, other selective labels domains may have other natural domain constraints: for example, randomly assigned human decision-makers (Kleinberg et al., 2018) or repeated measurements of the same individual (Lum et al., 2022). Beyond selective labels, our model represents a concrete example of how domain constraints can improve inference in the presence of distribution shift.

ACKNOWLEDGMENTS

The authors thank Gabriel Agostini, Sivaramakrishnan Balachandar, Serina Chang, Erica Chiang, Avi Feller, Eran Halperin, Andrew Ilyas, Pang Wei Koh, Ben Laufer, Zhi Liu, Smitha Milli, Sendhil Mullainathan, Josue Nassar, Kenny Peng, Ashesh Rambachan, Richa Rastogi, Evan Rose, Shuvom Sadhuka, Jacob Steinhardt, Robert Tillman, and Manolis Zampetakis for helpful conversations. This research was supported by a Google Research Scholar award, NSF CAREER #2142419, a CIFAR Azrieli Global scholarship, Optum, a LinkedIn Research Award, the Abby Joseph Cohen Faculty Fund, and NSF GRFP Grant DGE #2139899. This research has been conducted using the UK Biobank Resource under Application Number 72589. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

REFERENCES

Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pp. 129–138. PMLR, 2018.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Katy J.L. Bell, Chris Del Mar, Gordon Wright, James Dickinson, and Paul Glasziou. Prevalence of incidental prostate cancer: A systematic review of autopsy studies. *International Journal of Cancer*, 137(7):1749–1757, 2015.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Daniel Björkegren and Darrell Grissen. Behavior revealed in mobile phone usage predicts credit repayment. *The World Bank Economic Review*, 34(3):618–634, 2020.

Miranda Bogen. All the ways hiring algorithms can introduce bias. *Harvard Business Review*, 6, 2019.

William Cai, Johann Gaebler, Nikhil Garg, and Sharad Goel. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 22–28, 2020.

Judith Campisi. Aging, cellular senescence, and cancer. *Annual Review of Physiology*, 75:685–705, 2013.

Cancer Research UK. Breast cancer statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer.

Cancer Research UK. Breast screening. https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/screening/breast-screening, 2023.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

CDC. Prevalence of heart disease–United States, 2005. *MMWR. Morbidity and Mortality Weekly Report*, 56(6):113–118, 2007.

David C. Chan, Matthew Gentzkow, and Chuan Yu. Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, 137(2):729–783, 2022.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4:123–144, 2021.

Jennifer Chien, Margaret Roberts, and Berk Ustun. Algorithmic censoring in dynamic learning systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23. Association for Computing Machinery, 2023.

Santiago Cortes-Gomez, Mateo Dulce, Carlos Patino, and Bryan Wilder. Statistical inference under constrained selection bias. *arXiv preprint arXiv:2306.03302*, 2023.

Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 582–593, 2020.

Jonathan Crook and John Banasik. Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, 28(4):857–874, 2004.

Constantinos Daskalakis, Patroklos Stefanou, Rui Yao, and Emmanouil Zampetakis. Efficient truncated linear regression with unknown noise variance. *Advances in Neural Information Processing Systems*, 34:1952–1963, 2021.

N. Meltem Daysal, Sendhil Mullainathan, Ziad Obermeyer, Suproteem K Sarkar, and Mircea Trandafir. An economic approach to machine learning in health policy. *Univ. of Copenhagen Dept. of Economics Discussion, CEBI Working Paper*, (24), 2022.

Maria De-Arteaga, Vincent Jeanselme, Artur Dubrawski, and Alexandra Chouldechova. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648*, 2021.

Tracey J. Devonport, Gavin Ward, Hana Morrissey, Christine Burt, J. Harris, S. Burt, R. Patel, R. Manning, R. Paredes, and W. Nicholls. A systematic review of inequalities in the mental health experiences of Black African, Black Caribbean and Black-mixed UK populations: Implications for action. *Journal of Racial and Ethnic Health Disparities*, 10(4):1669–1681, 2023.

Liran Einav, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer. Predictive modeling of US health care spending in late life. *Science*, 360(6396):1462–1465, 2018.

Nasibeh Zanjirani Farahani, Divaakar Siva Baala Sundaram, Moein Enayati, Shivaram Poigai Arunachalam, Kalyan Pasupathy, and Adelaide M Arruda-Olson. Explanatory analysis of a machine learning model to identify hypertrophic cardiomyopathy patients from EHR using diagnostic codes. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1932–1937. IEEE, 2020.

Matt Franchi, J.D. Zamfirescu-Pereira, Wendy Ju, and Emma Pierson. Detecting disparities in police deployments using dashcam data. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 534–544, 2023.

Mitchell H Gail, Louise A Brinton, David P Byar, Donald K Corle, Sylvan B Green, Catherine Schairer, and John J Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

Irena Gao, Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Out-of-domain robustness via targeted augmentations.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 2013.

Shahrzad Gholami, Lily Xu, Sara Mc Carthy, Bistra Dilkina, Andrew Plumptre, Milind Tambe, Rohit Singh, Mustapha Nsubuga, Joshua Mabonga, Margaret Driciru, et al. Stay ahead of poachers: Illegal wildlife poaching prediction and patrol planning under uncertainty with field test evaluations. 2019.

Rolf H.H. Groenwold, A. Rogier T. Donders, Kit C.B. Roes, Frank E. Harrell Jr, and Karel G.M. Moons. Dealing with missing outcome data in randomized trials and observational studies. *American Journal of Epidemiology*, 175(3):210–217, 2012.

Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. Counterfactual prediction under outcome measurement error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1584–1598. Association for Computing Machinery, 2023.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

James J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement*, volume 5, pp. 475–492. National Bureau of Economic Research, 1976.

James J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, pp. 153–161, 1979.

Peter Henderson, Ben Chugg, Brandon Anderson, Kristen Altenburger, Alex Turk, John Guyton, Jacob Goldin, and Daniel E. Ho. Integrating reward maximization and population estimation: Sequential decision-making for internal revenue service audit selection. *arXiv preprint arXiv:2204.11910*, 2022.

Rob Hicks. The Heckman sample selection model. https://econ.pages.code.wm.edu/407/notes/docs/index.html, 2021.

Woo Suk Hong, Adrian Daniel Haimovich, and R. Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PLOS One*, 13(7), 2018.

Peter Hull. What marginal outcome tests can tell us about racially biased decision-making. Technical report, National Bureau of Economic Research, 2021.

Andrew Ilyas, Emmanouil Zampetakis, and Constantinos Daskalakis. A theoretical and practical framework for regression and classification from truncated samples. In *International Conference on Artificial Intelligence and Statistics*, pp. 4463–4473. PMLR, 2020.

Simon Jackman. *Bayesian analysis for the social sciences*. John Wiley & Sons, 2009.

Sneha S. Jain, Tony Sun, Kathleen Brown, Vijendra Ramlall, Nicholas Tatonetti, Noémie Elhadad, Fatima Rodriguez, Ronald Witteles, Mathew S. Maurer, Timothy Poterucha, et al. Antiracist AI: A machine learning model using electrocardiograms and echocardiograms can detect transthyretin cardiac amyloidosis and decrease racial bias in diagnostic testing. *Journal of the American College of Cardiology*, 81:338–338, 2023.

Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. Comparison of implicit and explicit feedback from an online music recommendation service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pp. 47–51, 2010.

Lara Jehi, Xinge Ji, Alex Milinovich, Serpil Erzurum, Brian P. Rubin, Steve Gordon, James B. Young, and Michael W. Kattan. Individualizing risk prediction for positive coronavirus disease 2019 testing: Results from 11,672 patients. *Chest*, 158(4):1364–1375, 2020.

Xiwen Jia, Allyson Lynch, Yuheng Huang, Matthew Danielson, Immaculate Lang'at, Alexander Milder, Aaron E. Ruby, Hao Wang, Sorelle A. Friedler, Alexander J. Norquist, et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*, 573(7773): 251–255, 2019.

Heinrich Jiang, Qijia Jiang, and Aldo Pacchiano. Learning the truth from only one side of the story. In *International Conference on Artificial Intelligence and Statistics*, pp. 2413–2421. PMLR, 2021.

Lawrence Joseph, Theresa W. Gyorkos, and Louis Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–272, 1995.

Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*, 2018.

Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):771–800, 2020a.

Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. Bayesian sensitivity analysis for offline policy evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 64–70. Association for Computing Machinery, 2020b.

Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: A systematic review. *JAMA*, 306(15):1688–1698, 2011.

Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

Wei-Yin Ko, Konstantinos C Siontis, Zachi I Attia, Rickey E Carter, Suraj Kapa, Steve R Ommen, Steven J Demuth, Michael J Ackerman, Bernard J Gersh, Adelaide M Arruda-Olson, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *Journal of the American College of Cardiology*, 75(7):722–733, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Susan G. Komen. Factors linked to breast cancer risk. https://www.komen.org/breast-cancer/risk-factor/factors-that-affect-risk/, 2023.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

Benjamin Laufer, Emma Pierson, and Nikhil Garg. End-to-end auditing of decision pipelines.

Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Arthur Lewbel. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, 2019.

Zhi Liu and Nikhil Garg. Equity in resident crowdsourcing: Measuring under-reporting without ground truth data. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, pp. 1016–1017. Association for Computing Machinery, 2022.

Kristian Lum, David B. Dunson, and James Johndrow. Closer than they appear: A Bayesian perspective on individual-level heterogeneity in risk assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(2):588–614, 2022.

Getrude Makurumidze, Connie Lu, and Babagbemi Kemi. Addressing disparities in breast cancer screening: A review. *Applied Radiology*, 51(6):24–28, 2022.

Charles F. Manski. Bounding the accuracy of diagnostic tests, with application to COVID-19 antibody tests. *Epidemiology*, 32(2):162–167, 2020.

Charles F. Manski and Francesca Molinari. Estimating the COVID-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics*, 220(1):181–192, 2021.

Samuel A. McDonald, Richard J. Medford, Mujeeb A. Basit, Deborah B. Diercks, and D. Mark Courtney. Derivation with internal validation of a multivariable predictive model to predict COVID-19 test results in emergency department patients. *Academic Emergency Medicine*, 28(2):206–214, 2021.

Michael P. McLaughlin. *A compendium of common probability distributions*. Michael P. McLaughlin, 2001.

Christopher S. McMahan, Stella Self, Lior Rennert, Corey Kalbaugh, David Kriebel, Duane Graves, Cameron Colby, Jessica A. Deaver, Sudeep C. Popat, Tanju Karanfil, et al. COVID-19 wastewater epidemiology: A model to estimate infected populations. *The Lancet Planetary Health*, 5(12), 2021.

Christopher J. McWilliams, Daniel J. Lawson, Raul Santos-Rodriguez, Iain D. Gilchrist, Alan Champneys, Timothy H. Gould, Mathew J.C. Thomas, and Christopher P. Bourdeaux. Towards a decision support tool for intensive care discharge: Machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. *BMJ Open*, 9(3), 2019.

Alan Mishler and Edward H. Kennedy. FADE: Fair double ensemble learning for observable and counterfactual outcomes. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1053. Association for Computing Machinery, 2022.

Rajiv Movva, Divya Shanmugam, Kaihua Hou, Priya Pathak, John Guttag, Nikhil Garg, and Emma Pierson. Coarse race data conceals disparities in clinical risk score performance. *arXiv preprint arXiv:2304.09270*, 2023.

John Mullahy, Atheendar Venkataramani, Daniel L. Millimet, and Charles F. Manski. Embracing uncertainty: The value of partial identification in public health and clinical research. *American Journal of Preventive Medicine*, 61(2), 2021.

Sendhil Mullainathan and Ziad Obermeyer. Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727, 2022.

James Nazroo, James Jackson, Saffron Karlsen, and Myriam Torres. The Black diaspora and health inequalities in the US and England: Does where you go and how you get there make a difference? *Sociology of Health & Illness*, 29(6):811–830, 2007.

NIH National Cancer Institute. The breast cancer risk assessment tool. https://bcrisktool.cancer.gov/, 2017.

NIH National Cancer Institute. COVID-19 SeroHub. https://covid19serohub.nih.gov/, 2023.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Akaninyene Otu, Bright Opoku Ahinkorah, Edward Kwabena Ameyaw, Abdul-Aziz Seidu, and Sanni Yaya. One country, two crises: What COVID-19 reveals about health inequalities among BAME communities in the United Kingdom and the sustainability of its health system? *International Journal for Equity in Health*, 19(1):1–6, 2020.

Clare Allison Parker, Nan Liu, Stella Xinzi Wu, Yuzeng Shen, Sean Shao Wei Lam, and Marcus Eng Hock Ong. Predicting hospital admission at the emergency department triage: A novel prediction model. *The American Journal of Emergency Medicine*, 37(8):1498–1504, 2019.

Neil J. Perkins, Stephen R. Cole, Ofer Harel, Eric J. Tchetgen Tchetgen, BaoLuo Sun, Emily M. Mitchell, and Enrique F. Schisterman. Principled approaches to missing data in epidemiologic studies. *American Journal of Epidemiology*, 187(3):568–575, 2018.

Emma Pierson. Assessing racial inequality in COVID-19 testing with Bayesian threshold tests. *Machine Learning for Health (ML4H) at NeurIPS 2020 - Extended Abstract*, 2020.

Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In *International Conference on Artificial Intelligence and Statistics*, pp. 96–105. PMLR, 2018.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4 (7):736–745, 2020.

Ashesh Rambachan, Amanda Coston, and Edward H. Kennedy. Counterfactual risk assessments under unmeasured confounding. 2022.

Richa Rastogi, Michela Meister, UC Obermeyer, Jon Kleinberg, Pang Wei Koh, and Emma Pierson. Learn from the patient, not the doctor: Predicting downstream outcomes versus specialist labels. 2023. URL https://github.com/RichRast/Predicting_downstream_outcomes_vs_specialist_labels.

Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. In *International Conference on Learning Representations*, 2022.

Roshni Sahoo, Lihua Lei, and Stefan Wager. Learning from a biased sample. *CoRR*, abs/2209.01754, 2022.

Srila Samphao, Amanda J. Wheeler, Elizabeth Rafferty, James S. Michaelson, Michelle C. Specht, Michele A. Gadd, Kevin S. Hughes, and Barbara L. Smith. Diagnosis of breast cancer in women age 40 and younger: Delays in diagnosis result from underuse of genetic testing and breast imaging. *The American Journal of Surgery*, 198(4):538–543, 2009.

Steven J. Schrodi, Andrea DeBarber, Max He, Zhan Ye, Peggy Peissig, Jeffrey J. Van Wormer, Robert Haws, Murray H. Brilliant, and Robert D. Steiner. Prevalence estimation for monogenic autosomal recessive diseases using population-based genetic data. *Human Genetics*, 134:659–669, 2015.

Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. *Advances in Neural Information Processing Systems*, 30, 2017.

Kelly Servik. Huge hole in COVID-19 testing data makes it harder to study racial disparities. *Science*, 2020.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Divya Shanmugam and Emma Pierson. Quantifying inequality in underreported medical conditions. *arXiv preprint arXiv:2110.04133*, 2021.

Yiqiu Shen, Farah E. Shamout, Jamie R. Oliver, Jan Witowski, Kawshik Kannan, Jungkyu Park, Nan Wu, Connor Huddleston, Stacey Wolfson, Alexandra Millet, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nature Communications*, 12(1):1–13, 2021.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

StataCorp. *Stata 18 Base Reference Manual*, pp. 1089–1097. Stata Press, 2023.

Walter W. Stroup. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, 2012.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3), 2015.

Yan Sun, Bee Hoon Heng, Seow Yian Tay, and Eillyne Seow. Predicting hospital admissions at emergency department triage using routine administrative data. *Academic Emergency Medicine*, 18(8):844–850, 2011.

David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019.

Ott Toomet and Arne Henningsen. Sample selection models in R: Package sampleselection. *Journal of Statistical Software*, 27:1–23, 2008.

Peter Townsend, Peter Phillimore, and Alastair Beattie. *Health and Deprivation: Inequality and the North*. Routledge, 1988.

Jonathan Tyrer, Stephen W Duffy, and Jack Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine*, 23(7):1111–1130, 2004.

US Cancer Statistics Working Group et al. US cancer statistics: 1999–2009 incidence and mortality web-based report. *Atlanta GA: USDHHS, CDC and National Cancer Institute*, 2013.

Wynand P.M.M. Van de Ven and Bernard M.S. Van Praag. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics*, 17(2):229–252, 1981.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Austin Waters and Risto Miikkulainen. GRADE: Machine learning support for graduate admissions. *AI Magazine*, 35(1):64–64, 2014.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022.

Ann-Britt E. Wiréhn, H. Mikael Karlsson, and John M. Carstensen. Estimating disease prevalence using a population-based administrative healthcare database. *Scandinavian Journal of Public Health*, 35(4):424–431, 2007.

Mike Wu, Marzyeh Ghassemi, Mengling Feng, Leo A. Celi, Peter Szolovits, and Finale Doshi-Velez. Understanding vasopressor intervention and weaning: Risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association*, 24(3): 488–495, 2017.

Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60–66, 2019.

Adam Yala, Peter G. Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578), 2021.

Adam Yala, Peter G. Mikhael, Constance Lehman, Gigin Lin, Fredrik Strand, Yung-Liang Wan, Kevin Hughes, Siddharth Satuluru, Thomas Kim, Imon Banerjee, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nature Medicine*, 28(1):136–143, 2022.

Tatiane Yanes, Mary-Anne Young, Bettina Meiser, and Paul A. James. Clinical applications of polygenic breast cancer risk: A critical review and perspectives of an emerging field. *Breast Cancer Research*, 22(21), 2020.

J.D. Zamfirescu-Pereira, Jerry Chen, Emily Wen, Allison Koenecke, Nikhil Garg, and Emma Pierson. Trucks don't mean trump: Diagnosing human error in image analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 799–813, 2022.

Anna Zink, Ziad Obermeyer, and Emma Pierson. Race corrections in clinical models: Examining family history and cancer risk. *medRxiv*, pp. 2023–03, 2023.

A CALCULATING DISEASE PREVALENCE

To implement the prevalence constraint, we assume that the *disease prevalence*, or average value of Y across the population, is at least approximately known. This assumption is plausible in medical settings because estimating prevalence is the focus of substantial public health research. Methods to calculate prevalence include serology, where blood samples are used to detect specific antibodies or antigens of a disease (Joseph et al., 1995); stool or wastewater testing for disease markers (Joseph et al., 1995; McMahan et al., 2021); genetic methods, where genomic registries can be analyzed to calculate allele frequency and estimate disease prevalence (Schrodi et al., 2015); autopsy reports for a particular disease (Bell et al., 2015); and administrative data collected by primary, outpatient, and inpatient care centers (Wiréhn et al., 2007). Additionally, our Bayesian formulation can incorporate approximate prevalence estimates (e.g. bounded estimates), and these bounds can be estimated using the sensitivity and specificity of the prevalence estimation method (Manski & Molinari, 2021; Manski, 2020; Mullahy et al., 2021).

B PROOFS

Proof outline: In this section, we provide three proofs to show why domain constraints improve parameter inference. We start by showing that the well-studied Heckman correction model (Heckman, 1976; 1979) is a special case of the general model in equation 1 (Proposition 3.1). It is known that placing constraints on the Heckman model can improve parameter inference (Lewbel, 2019). We show that our proposed prevalence and expertise constraints have a similar effect by proving that our proposed constraints never worsen the precision of parameter inference (Proposition 3.2). We then provide conditions under which our constraints strictly improve precision (Proposition B.2).

Notation and assumptions: Below, we use Φ to denote the normal CDF, ϕ the normal PDF, and $\boldsymbol{\beta_T} = \alpha\boldsymbol{\beta_Y} + \boldsymbol{\beta_\Delta}$. Let X be the matrix of observable features. We assume that the first column of X corresponds to the intercept; X is zero mean for all columns except the intercept; and the standard identifiability condition that our data matrix is full rank, i.e., $X^T X$ is invertible. We also assume that $\alpha > 0$.

We start by defining the Heckman correction model.

Definition 1 (Heckman correction model)**.** *The Heckman model can be written in the following form (Hicks, 2021):*

$$\begin{aligned} T_i &= \mathbb{1}[X_i^T \tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} + u_i > 0] \\ Y_i &= X_i^T \tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}} + Z_i \\ \begin{bmatrix} u_i \\ Z_i \end{bmatrix} &\sim \textit{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \tilde{\rho} \\ \tilde{\rho} & \tilde{\sigma}^2 \end{bmatrix} \right). \end{aligned} \tag{2}$$

In other words, $T_i = 1$ if a linear function of X_i plus some unit normal noise u_i exceeds zero. Y_i is a linear function of X_i plus normal noise Z_i with variance $\tilde{\sigma}^2$. Importantly, the noise terms Z_i and u_i are *correlated*, with covariance $\tilde{\rho}$. The model parameters are $\tilde{\theta} \triangleq (\tilde{\rho}, \tilde{\sigma}^2, \tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}})$. We use tildes over the Heckman model parameters to distinguish them from the parameters in our original model in equation 1. We now prove Proposition 3.1.

Proposition 3.1. *The Heckman model (Definition 1) is equivalent to the following special case of the general model in equation 1:*

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, \sigma^2) \\ r_i &= X_i^T \boldsymbol{\beta_Y} + Z_i \\ Y_i &= r_i \\ T_i &\sim \textit{Bernoulli}(\Phi(\alpha r_i + X_i^T \boldsymbol{\beta_\Delta})). \end{aligned} \tag{3}$$

Proof. If we substitute in the value of r_i, the equation for Y_i is equivalent to that in the Heckman model. So it remains only to show that T_i in equation 3 can be rewritten in the form in equation 2.

We first rewrite equation 3 in slightly more convenient form:

$$
\begin{aligned}
&T_i \sim \text{Bernoulli}(\Phi(\alpha r_i + X_i^T \boldsymbol{\beta_\Delta})) \rightarrow \\
&T_i \sim \text{Bernoulli}(\Phi(\alpha(X_i^T \boldsymbol{\beta_Y} + Z_i) + X_i^T \boldsymbol{\beta_\Delta})) \rightarrow \\
&T_i \sim \text{Bernoulli}(\Phi(X_i^T(\alpha \boldsymbol{\beta_Y} + \boldsymbol{\beta_\Delta}) + \alpha Z_i)) \rightarrow \\
&T_i \sim \text{Bernoulli}(\Phi(X_i^T \boldsymbol{\beta_T} + \alpha Z_i))\,.
\end{aligned}
$$

We then apply the latent variable formulation of the probit link:

$$
\begin{aligned}
&T_i \sim \text{Bernoulli}(\Phi(X_i^T \boldsymbol{\beta_T} + \alpha Z_i)) \rightarrow \\
&T_i = \mathbb{1}[X_i^T \boldsymbol{\beta_T} + \alpha Z_i + \epsilon_i > 0], \epsilon_i \sim \mathcal{N}(0,1)\,,
\end{aligned}
$$

where $\alpha Z_i + \epsilon_i$ is a normal random variable with standard deviation $\sqrt{\alpha^2\sigma^2+1}$. We divide through by this factor to rewrite the equation for T_i:

$$
T_i = \mathbb{1}[X_i^T \tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} + u_i > 0]\,,
$$

which is equivalent to equation 2. Here, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} = \frac{\beta_T}{\sqrt{\alpha^2\sigma^2+1}}$ and $u_i = \frac{\alpha Z_i + \epsilon_i}{\sqrt{\alpha^2\sigma^2+1}}$ is a unit-scale normal random variable whose covariance with Z_i is

$$
\begin{aligned}
\text{cov}\left(\frac{\alpha Z_i + \epsilon_i}{\sqrt{\alpha^2\sigma^2+1}}, Z_i\right) &= \mathbb{E}\left(\frac{\alpha Z_i + \epsilon_i}{\sqrt{\alpha^2\sigma^2+1}} \cdot Z_i\right) - \mathbb{E}\left(\frac{\alpha Z_i + \epsilon_i}{\sqrt{\alpha^2\sigma^2+1}}\right)\mathbb{E}\left(Z_i\right) \\
&= \frac{\alpha \mathbb{E}\left(Z_i^2\right)}{\sqrt{\alpha^2\sigma^2+1}} \\
&= \frac{\alpha\sigma^2}{\sqrt{\alpha^2\sigma^2+1}}\,.
\end{aligned}
$$

Thus, the special case of our model in equation 3 is equivalent to the Heckman model, where the mapping between the parameters is:

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}} &= \boldsymbol{\beta_Y} \\
\tilde{\sigma}^2 &= \sigma^2 \\
\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} &= \frac{\boldsymbol{\beta_T}}{\sqrt{\alpha^2\sigma^2+1}} \\
\tilde{\rho} &= \frac{\alpha\sigma^2}{\sqrt{\alpha^2\sigma^2+1}}\,.
\end{aligned}
\tag{5}
$$

□

As described in Lewbel (2019), the Heckman correction model is identified without any further assumptions. It then follows that the special case of our model in equation 3 is identified without further constraints. One can simply estimate the Heckman model, which by the mapping in equation 5 immediately yields estimates of $\boldsymbol{\beta_Y}$ and σ^2. Then, the equation for $\tilde{\rho}$ can be solved for α, yielding a unique value since $\alpha > 0$. Similarly the equation for $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$ yields the estimate for $\boldsymbol{\beta_T}$ (and thus $\boldsymbol{\beta_\Delta}$).

While the Heckman model is identified without further constraints, this identification is known to be very weak, relying on functional form assumptions (Lewbel, 2019). To mitigate this problem, when the Heckman model is used in the econometrics literature it is typically estimated with constraints on the parameters. In particular, a frequently used constraint is an *exclusion restriction*: there must be at least one feature with a non-zero coefficient in the equation for T but not Y. While this constraint differs from the ones we propose, one might expect our proposed prevalence and expertise constraints to have a similar effect and improve the precision of parameter inference. We make this precise through Proposition 3.2.

Throughout the results below, we analyze the posterior distribution of model parameters given the observed data: $g(\theta) \triangleq p(\theta|X,T,Y)$. We show that constraining the value of any one parameter (through the prevalence or expertise constraint) will not worsen the posterior variance of the other parameters. In particular, constraining a parameter θ_{con} to a value drawn from its posterior distribution will not in expectation increase the posterior variance of any other unconstrained parameters θ_{unc}. To formalize this, we define the *expected conditional variance*:

Definition 2 (Expected conditional variance)**.** *Let the distribution over model parameters $g(\theta) \triangleq p(\theta|X,T,Y)$ be the posterior distribution of the parameters θ given the observed data $\{X,T,Y\}$. We define the expected conditional variance of an unconstrained parameter θ_{unc}, conditioned on the value of a constrained parameter θ_{con}, to be $\mathbb{E}[Var(\theta_{unc}|\theta_{con})] \triangleq \mathbb{E}_{\theta^*_{con}\sim g}[Var(\theta_{unc}|\theta_{con} = \theta^*_{con})]$.*

Proposition 3.2. *In expectation, constraining the parameter θ_{con} does not increase the variance of any other parameter θ_{unc}. In other words, $\mathbb{E}[Var(\theta_{unc}|\theta_{con})] \leq Var(\theta_{unc})$. Moreover, the inequality is strict as long as $\mathbb{E}[\theta_{unc}|\theta_{con}]$ is non-constant in θ_{con} (i.e., $Var(\mathbb{E}[\theta_{unc}|\theta_{con}]) > 0$).*

Proof. The proof follows from applying the law of total variance to the posterior distribution g. The law of total variance states that:

$$\text{Var}(\theta_{\text{unc}}) = \mathbb{E}[\text{Var}(\theta_{\text{unc}}|\theta_{\text{con}})] + \text{Var}(\mathbb{E}[\theta_{\text{unc}}|\theta_{\text{con}}])\,.$$

Since $\text{Var}(\mathbb{E}[\theta_{\text{unc}}|\theta_{\text{con}}])$ is non-negative,

$$\mathbb{E}[\text{Var}(\theta_{\text{unc}}|\theta_{\text{con}})] \leq \text{Var}(\theta_{\text{unc}})\,.$$

Additionally, if $\mathbb{E}[\theta_{\text{unc}}|\theta_{\text{con}}]$ is non-constant in θ_{con} then $\text{Var}(\mathbb{E}[\theta_{\text{unc}}|\theta_{\text{con}}])$ is strictly positive. Thus the strict inequality follows. □

We now discuss how Proposition 3.2 applies to our proposed constraints and the Heckman model. Both the prevalence and expertise constraints fix the value of at least one parameter. The prevalence constraint fixes the value of $\boldsymbol{\beta_Y}_0$ and the expertise constraint fixes the value of $\boldsymbol{\beta_\Delta}_d$ for some d. Thus by Proposition 3.2, we know that the prevalence and expertise constraints will not increase the variance of any model parameters, and will strictly reduce them as long as the posterior expectations of the unconstrained parameters are non-constant in the constrained parameters.

We now show that when $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$ is known, the prevalence constraint strictly reduces variance. The setting where $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$ is known is a natural one because $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$ can be immediately estimated from the observed data X and T, and previous work in both econometrics and statistics thus have also considered this setting (Heckman, 1976; Ilyas et al., 2020). With additional assumptions, we also show that the expertise constraint strictly reduces variance. We derive these results in the setting with flat priors for algebraic simplicity. However, analogous results also hold under other natural choices of prior (e.g., standard conjugate priors for Bayesian linear regression (Jackman, 2009)). In the results below, we analyze the conditional mean of Y conditioned on $T = 1$. Thus, we start by defining this value.

Lemma B.1 (Conditional mean of Y conditioned on $T = 1$)**.** *Past work has shown that the expected value of Y_i when $T_i = 1$ is (Hicks, 2021):*

$$\begin{aligned}\mathbb{E}[Y_i|T_i = 1] &= \mathbb{E}[Y_i|X_i^T\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} + u > 0] \\ &= X_i\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}} + \tilde{\rho}\tilde{\sigma}\frac{\phi(X_i\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}{\Phi(X_i\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}\,,\end{aligned}$$

where Φ denotes the normal CDF, ϕ the normal PDF, and $\frac{\phi(X\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}{\Phi(X\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}$ the inverse Mills ratio. This can be more succinctly represented in matrix notation as

$$\mathbb{E}[Y_i|T_i = 1] = M\theta\,,$$

where $M = [X_{T=1}; \frac{\phi(X_{T=1}\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}{\Phi(X_{T=1}\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}] \in \mathbb{R}^{N_{T=1}\times(d+1)}$, $\theta = [\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}}, \tilde{\rho}\tilde{\sigma}] \in \mathbb{R}^{d+1}$, $X_{T=1}$ denotes the rows of X corresponding to $T = 1$, and $N_{T=1}$ is the number of rows of X for which $T = 1$.

Proposition B.2. *Assume $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$ is fixed and flat priors on all parameters. Additionally, assume the standard identifiability condition that the matrix $M = [X_{T=1}; \frac{\phi(X_{T=1}\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}{\Phi(X_{T=1}\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}]$ is full rank. Then, in expectation, constraining a component of $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}}$ in the Heckman correction model strictly reduces the posterior variance of the other model parameters. The prevalence constraint does this without any further assumptions, and the expertise constraint does this if $\tilde{\rho}$ and $\tilde{\sigma}^2$ are fixed.*

Proof. We will start by showing that when $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$ is fixed, constraining a component of $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}}$ strictly reduces the variance of the other model parameters. From the definition of the conditional mean of Y conditioned on $T = 1$ (Lemma B.1), we get

$$\mathbb{E}[Y_i|T_i = 1] = M\theta\,.$$

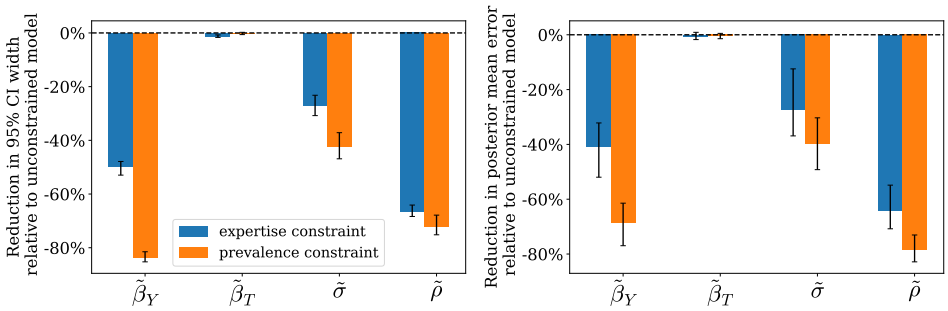


Figure S1: Results using synthetic data from the Heckman model. The prevalence and expertise constraints each produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

Under flat priors on all parameters, the posterior expectation of the model parameters given the observed data $\{X, T, Y\}$ is simply the standard ordinary least squares solution given by the normal equation (Jackman, 2009):

$$\mathbb{E}[\theta | X, T, Y] = (M^T M)^{-1} M^T Y \,.$$

By assumption, M is full rank, so $M^T M$ is invertible.

When $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}_d}$ is constrained to equal to $\tilde{\boldsymbol{\beta}}^*_{\boldsymbol{Y}_d}$ for some component d, the equation instead becomes:

$$\mathbb{E}[\theta_{-d} | \tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}_d} = \tilde{\boldsymbol{\beta}}^*_{\boldsymbol{Y}_d}, X, T, Y] = (M_{-d}^T M_{-d})^{-1} M_{-d}^T (Y - X_{T=1_d} \tilde{\boldsymbol{\beta}}^*_{\boldsymbol{Y}_d}) \,.$$

We use the subscript $-d$ notation to indicate that we no longer estimate the component d. Here, $M_{-d} = [X_{T=1_{-d}}; \frac{\phi(X_{T=1}\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}{\Phi(X_{T=1}\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}})}] \in \mathbb{R}^{N_{T=1} \times d}$ and $\theta_{-d} = [\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}_{-d}}, \tilde{\rho}\tilde{\sigma}] \in \mathbb{R}^d$. Since $X_{T=1_d}$ is nonzero and M is full rank, it follows that $\mathbb{E}[\theta_{-d} | \tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}_d} = \tilde{\boldsymbol{\beta}}^*_{\boldsymbol{Y}_d}, X, T, Y]$ is not constant in $\tilde{\boldsymbol{\beta}}^*_{\boldsymbol{Y}_d}$. Thus by Proposition 3.2, constraining $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}_d}$ reduces the variance of the parameters in θ_{-d} ($\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}'_d}$ for $d' \neq d$ and $\tilde{\rho}\tilde{\sigma}$).

We will now show that both the prevalence and expertise constraints constrain a component of $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}}$. Assuming the standard condition that columns of X are zero-mean except for an intercept column of ones, the prevalance constraint fixes

$$\begin{aligned} \mathbb{E}_Y[Y] &= \mathbb{E}_Y[\mathbb{E}_X[\mathbb{E}_Z[Y|X, Z]]] \\ &= \mathbb{E}_X[\mathbb{E}_Z[X^T \boldsymbol{\beta}_{\boldsymbol{Y}} + Z]] \\ &= \boldsymbol{\beta}_{\boldsymbol{Y}_0} \,, \end{aligned}$$

where $\boldsymbol{\beta}_{\boldsymbol{Y}_0}$ is the 0th index (intercept term) of $\boldsymbol{\beta}_{\boldsymbol{Y}}$. The expertise constraint also fixes a component of $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}}$ if $\tilde{\rho}$ and $\tilde{\sigma}^2$ are fixed. This can be shown by algebraically rearranging equation 5 to yield

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}} = \tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} \frac{\tilde{\sigma}^2}{\tilde{\rho}} - \boldsymbol{\beta}_{\boldsymbol{\Delta}} \frac{\tilde{\sigma}\sqrt{\tilde{\sigma}^2 - \tilde{\rho}^2}}{\tilde{\rho}} \,.$$

□

While we derive our theoretical results for the Heckman correction model, in both our synthetic experiments (§4) and our real-world case study (§5) we validate that our constraints improve parameter inference beyond the special Heckman case.

C Derivation of the closed-form uniform unobservables model

Conducting sampling for our general model described by equation 1 is faster if the distribution of unobservables f and link functions h_Y and h_T allow one to marginalize out Z_i through closed-form integrals, since otherwise Z_i must be sampled for each datapoint i, producing a high-dimensional

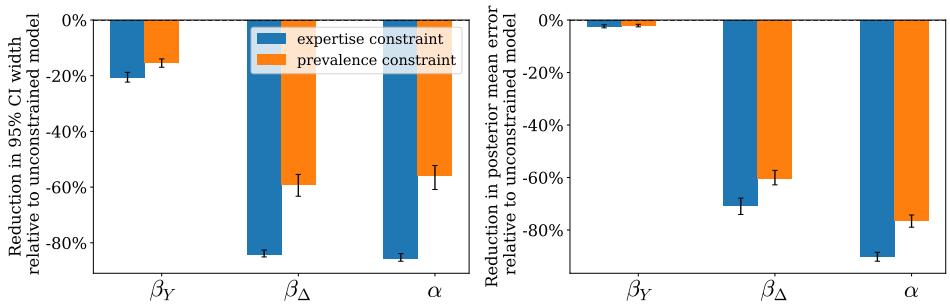


Figure S2: Results using synthetic data from the Bernoulli-sigmoid model with normal unobservables and fixed σ^2. The prevalence and expertise constraints each produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

latent variable which slows computation and convergence. Many distributions do not produce closed-form integrals when combined with a sigmoid or probit link function, which are two of the most commonly used links with binary variables.[5] However, we *can* derive closed forms for the special *uniform unobservables* case described by equation 4.

Below, we leave the i subscript implicit to keep the notation concise. When computing the log likelihood of the data, to marginalize out Z, we must be able to derive closed forms for the following three integrals:

$$p(Y=1,T=1|X) = \int_Z p(Y=1,T=1|X,Z)f(Z)dZ$$
$$p(Y=0,T=1|X) = \int_Z p(Y=0,T=1|X,Z)f(Z)dZ$$
$$p(T=0|X) = \int_Z p(T=0|X,Z)f(Z)dZ\,,$$

since the three possibilities for an individual datapoint are $\{Y=1,T=1\}$, $\{Y=0,T=1\}$, $\{T=0\}$. To implement the prevalence constraint (which fixes the $\mathbb{E}[Y]$), we also need a closed form for the following integral:

$$p(Y=1|X) = \int_Z p(Y=1|X,Z)f(Z)dZ\,.$$

For the uniform unobservables model with $\alpha=1$, the four integrals have the following closed forms, where below we define $A = e^{X^T\boldsymbol{\beta_T}}$ and $B = e^{X^T\boldsymbol{\beta_Y}}$:

$$\begin{aligned}
p(Y=1,T=1|X) = \frac{1}{\sigma\left(A-B\right)}\Big(&\sigma\left(A-B\right) - A\log\left(\left(B+1\right)A^{-1}\right)\\
&+ A\log\left(\left(Be^{\sigma}+1\right)A^{-1}e^{-\sigma}\right) + B\log\left(\left(A+1\right)A^{-1}\right)\\
&- B\log\left(\left(Ae^{\sigma}+1\right)A^{-1}e^{-\sigma}\right)\Big)\\
p(Y=0,T=1|X) = \frac{1}{\sigma\left(A-B\right)}\Big(&\big(-\log\left(\left(A+1\right)A^{-1}\right) + \log\left(\left(B+1\right)A^{-1}\right)\\
&+ \log\left(\left(Ae^{\sigma}+1\right)A^{-1}e^{-\sigma}\right) - \log\left(\left(Be^{\sigma}+1\right)A^{-1}e^{-\sigma}\right)\big)A\Big)\\
p(T=0|X) = &\frac{\log\left(1+A^{-1}\right) - \log\left(A^{-1}e^{-\sigma}+1\right)}{\sigma}\\
p(Y=1|X) = &\frac{\sigma - \log\left(1+B^{-1}\right) + \log\left(B^{-1}e^{-\sigma}+1\right)}{\sigma}\,.
\end{aligned}$$

The integrals also have closed forms for other integer values of α (e.g., $\alpha=2$) allowing one to perform robustness checks with alternate model specifications (see Appendix F.2 Figure S8).

[5] Specifically, we search over the distributions in McLaughlin (2001), combined with logit or probit links, and find that most combinations do not yield closed forms.

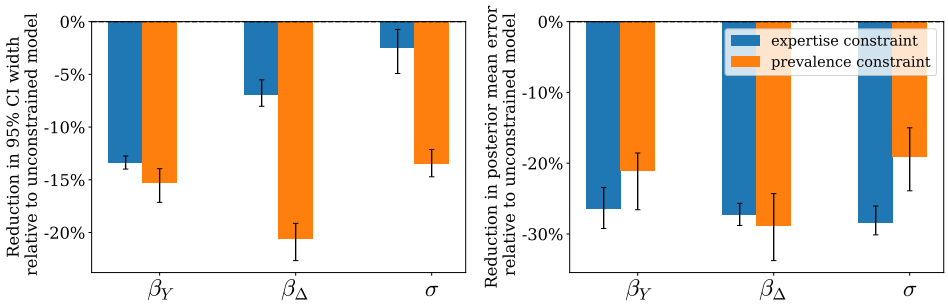


Figure S3: Results using synthetic data from the Bernoulli-sigmoid model with normal unobservables and fixed α. The prevalence and expertise constraints each produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

D SYNTHETIC EXPERIMENTS

We first validate that the prevalence and expertise constraints improve the precision and accuracy of parameter inference for the Heckman model described in equation 2. We then extend beyond this special case and examine various Bernoulli-sigmoid instantiations of our general model in equation 1, which assume a binary outcome variable Y. With a binary outcome, models are known to be more challenging to fit: for example, one cannot simultaneously estimate both α and σ^2 (so we must fix either α or σ^2), and models fit without constraints may fail to recover the correct parameters (StataCorp, 2023; Van de Ven & Van Praag, 1981; Toomet & Henningsen, 2008). We assess whether our proposed constraints improve model estimation even in this more challenging case. Specifically, we extend beyond the Heckman model to the following data generating settings: (i) uniform unobservables and fixed α, (ii) normal unobservables and fixed σ^2; (iii) normal unobservables and fixed α; and (iv) other more complex models. For the uniform model, we conduct experiments only with fixed α (not fixed σ^2) because, as discussed above, this allows us to marginalize out Z.

In all models, to incorporate the prevalence constraint into the model, we add a quadratic penalty to the model penalizing it for inferences that produce an inferred $\mathbb{E}[Y]$ that deviates from the true $\mathbb{E}[Y]$. To incorporate the expertise constraint into the model, we set the model parameters $\boldsymbol{\beta}_{\Delta_d}$ to be equal to 0 for all dimensions d to which the expertise constraint applies.

D.1 HECKMAN MODEL

We first conduct synthetic experiments using the Heckman model defined in equation 2. This model is identifiable without any further constraints, thus we estimate parameters $\theta \triangleq (\tilde{\rho}, \tilde{\sigma}^2, \tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}})$.

In the simulation, we use 5000 datapoints; 5 features (including the intercept column of 1s); X, $\boldsymbol{\beta}_Y$, and $\boldsymbol{\beta}_T$ drawn from unit normal distributions; and $\sigma \sim \mathcal{N}(2, 0.1)$. We draw the intercept terms $\boldsymbol{\beta}_{Y_0} \sim \mathcal{N}(-2, 0.1)$ and $\boldsymbol{\beta}_{T_0} \sim \mathcal{N}(2, 0.1)$. We assume the expertise constraint applies to $\boldsymbol{\beta}_{\Delta_2} = \boldsymbol{\beta}_{\Delta_3} = \boldsymbol{\beta}_{\Delta_4} = 0$. Thus, by rearranging equation 5, we fix $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}} = \tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} \frac{\tilde{\sigma}^2}{\tilde{\rho}}$. When calculating the results for $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$ and $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}}$, we do not include the dimensions along which we assume expertise since these dimensions are assumed to be fixed for the model with the expertise constraint.

We show results in Figure S1. Both constraints generally produce more precise and accurate inferences for all parameters relative to the unconstrained model. The only exception is $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$, for which both models produce equivalently accurate and precise inferences. This is consistent with our theoretical results, which do not imply that the precision of inference for $\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}}$ should improve.

D.2 UNIFORM UNOBSERVABLES MODEL

We now discuss our synthetic experiments using the Bernoulli-sigmoid model with uniform unobservables and $\alpha = 1$ in equation 4. Our simulation parameters are similar to the Heckman model experiments. We use 5000 datapoints; 5 features (including the intercept column of 1s); X, $\boldsymbol{\beta}_Y$, and $\boldsymbol{\beta}_\Delta$ drawn from unit normal distributions; and $\sigma \sim \mathcal{N}(2, 0.1)$. We draw the intercept terms $\boldsymbol{\beta}_{Y_0} \sim \mathcal{N}(-2, 0.1)$ and $\boldsymbol{\beta}_{\Delta_0} \sim \mathcal{N}(2, 0.1)$ to approximately match $p(Y)$ and $p(T)$ in realistic medical

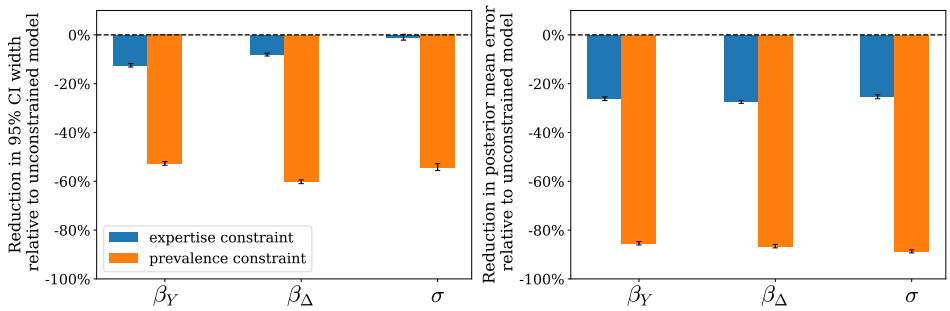


Figure S4: The prevalence and expertise constraints still improve parameter inference when quadrupling the number of features relative to Figure 2. Results are shown using synthetic data from the Bernoulli-sigmoid model with uniform unobservables. Both constraints produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

settings, where disease prevalence is relatively low, but a large fraction of the population is tested because false negatives are more costly than false positives. We assume the expertise constraint applies to $\boldsymbol{\beta}_{\Delta_2} = \boldsymbol{\beta}_{\Delta_3} = \boldsymbol{\beta}_{\Delta_4} = 0$. We show results in Figure 2. When calculating the results for $\boldsymbol{\beta}_{\boldsymbol{\Delta}}$, we do not include the dimensions along which we assume expertise since these dimensions are assumed to be fixed for the model with the expertise constraint.

D.3 Normal unobservables model

We also conduct synthetic experiments using the following Bernoulli-sigmoid model with normal unobservables:

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, \sigma^2) \\ r_i &= X_i^T \boldsymbol{\beta_Y} + Z_i \\ Y_i &\sim \text{Bernoulli}(\text{sigmoid}(r_i)) \\ T_i &\sim \text{Bernoulli}(\text{sigmoid}(\alpha r_i + X_i^T \boldsymbol{\beta_\Delta}))\,. \end{aligned} \tag{6}$$

We show results for two cases: when σ^2 is fixed and when α is fixed. Because this distribution of unobservables does not allow us to marginalize out Z, it converges more slowly than the uniform unobservables model and we must use a smaller sample size for computational tractability.

Fixed σ^2: We use the same simulation parameters as the uniform model. We fix $\sigma^2 = 2$ and we draw $\alpha \sim N(1, 0.1)$. We show results in Figure S2. Both the prevalence and expertise constraints produce more precise and accurate inferences for all parameters relative to the unconstrained model.

Fixed α: We use the same simulation parameters as the uniform model, except we reduce the number of datapoints to 200. We fix $\alpha = 1$ and we draw $\sigma^2 \sim N(2, 0.1)$. We show results in Figure S3. Both the prevalence and expertise constraints produce more precise and accurate inferences for all parameters relative to the unconstrained model.

D.4 More complex models

To show our constraints are useful with more complex models, we ran two additional synthetic experiments on the Bernoulli-sigmoid model with uniform unobservables. First, we demonstrated applicability to higher-dimensional features. We show results in Figure S4. Even after quadrupling the number of features (which increases the runtime by a factor of three), both constraints still improve precision and accuracy. Secondly, we evaluate a more complex model with pairwise nonlinear interactions between features. We show results in Figure S5. Again both constraints generally improve precision and accuracy. We note our implementation relies on MCMC which is known to be less scalable than approaches like variational inference (Wainwright & Jordan, 2008) and would likely not scale to very high-dimensional features. However, our approach does not intrinsically rely

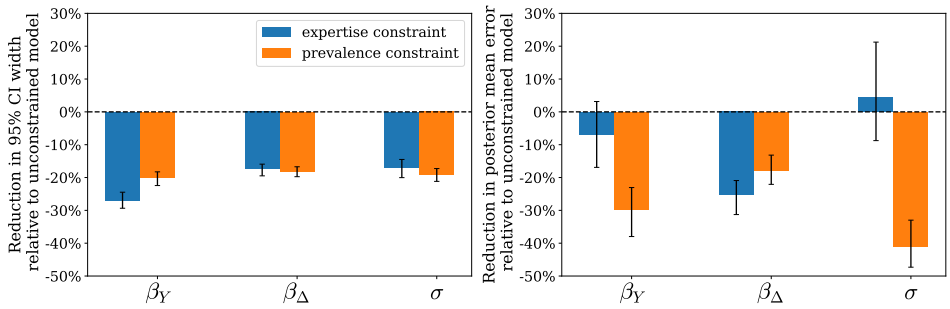


Figure S5: The prevalence and expertise constraints still improve parameter inference even when using pairwise nonlinear interactions between features (rather than only linear terms, as shown in Figure 2). Results are shown using synthetic data from the Bernoulli-sigmoid model with uniform unobservables. Both constraints generally produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

on MCMC, and incorporating more scalable estimation methods is a natural direction for future work.[6]

E UK BIOBANK DATA

Label processing: In the UK Biobank (UKBB), each person's data is collected at their baseline visit. The time period we study is the 10 years preceding each person's baseline visit. $T_i \in \{0, 1\}$ denotes whether the person receives a mammogram in the 10 year period. $Y_i \in \{0, 1\}$ denotes whether the person receives a breast cancer diagnosis in the 10 year period. We verify that very few people in the dataset have $T = 0$ and $Y = 1$ (i.e., are diagnosed with no record of a test): $p(Y = 1|T = 0) = 0.0005$. We group these people with the untested $T = 0$ population, since they did not receive a breast cancer test.

Feature processing: We include features which satisfy two desiderata. First, we use features that previous work has found to be predictive of breast cancer (NIH National Cancer Institute, 2017; Komen, 2023; Yanes et al., 2020). Second, since features are designed to be used in predicting T_i and Y_i, they must be measured prior to T_i and Y_i (i.e., at the beginning of the 10 year study period). Since the start of our 10 year study period occurs before the date of data collection, we choose features that are either largely time invariant (e.g. polygenic risk score) or that can be recalculated at different points in time (e.g. age). The full list of features that we include is: breast cancer polygenic risk score, previous biopsy procedure (based on OPCS4 operation codes), age at first period (menarche), height, Townsend deprivation index[7], race (White, Black/mixed Black, and Asian/mixed Asian), and age at the beginning of the study period (<35, 35-39, and 40-45). We normalize all features to have mean 0 and standard deviation 1.

Sample filtering: We filtered our sample based on four conditions. (i) We removed everyone without data on whether or not they received breast cancer testing, which automatically removed all men because UKBB does not have any recorded data on breast cancer tests for men. (ii) We removed everyone who was missing data (e.g. responded "do not know") for breast cancer polygenic risk score; previous biopsy procedure; menarche; height; Townsend deprivation index; race; age; duration of moderate physical activity; cooked, salad, and raw vegetable intake; weight; use of the following medication: aspirin, ibuprofen, celebrex, and naproxen; family history of breast cancer; and previous detection of carcinoma in breast. (iii) We removed everyone who did not self report being of White, Black/mixed Black, or Asian/mixed Asian race. (iv) We remove patients who were diagnosed with breast cancer before the start of our 10 year study period, as is standard in previous work (Zink et al., 2023). (v) We removed everyone above the age of 45 at the beginning of the observation period, since

[6]We use the same simulation parameters as our standard uniform model experiments. We set the expertise constraint to apply to a random subset of 60% of the features to match the standard uniform model experiments where expertise is assumed for 3 out of the 5 features.

[7]The Townsend deprivation index is a measure of material deprivation that incorporates unemployment, non-car ownership, non-home ownership, and household overcrowding (Townsend et al., 1988).

the purpose of our case study is to assess how the model performs in the presence of the distribution shift induced by the fact that young women tested for breast cancer are non-representative.[8]

Model fitting: We divide the data into train and test sets with a 70-30 split. We use the train set to fit our model. We use the test set to validate our risk predictions on the tested population ($T = 1$). We validate our risk predictions for the $T = 1$ population on a test set because the model is provided both Y and X for the train set, so using a test set replicates standard machine learning practice. We do not run the other validations (predicting risk among the $T = 0$ population and inference of unobservables) on a test set because in all these cases the target variable is unseen by the model during training. Overfitting concerns are minimal because we use a large dataset and few features.

Inferred risk predicts breast cancer diagnoses among the untested population: When verifying that inferred risk predicts future cancer diagnoses for the people who were untested ($T_i = 0$) at the baseline, we use data from the three UKBB follow-up visits. We only consider the subset of people who attended at least one of the follow-up visits. We mark a person as having a future breast cancer diagnosis if they report receiving a breast cancer diagnosis at a date after their baseline visit.

Inferred unobservables correlate with known unobservables: We verify that across people, our inferred posterior mean of unobservables correlates with a true unobservable—whether the person has a family history of breast cancer. We define a family history of breast cancer as either the person's mother or sisters having breast cancer. We do not include this data as a feature because we cannot be sure that the measurement of family history precedes the measurement of T_i and Y_i. This allows us to hold out this feature as a validation.

IRB: Our institution's IRB determined that our research did not meet the regulatory definition of human subjects research. Therefore, no IRB approval or exemption was required.

F Additional experiments on cancer data

Here we provide additional sets of experiments. We provide a comparison to various baseline models (Appendix F.1) and robustness experiments (Appendix F.2).

F.1 Comparison to baseline models

We provide comparisons to three different types of baseline models: (i) a model trained solely on the tested population, (ii) a model which assumes the untested group is negative, and (iii) other selective labels baselines.

Comparison to models trained solely on the tested population: The first baseline that we consider is a model which estimates $p(Y_i = 1|T_i = 1, X_i)$: i.e., a model which predicts outcomes without unobservables using only the tested population.[9] This is a widely used approach in medicine and other selective labels settings. In medicine, it has been used to predict COVID-19 test results among people who were tested (Jehi et al., 2020; McDonald et al., 2021); to predict hypertrophic cardiomyopathy among people who received gold-standard imaging tests (Farahani et al., 2020); and to predict discharge outcomes among people deemed ready for ICU discharge (McWilliams et al., 2019). It has also been used in the settings of policing (Lakkaraju et al., 2017), government inspections (Laufer et al.), and lending (Björkegren & Grissen, 2020).

[8]To confirm that our predictive performance remains good when looking at patients of all ages, we conduct an additional analysis fitting our model on a dataset without the age filter, but keeping the other filters. (For computational tractability, we downsample this dataset to approximately match the size of the original age-filtered dataset.) We fit this dataset using the same model as that used in our main analyses, but add features to capture the additional age categories (the full list of age categories are: <35, 35-39, 40-44, 45-49, 50-54, ≥55). We find that if anything, predictive performance when using the full cohort is better than when using only the younger cohort from our main analyses in §5.2. Specifically, the model's quintile ratio is 4.6 among the tested population ($T_i = 1$) and 7.0 among the untested population ($T_i = 0$) that attended a follow-up visit.

[9]We estimate this using a logistic regression model, which is linear in the features. To confirm that non-linear methods yield similar results, we also fit random forest and gradient boosting classifiers. These methods achieve similar predictive performance to the linear model and they also predict an implausible age trend.

As shown in Figure S6, we find that the model trained solely on the tested population learns that cancer risk first increases with age and then falls sharply, contradicting prior epidemiological and physiological evidence (Komen, 2023; Cancer Research UK; US Cancer Statistics Working Group et al., 2013; Campisi, 2013). We see this same trend for a model fit without a prevalence constraint in §5.4. This indicates that these models do not predict plausible inferences consistent with prior work.

Comparison to a model which treats the untested group as negative: We also consider a baseline model which treats the untested group as negative; this is equivalent to predicting $p(T_i = 1, Y_i = 1|X_i)$, an approach used in prior selective labels work (Shen et al., 2021; Ko et al., 2020; Rastogi et al., 2023). We find that, though this baseline no longer learns an implausible age trend, it underperforms our model in terms of AUC (AUC is 0.60 on the tested population vs. 0.63 for our model; AUC is 0.60 on the untested population vs. 0.63 for our model) and quintile ratio (quintile ratio on the tested population is 2.4 vs. 3.3 for our model; quintile ratio for both models is 2.5 on the untested population). This baseline is a special case of our model with the prevalence constraint set to $p(Y = 1|T = 0) = 0$, an implausibly low prevalence constraint. In light of this, it makes sense that this baseline learns a more plausible age trend, but underperforms our model overall.

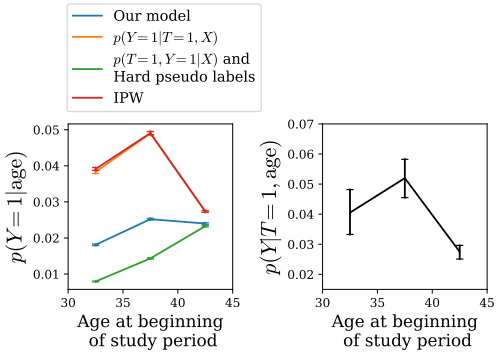


Figure S6: We run three sets of baseline models: (i) models trained solely on the tested population, estimating $p(Y_i = 1|T_i = 1, X_i)$; (ii) models which treat the untested group as negative, estimating $p(T_i = 1, Y_i = 1|X_i)$; and (iii) other selective labels baselines (IPW and hard pseudo labels). Both IPW and the model estimating $p(Y_i = 1|T_i = 1, X_i)$ learn that cancer risk first increases and then decreases with age, contradicting prior literature. This implausible inference occurs because the tested population has the same misleading age trend (right plot). In contrast, our Bayesian model learns a more plausible age trend (left plot, blue line). Hard pseudo labels and the model estimating $p(T_i = 1, Y_i = 1|X_i)$ also learn plausible age trends, but they underperform our Bayesian model in predictive performance.

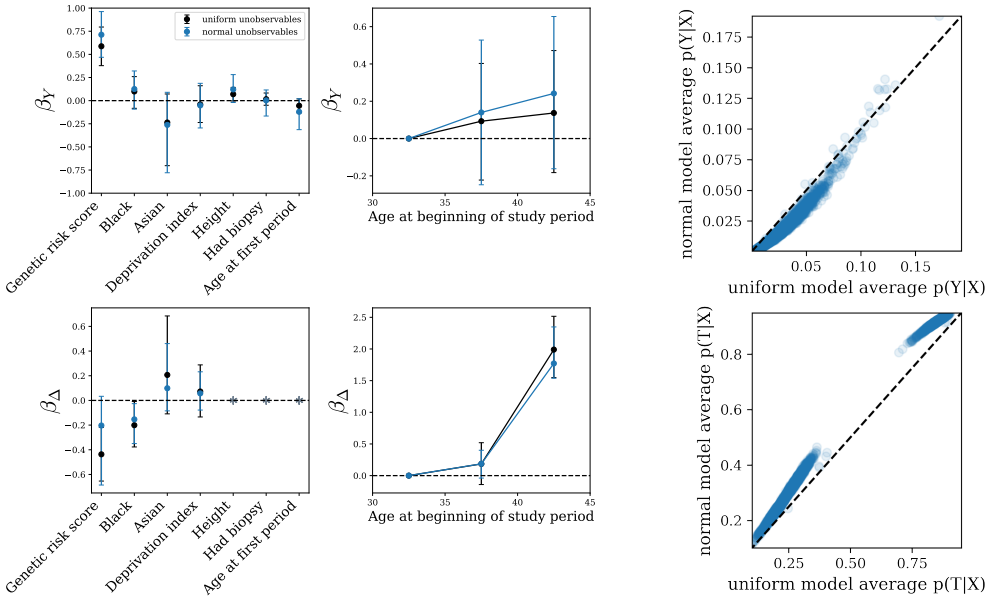
Comparison to other selective labels baselines: We also consider two other common selective labels baselines (Rastogi et al., 2023). First, we predict hard pseudo labels for the untested population (Lee, 2013): i.e., we train a classifier on the tested population and use its outputs as pseudo labels for the untested population. Due to the low prevalence of breast cancer in our dataset, the pseudo labels are all $Y_i = 0$, so this model is equivalent to treating the untested group as negative and similarly underperforms our model in predictive performance. Second, we use inverse propensity weighting (IPW) (Shimodaira, 2000): i.e., we train a classifier on the tested population but reweight each sample by the inverse propensity weight $\frac{1}{p(T_i=1|X_i)}$.[10] As shown in Figure S6, this baseline also learns the implausible age trend that cancer risk first increases and then decreases with age: this is because merely reweighting the sample, without encoding that the untested patients are less likely to have cancer via a prevalence constraint, is insufficient to correct the misleading age trend.

F.2 Robustness checks for the breast cancer case study

Our primary breast cancer results (§5) are computed using the Bernoulli-sigmoid model in equation 4. In this model, unobservables are drawn from a uniform distribution, α is set to 1, and the prevalence constraint is set to $p(Y = 1) = 0.02$ based on previously reported breast cancer incidence statistics (Cancer Research UK). In order to assess the robustness of our results, we show that they remain consistent when altering all three of these aspects to plausible alternative specifications.

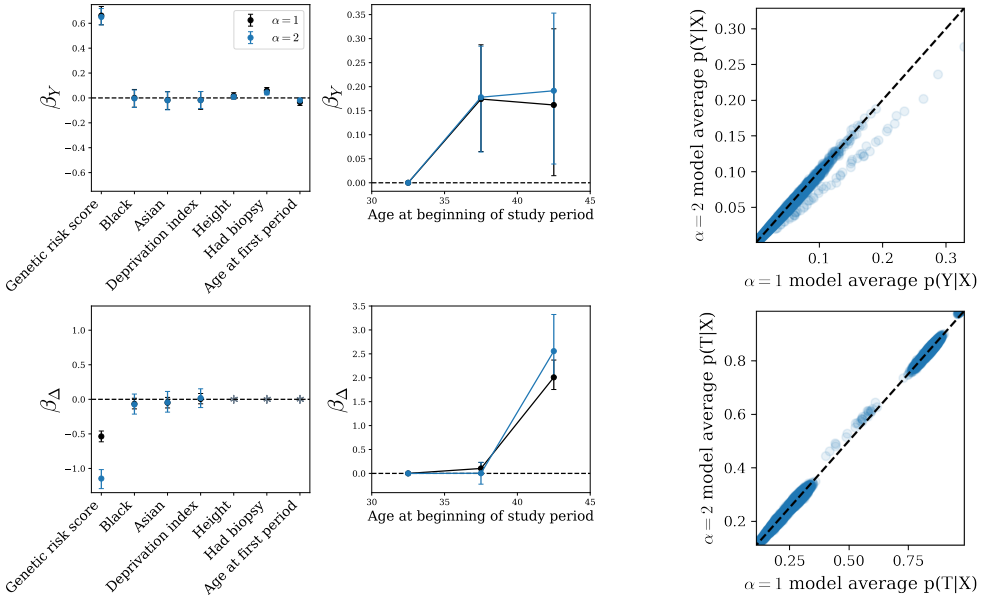
Consistency across different distributions of unobservables: We compare the uniform unobservables model (equation 4) to the normal unobservables model (equation 6). As described in Appendix

[10] We clip $p(T_i = 1|X_i)$ to be between [0.05, 0.95], consistent with previous work.



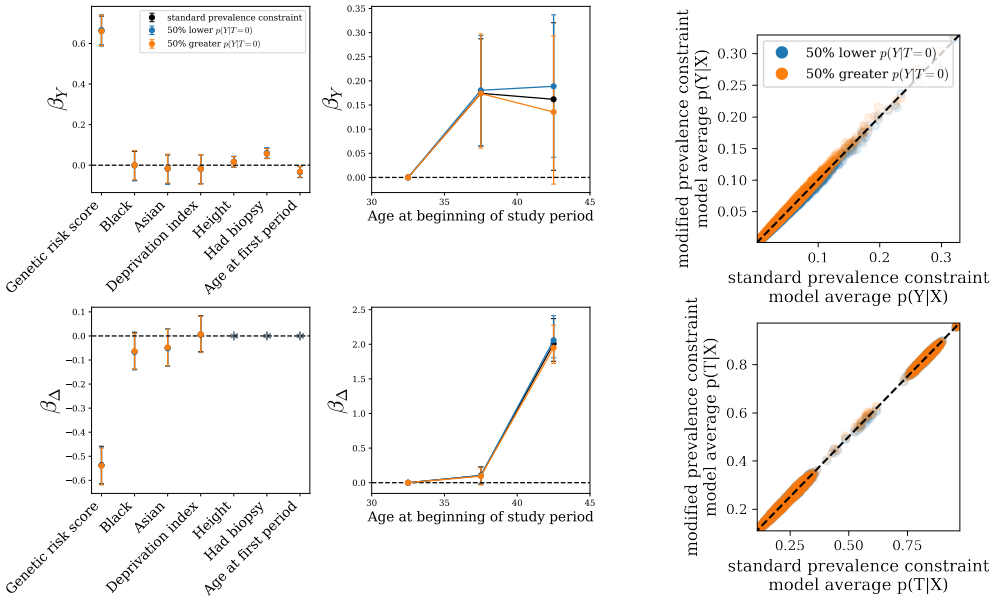
(a) (b)

Figure S7: We compare the results from the uniform unobservable model in equation 4 (black) and the normal unobservable model in equation 6 (blue). Figure S7a: The estimated $\boldsymbol{\beta_Y}$ and $\boldsymbol{\beta_\Delta}$ coefficients remain similar for both models, with similar trends in the point estimates and overlapping confidence intervals. Figure S7b: Both models predict highly correlated values for $p(Y_i|X_i)$ and $p(T_i|X_i)$. Perfect correlation is represented by the dashed line.



(a) (b)

Figure S8: We compare the results from the uniform unobservable model with $\alpha = 1$ (black) and $\alpha = 2$ (blue). Figure S8a: The inferred $\boldsymbol{\beta_Y}$ and $\boldsymbol{\beta_\Delta}$ coefficients are generally very similar, with similar trends in the point estimates and overlapping confidence intervals. The only exception is the estimate of $\boldsymbol{\beta_\Delta}$ for genetic risk, which is explained by the fact that the prediction of $\boldsymbol{\beta_\Delta}$ depends on the value of α. Figure S8b: Both models predict highly correlated values for $p(Y_i|X_i)$ and $p(T_i|X_i)$. Perfect correlation is represented by the dashed line.



(a) (b)

Figure S9: We compare the results from the uniform unobservables model with a prevalence constraint of $\mathbb{E}[Y] = 0.02$ informed by cancer statistics (Cancer Research UK) (black), a prevalence constraint which corresponds to 50% less of the untested population having the disease (blue), and a prevalence constraint which corresponds to 50% more of the untested population having the disease (orange). Figure S9a: The predictions for all three models are similar as seen by the similar trends in the point estimates and overlapping confidence intervals. Figure S9b: All three models predict correlated values for $p(Y_i|X_i)$ and $p(T_i|X_i)$. Perfect correlation is represented by the dashed line.

D, the normal unobservables model does not allow us to marginalize out Z_i and thus converges more slowly. Hence, for computational tractability, we run the model on a random subset of $\frac{1}{8}$ of the full dataset. In Figure S7a, we see that the estimated coefficients for both models remain similar, with similar trends in the point estimates and overlapping confidence intervals. Figure S7b shows that the inferred values of $p(Y_i|X_i)$ and $p(T_i|X_i)$ for each data point also remain correlated, indicating that the models infer similar testing probabilities and disease risks for each person.

Consistency across different α: We compare the uniform unobservables model with $\alpha = 1$ to a uniform unobservables model with $\alpha = 2$. In Figure S8a, we see that the inferred coefficients for both models are generally very similar, with similar trends in the point estimates and overlapping confidence intervals. The only exception is $\boldsymbol{\beta_\Delta}$ for the genetic risk score. While both models find a negative $\boldsymbol{\beta_\Delta}$ for the genetic risk score, indicating genetic information is underused, the coefficient is less negative when $\alpha = 1$. This difference occurs because altering α changes the assumed relationship between the risk score and the testing probability under purely risk-based allocation, and thus changes the estimated deviations from this relationship (which $\boldsymbol{\beta_\Delta}$ captures). Past work also makes assumptions about the relationship between risk and human decision-making (Pierson, 2020; Simoiu et al., 2017; Pierson et al., 2018; 2020). We can restrict the plausible values of α, and thus $\boldsymbol{\beta_\Delta}$, using the following approaches: (i) restricting α to a range of reasonable values based on domain knowledge; (ii) setting α to the value predicted by a model with σ^2 pinned; or (iii) fitting α and σ^2 in a model with non-binary Y_i outcomes when both parameters can be simultaneously identified.

To confirm model consistency, we compare the inferred values of $p(Y_i|X_i)$ and $p(T_i|X_i)$ for each data point. As shown in Figure S8b, these estimates remain highly correlated across both models, indicating that the models infer similar testing probabilities and disease risks for each person.

Consistency across different prevalence constraints: The prevalence constraint fixes the estimate of $p(Y = 1)$. Because the proportion of tested individuals who have the disease, $p(Y = 1|T = 1)$, is known from the observed data, fixing $p(Y = 1)$ is equivalent to fixing the proportion of *untested* individuals with the disease, $p(Y = 1|T = 0)$. For the model in §5, we set the prevalence constraint to 0.02 based on cancer incidence statistics (Cancer Research UK). However, disease prevalence may not be exactly known (Manski & Molinari, 2021; Manski, 2020; Mullahy et al., 2021). To check the

robustness of our results to plausible variations in the prevalence constraint, we compare to two other prevalence constraints that correspond to 50% lower and 50% higher values of $p(Y = 1|T = 0)$.[11] This yields overall prevalence constraints of $\mathbb{E}[Y] \approx 0.018$ and 0.022, respectively. In Figure S9a, we compare the $\boldsymbol{\beta_Y}$ and $\boldsymbol{\beta_\Delta}$ coefficients for these three different prevalence constraints. Across all three models, the estimated coefficients remain similar, with similar trends in the point estimates and overlapping confidence intervals. In particular, the age trends also remain similar in all three models, in contrast to the model fit without a prevalence constraint (§5.4). In Figure S9b, we compare the inferred values of $p(Y_i|X_i)$ and $p(T_i|X_i)$ for each data point and confirm that these estimates remain highly correlated across all three models, indicating that the models infer very similar testing probabilities and disease risks for each person.

[11]While our results are robust to significant alterations of the prevalence constraint, we do note that if the model is run with a wildly misspecified prevalence constraint — for example, $p(Y = 1|T = 0) = 0$ — it could produce incorrect results. To avoid this issue, our Bayesian framework also accommodates approximate constraints, if the prevalence is only approximately known.