Domain constraints improve risk prediction when outcome data is missing

Sidhika Balachandar
Cornell Tech

Nikhil Garg
Jacobs Technion-Cornell Institute, Cornell Tech

Emma Pierson
Jacobs Technion-Cornell Institute, Cornell Tech

Abstract

Machine learning models often predict the outcome resulting from a human decision. For example, if a doctor tests a patient for disease, will the patient test positive? A challenge is that the human decision *censors* the outcome data: 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 describe a Bayesian model to capture this setting whose purpose is to estimate risk for both tested and untested patients. To aid model estimation, we propose two *domain-specific* 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 the constraints can improve parameter inference. We apply our model to a case study of cancer risk prediction, showing that the model can identify suboptimalities in test allocation and that the prevalence constraint increases the plausibility of inferences.

1 Introduction

Machine learning models often predict outcomes in settings where a human makes a high-stakes decision. In healthcare, a doctor decides whether to test a patient for disease, and machine learning models predict whether the patient will test positive [1–3]. A fundamental challenge in all these settings is that the human decision *censors* the data the model can learn from: e.g., 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.

Overall, there is a challenging distribution shift between the tested and untested populations: they may differ 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—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 [4], occurs in high-stakes settings including medical testing, hiring, and lending and has been the subject of wide academic interest (related work in Appendix A).

Without further constraints, there are a wide range of possibilities for the untested patients. However, we often can limit these possibilities: e.g., in medicine, information about overall disease prevalence constrains the proportion of untested patients who can have the disease. Motivated by this, we make the following contributions. First, we describe a Bayesian model which captures this setting and nests as special cases classic models from econometrics. Second, we propose two *domain constraints*

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

informed by the medical domain: a *prevalence constraint* and an *expertise constraint*. We show theoretically and on synthetic data that the constraints improve inference. Finally, we apply our model to estimate breast cancer risk. We show the model can identify suboptimalities in test allocation and that the prevalence constraint increases the plausibility of inferences. While our feature vector is low-dimensional, our approach extends to supervised learning tasks with more complex inputs: e.g., medical images or embeddings from foundation models [5–7].

2 Model

We now describe our Bayesian model. Following previous work [3], 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 (i) to accurately estimate risk for both the tested and untested patients and (ii) to quantify deviations from purely risk-based test allocation.

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 further 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 with a continuous Y_i. If $T_i = 0$ we do not observe Y_i.

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, $Z_i \in \mathbb{R}$ represents *unobservables* [8, 9], that affect *both* T_i and Y_i, but are not in the dataset – e.g., whether the doctor observes that the person is in pain. Z_i is drawn from a distribution f with scale parameter σ^2, capturing the importance of unobservables.

$r_i \in \mathbb{R}$ represents a person's *risk score*, which captures their risk of having a disease. r_i 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., $Y_i \sim \text{Bernoulli}(\text{sigmoid}(r_i))$.

Whether a person is tested ($T_i = 1$) is drawn from a distribution h_T parameterized by $\alpha r_i + X_i^T \boldsymbol{\beta_\Delta}$. This function captures that testing decisions depend not only on r_i but also on human and policy factors: for example, screening policies or socioeconomic disparities (captured by $\boldsymbol{\beta_\Delta} \in \mathbb{R}^D$). Putting things together, the model parameters are $\theta \triangleq (\alpha, \sigma^2, \boldsymbol{\beta_\Delta}, \boldsymbol{\beta_Y})$.

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

Prevalence constraint: We assume disease prevalence across the entire population (not just the tested population), $\mathbb{E}[Y]$, is known, as is true in many health settings: for example, cancer [10], COVID-19 [11], and heart disease [12]. In some cases, the prevalence is only *approximately* known [13–15]; 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. However, we note that we still estimate ${\boldsymbol{\beta_Y}}_d$ for the features on which we assume expertise.

3 Theoretical Analysis

We summarize our proofs here and provide details in Appendix B. We prove why our proposed constraints improve parameter inference by analyzing a special case of our general model in eq. (1). We show that this case is equivalent to the Heckman model [16, 17], used to correct bias from non-randomly selected samples (Proposition B.1). It is known that placing constraints on the Heckman model *improves the precision of parameter inference* [18], suggesting that our proposed constraints can do so as well. We show that our constraints *never worsen* the precision of parameter inference (Proposition B.2) and provide conditions under which they strictly *improve* it (Proposition B.4).

3.1 Empirical extension beyond the Heckman special case

In our experiments, we validate that our theoretical results hold beyond the Heckman setting. Specifically, we conduct experiments using the following Bernoulli-sigmoid model:

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

We draw Z from a uniform distribution and fix α because this allows us to marginalize out Z, accelerating model-fitting (see Appendix C). However, our approach is applicable to other distributions of unobservables: in Appendix D.3 and Appendix E.3 we show similar results for a normal distribution.

3.2 Synthetic experiments

In Appendix D, we validate our approach on synthetic data. We find that our theory agrees with our experimental results on both the Heckman and Bernoulli-sigmoid model. The constraints produce narrower posterior confidence intervals (improving precision). The constraints also produce posterior means which lie closer to the true parameter values (improving accuracy). The code for these experiments is here: `https://github.com/sidhikabalachandar/domain_constraints`.

4 Real-world case study: Breast cancer testing

In the following sections, we describe our experimental set up and the model we fit (§4.1), we conduct four validations on the fitted model (§4.2), we use the model to assess historical testing decisions (§4.3), and we compare to a model fit without a prevalence constraint (§4.4).

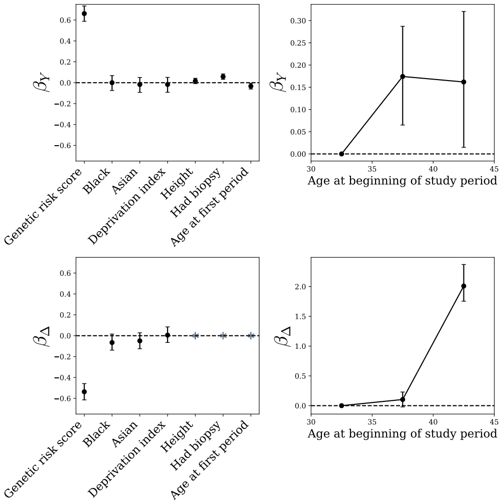


Figure 1: Estimated $\boldsymbol{\beta_Y}$ (top) capture known cancer risk factors. Estimated $\boldsymbol{\beta_\Delta}$ (bottom) capture the underuse of genetic information (left) and known age-based testing policies (right). Points indicate posterior means and vertical lines indicate 95% confidence intervals. Gray asterisks indicate coefficients set by the expertise constraint.

4.1 Experimental setup

We apply our model to a breast cancer dataset of 54,746 people from the UK Biobank [19] (see Appendix F for details). Our X_i consist of 7 features predictive of breast cancer; $T_i \in \{0, 1\}$ denotes whether the person receives a mammogram (the most common breast cancer test) in the 10 years following the measurement of features; and $Y_i \in \{0, 1\}$ denotes whether the person is diagnosed with breast cancer in the 10 year period. We analyze a younger female population (age ≤ 45) because it creates a challenging distribution shift: younger people are generally not tested for cancer [20], so the tested and untested populations differ. We include a prevalence constraint $\mathbb{E}[Y] = 0.02$, based on incidence statistics in the UK [10]. We include an expertise constraint by allowing $\boldsymbol{\beta_\Delta}$ to deviate from 0 only for (i) racial/socioeconomic features, due to disparities in healthcare access [21–23]; (ii) genetic features, since genetic information may be underused [24]; and (iii) age, due to age-based testing policies [20]. In Appendix E.3, we run robustness experiments.

In Figure 1, 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 E.2 Figure S8, we also compare our model's performance to a suite of additional baselines. This includes (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. Collectively, these baselines all suffer from various issues our model does not, including learning implausible age trends inconsistent with prior literature or worsening predictive performance.

4.2 Validating the model

Validating models in selective labels settings is difficult because outcomes are not observed for the untested. Still, we conduct a suite of validations. Below we show the first validation: the model's inferred risks predict cancer diagnoses. In Appendix E.1 we present three more validations. First, we show that the inferred unobservables correlate with a true unobservable—family history of breast cancer. This is an unobservable because it influences both T and Y but is not included in the data given to the model. Second, we show the estimated $\boldsymbol{\beta_Y}$ coefficients capture known cancer risk factors: genetic risk, previous biopsy, age at first period, and age [25, 26]. Third, we show the inferred age-based testing policy correlates with known public health policies.

Inferred risk predicts breast cancer diagnoses: Verifying that inferred risk predicts cancer 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[1] 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 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 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%).[2]

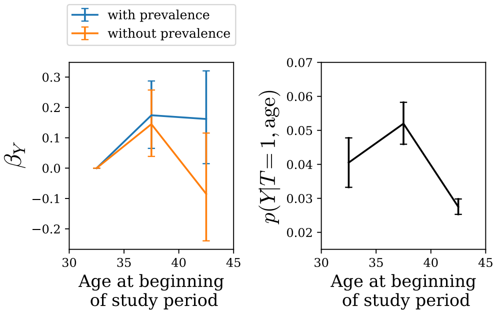


Figure 2: Without the prevalence constraint, the model learns that cancer risk first increases and then decreases with age (left, orange), contradicting prior literature. This incorrect inference occurs because the tested population has the same misleading age trend (right). In contrast, the prevalence constraint encodes that the (younger) untested population has lower risk, allowing the model to learn a more accurate age trend (left, blue).

4.3 Assessing historical testing decisions

Non-zero components of $\boldsymbol{\beta_\Delta}$ indicate features that affect a person's probability of being tested even when controlling for their disease risk. The bottom left plot in Figure 1 plots the inferred $\boldsymbol{\beta_\Delta}$, revealing that genetic information is underused. While genetic risk is strongly predictive of Y,

[1] Reporting outcome rates by inferred risk quintile or decile is a common metric in health risk prediction settings [3, 27, 28].

[2] We also note that the AUC amongst the tested population is 0.63 and amongst the untested population that attended a followup visit is 0.63. These AUCs are similar to past predictions which use similar feature sets[29]. For instance, the Tyrer-Cuzick [30] and Gail [31] models achieved AUCs of 0.62 and 0.59.

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 [32]. However, both confidence intervals overlap zero (due to the small size of these groups in our dataset).

4.4 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 2, without the prevalence constraint, the model learns that cancer risk first increases with age and then falls, contradicting prior epidemiological and physiological evidence [10, 33–35]. This is because, due to the age-based testing policy in the UK [20], being tested for breast cancer before age 50 is unusual, so patients under the age of 50 are tested only if they are of very high risk for breast cancer. Therefore, in our setting the tested population under age 50 is non-representative because their risk is much higher than the corresponding untested population. Thus, 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.

5 Discussion

We describe a Bayesian model to infer risk and assess historical human decision-making in selective labels settings, which commonly occur in healthcare and other domains. Such models are challenging to estimate because the untested population may differ from the tested population. To overcome this, we propose two domain constraints—a prevalence constraint and an expertise constraint—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.

References

[1] L. Jehi, X. Ji, A. Milinovich, S. Erzurum, B. P. Rubin, S. Gordon, J. B. Young, and M. W. Kattan, “Individualizing risk prediction for positive coronavirus disease 2019 testing: Results from 11,672 patients,” *Chest*, vol. 158, no. 4, pp. 1364–1375, 2020.

[2] S. A. McDonald, R. J. Medford, M. A. Basit, D. B. Diercks, and D. M. Courtney, “Derivation with internal validation of a multivariable predictive model to predict COVID-19 test results in emergency department patients,” *Academic Emergency Medicine*, vol. 28, no. 2, pp. 206–214, 2021.

[3] S. Mullainathan and Z. Obermeyer, “Diagnosing physician error: A machine learning approach to low-value health care,” *The Quarterly Journal of Economics*, vol. 137, no. 2, pp. 679–727, 2022.

[4] H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, and S. 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.

[5] C. Pang, X. Jiang, K. S. Kalluri, M. Spotnitz, R. Chen, A. Perotte, and K. Natarajan, “CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks,” in *Machine Learning for Health*, pp. 239–260, PMLR, 2021.

[6] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, “Doctor AI: Predicting clinical events via recurrent neural networks,” in *Machine Learning for Health*, pp. 301–318, PMLR, 2016.

[7] P. Prakash, S. Chilukuri, N. Ranade, and S. Viswanathan, “RareBERT: Transformer architecture for rare disease patient identification using administrative claims,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 453–460, 2021.

[8] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

[9] A. Rambachan, A. Coston, and E. H. Kennedy, "Counterfactual risk assessments under unmeasured confounding," 2022.

[10] Cancer Research UK, "Breast cancer statistics." https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer.

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

[12] CDC, "Prevalence of heart disease–United States, 2005," *MMWR. Morbidity and Mortality Weekly Report*, vol. 56, no. 6, pp. 113–118, 2007.

[13] C. F. Manski and F. Molinari, "Estimating the COVID-19 infection rate: Anatomy of an inference problem," *Journal of Econometrics*, vol. 220, no. 1, pp. 181–192, 2021.

[14] C. F. Manski, "Bounding the accuracy of diagnostic tests, with application to COVID-19 antibody tests," *Epidemiology*, vol. 32, no. 2, pp. 162–167, 2020.

[15] J. Mullahy, A. Venkataramani, D. L. Millimet, and C. F. Manski, "Embracing uncertainty: The value of partial identification in public health and clinical research," *American Journal of Preventive Medicine*, vol. 61, no. 2, 2021.

[16] J. 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*, vol. 5, pp. 475–492, National Bureau of Economic Research, 1976.

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

[18] A. Lewbel, "The identification zoo: Meanings of identification in econometrics," *Journal of Economic Literature*, vol. 57, no. 4, pp. 835–903, 2019.

[19] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. 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*, vol. 12, no. 3, 2015.

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

[21] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical machine learning in healthcare," *Annual Review of Biomedical Data Science*, vol. 4, pp. 123–144, 2021.

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

[23] D. Shanmugam and E. Pierson, "Quantifying inequality in underreported medical conditions," *arXiv preprint arXiv:2110.04133*, 2021.

[24] S. Samphao, A. J. Wheeler, E. Rafferty, J. S. Michaelson, M. C. Specht, M. A. Gadd, K. S. Hughes, and B. 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*, vol. 198, no. 4, pp. 538–543, 2009.

[25] T. Yanes, M.-A. Young, B. Meiser, and P. A. James, "Clinical applications of polygenic breast cancer risk: A critical review and perspectives of an emerging field.," *Breast Cancer Research*, vol. 22, no. 21, 2020.

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

[27] L. Einav, A. Finkelstein, S. Mullainathan, and Z. Obermeyer, “Predictive modeling of US health care spending in late life,” *Science*, vol. 360, no. 6396, pp. 1462–1465, 2018.

[28] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[29] A. Yala, P. G. Mikhael, F. Strand, G. Lin, K. Smith, Y.-L. Wan, L. Lamb, K. Hughes, C. Lehman, and R. Barzilay, “Toward robust mammography-based models for breast cancer risk,” *Science Translational Medicine*, vol. 13, no. 578, 2021.

[30] J. Tyrer, S. W. Duffy, and J. Cuzick, “A breast cancer prediction model incorporating familial and personal risk factors,” *Statistics in Medicine*, vol. 23, no. 7, pp. 1111–1130, 2004.

[31] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill, “Projecting individualized probabilities of developing breast cancer for white females who are being examined annually,” *JNCI: Journal of the National Cancer Institute*, vol. 81, no. 24, pp. 1879–1886, 1989.

[32] G. Makurumidze, C. Lu, and B. Kemi, “Addressing disparities in breast cancer screening: A review.,” *Applied Radiology*, vol. 51, no. 6, pp. 24–28, 2022.

[33] S. G. Komen, “Factors linked to breast cancer risk.” https://www.komen.org/breast-cancer/risk-factor/factors-that-affect-risk/, 2023.

[34] 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.

[35] J. Campisi, “Aging, cellular senescence, and cancer,” *Annual Review of Physiology*, vol. 75, pp. 685–705, 2013.

[36] J. Jung, C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein, “Simple rules to guide expert classifications,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 183, no. 3, pp. 771–800, 2020.

[37] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, “Human decisions and machine predictions,” *The Quarterly Journal of Economics*, vol. 133, no. 1, pp. 237–293, 2018.

[38] D. Björkegren and D. Grissen, “Behavior revealed in mobile phone usage predicts credit repayment,” *The World Bank Economic Review*, vol. 34, no. 3, pp. 618–634, 2020.

[39] J. Crook and J. Banasik, “Does reject inference really improve the performance of application scoring models?,” *Journal of Banking & Finance*, vol. 28, no. 4, pp. 857–874, 2004.

[40] J. Jung, S. Corbett-Davies, R. Shroff, and S. Goel, “Omitted and included variable bias in tests for disparate impact,” *arXiv preprint arXiv:1809.05651*, 2018.

[41] B. Laufer, E. Pierson, and N. Garg, “End-to-end auditing of decision pipelines,”

[42] C. J. McWilliams, D. J. Lawson, R. Santos-Rodriguez, I. D. Gilchrist, A. Champneys, T. H. Gould, M. J. Thomas, and C. 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*, vol. 9, no. 3, 2019.

[43] W. S. Hong, A. D. Haimovich, and R. A. Taylor, “Predicting hospital admission at emergency department triage using machine learning,” *PLOS One*, vol. 13, no. 7, 2018.

[44] C. A. Parker, N. Liu, S. X. Wu, Y. Shen, S. S. W. Lam, and M. E. H. Ong, “Predicting hospital admission at the emergency department triage: A novel prediction model,” *The American Journal of Emergency Medicine*, vol. 37, no. 8, pp. 1498–1504, 2019.

[45] Y. Sun, B. H. Heng, S. Y. Tay, and E. Seow, "Predicting hospital admissions at emergency department triage using routine administrative data," *Academic Emergency Medicine*, vol. 18, no. 8, pp. 844–850, 2011.

[46] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani, "Risk prediction models for hospital readmission: A systematic review," *JAMA*, vol. 306, no. 15, pp. 1688–1698, 2011.

[47] A. Waters and R. Miikkulainen, "GRADE: Machine learning support for graduate admissions," *AI Magazine*, vol. 35, no. 1, pp. 64–64, 2014.

[48] M. Bogen, "All the ways hiring algorithms can introduce bias," *Harvard Business Review*, vol. 6, 2019.

[49] G. Jawaheer, M. Szomszor, and P. 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.

[50] M. Wu, M. Ghassemi, M. Feng, L. A. Celi, P. Szolovits, and F. Doshi-Velez, "Understanding vasopressor intervention and weaning: Risk prediction in a public heterogeneous clinical time series database," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 488–495, 2017.

[51] A. Coston, A. Mishler, E. H. Kennedy, and A. Chouldechova, "Counterfactual risk assessments, evaluation, and fairness," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 582–593, 2020.

[52] M. De-Arteaga, V. Jeanselme, A. Dubrawski, and A. Chouldechova, "Leveraging expert consistency to improve algorithmic decision support," *arXiv preprint arXiv:2101.09648*, 2021.

[53] E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, *et al.*, "A large-scale analysis of racial disparities in police stops across the United States," *Nature Human Behaviour*, vol. 4, no. 7, pp. 736–745, 2020.

[54] C. Simoiu, S. Corbett-Davies, and S. Goel, "The problem of infra-marginality in outcome tests for discrimination," *The Annals of Applied Statistics*, vol. 11, no. 3, pp. 1193–1216, 2017.

[55] P. Henderson, B. Chugg, B. Anderson, K. Altenburger, A. Turk, J. Guyton, J. Goldin, and D. E. Ho, "Integrating reward maximization and population estimation: Sequential decision-making for internal revenue service audit selection," *arXiv preprint arXiv:2204.11910*, 2022.

[56] S. Gholami, L. Xu, S. Mc Carthy, B. Dilkina, A. Plumptre, M. Tambe, R. Singh, M. Nsubuga, J. Mabonga, M. Driciru, *et al.*, "Stay ahead of poachers: Illegal wildlife poaching prediction and patrol planning under uncertainty with field test evaluations," 2019.

[57] N. Z. Farahani, D. S. B. Sundaram, M. Enayati, S. P. Arunachalam, K. Pasupathy, and A. 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.

[58] Z. Liu and N. 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, p. 1016–1017, Association for Computing Machinery, 2022.

[59] W. Cai, J. Gaebler, N. Garg, and S. Goel, "Fair allocation through selective information acquisition," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 22–28, 2020.

[60] N. M. Daysal, S. Mullainathan, Z. Obermeyer, S. K. Sarkar, and M. Trandafir, "An economic approach to machine learning in health policy," *Univ. of Copenhagen Dept. of Economics Discussion, CEBI Working Paper*, no. 24, 2022.

[61] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, *et al.*, "WILDS: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*, pp. 5637–5664, PMLR, 2021.

[62] S. Sagawa, P. W. Koh, T. Lee, I. Gao, S. M. Xie, K. Shen, A. Kumar, W. Hu, M. Yasunaga, H. Marklund, S. Beery, E. David, I. Stavness, W. Guo, J. Leskovec, K. Saenko, T. Hashimoto, S. Levine, C. Finn, and P. Liang, "Extending the WILDS benchmark for unsupervised adaptation," in *International Conference on Learning Representations*, 2022.

[63] J. N. Kaur, E. Kiciman, and A. Sharma, "Modeling the data-generating process is necessary for out-of-distribution generalization," in *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

[64] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[65] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[66] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[67] R. Sahoo, L. Lei, and S. Wager, "Learning from a biased sample," *CoRR*, vol. abs/2209.01754, 2022.

[68] P. Hull, "What marginal outcome tests can tell us about racially biased decision-making," tech. rep., National Bureau of Economic Research, 2021.

[69] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[70] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *International Conference on Machine Learning*, pp. 3076–3085, PMLR, 2017.

[71] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[72] A. Alaa and M. Schaar, "Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design," in *International Conference on Machine Learning*, pp. 129–138, PMLR, 2018.

[73] A. Ilyas, E. Zampetakis, and C. 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.

[74] C. Daskalakis, P. Stefanou, R. Yao, and E. Zampetakis, "Efficient truncated linear regression with unknown noise variance," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1952–1963, 2021.

[75] A. Mishler and E. H. Kennedy, "FADE: Fair double ensemble learning for observable and counterfactual outcomes," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, p. 1053, Association for Computing Machinery, 2022.

[76] J. Jung, R. Shroff, A. Feller, and S. Goel, "Bayesian sensitivity analysis for offline policy evaluation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, p. 64–70, Association for Computing Machinery, 2020.

[77] R. H. Groenwold, A. R. T. Donders, K. C. Roes, F. E. Harrell Jr, and K. G. Moons, "Dealing with missing outcome data in randomized trials and observational studies," *American Journal of Epidemiology*, vol. 175, no. 3, pp. 210–217, 2012.

[78] N. J. Perkins, S. R. Cole, O. Harel, E. J. Tchetgen Tchetgen, B. Sun, E. M. Mitchell, and E. F. Schisterman, "Principled approaches to missing data in epidemiologic studies," *American Journal of Epidemiology*, vol. 187, no. 3, pp. 568–575, 2018.

[79] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC Press, 2013.

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

[81] K. Lum, D. B. Dunson, and J. 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*, vol. 185, no. 2, pp. 588–614, 2022.

[82] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *International Conference on Learning Representations*, 2021.

[83] O. Wiles, S. Gowal, F. Stimberg, S.-A. Rebuffi, I. Ktena, K. Dvijotham, and A. T. Cemgil, "A fine-grained analysis on distribution shift," in *International Conference on Learning Representations*, 2022.

[84] I. Gao, S. Sagawa, P. W. Koh, T. Hashimoto, and P. Liang, "Out-of-domain robustness via targeted augmentations,"

[85] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. Van Der Laak, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, vol. 58, p. 101544, 2019.

[86] S. Cortes-Gomez, M. Dulce, C. Patino, and B. Wilder, "Statistical inference under constrained selection bias," *arXiv preprint arXiv:2306.03302*, 2023.

[87] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, 2019.

[88] A. Yala, P. G. Mikhael, C. Lehman, G. Lin, F. Strand, Y.-L. Wan, K. Hughes, S. Satuluru, T. Kim, I. Banerjee, *et al.*, "Optimizing risk-based breast cancer screening policies with reinforcement learning," *Nature Medicine*, vol. 28, no. 1, pp. 136–143, 2022.

[89] Y. Shen, F. E. Shamout, J. R. Oliver, J. Witowski, K. Kannan, J. Park, N. Wu, C. Huddleston, S. Wolfson, A. Millet, *et al.*, "Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams," *Nature Communications*, vol. 12, no. 1, pp. 1–13, 2021.

[90] R. Hicks, "The Heckman sample selection model." https://econ.pages.code.wm.edu/407/notes/docs/index.html, 2021.

[91] S. Jackman, *Bayesian analysis for the social sciences*. John Wiley & Sons, 2009.

[92] M. P. McLaughlin, *A compendium of common probability distributions*. Michael P. McLaughlin, 2001.

[93] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software*, vol. 76, no. 1, 2017.

[94] M. Betancourt, "A conceptual introduction to Hamiltonian Monte Carlo," *arXiv preprint arXiv:1701.02434*, 2017.

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

[96] W. P. Van de Ven and B. M. Van Praag, "The demand for deductibles in private health insurance: A probit model with sample selection," *Journal of Econometrics*, vol. 17, no. 2, pp. 229–252, 1981.

[97] O. Toomet and A. Henningsen, “Sample selection models in R: Package sampleselection,” *Journal of Statistical Software*, vol. 27, pp. 1–23, 2008.

[98] M. J. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference,” *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[99] W.-Y. Ko, K. C. Siontis, Z. I. Attia, R. E. Carter, S. Kapa, S. R. Ommen, S. J. Demuth, M. J. Ackerman, B. J. Gersh, A. M. Arruda-Olson, *et al.*, “Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram,” *Journal of the American College of Cardiology*, vol. 75, no. 7, pp. 722–733, 2020.

[100] R. Rastogi, M. Meister, Z. Obermeyer, J. Kleinberg, P. W. Koh, and E. Pierson, “Learn from the patient, not the doctor: Predicting downstream outcomes versus specialist labels,” *Working paper*, 2023.

[101] D.-H. Lee, “Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks,” 2013.

[102] E. Pierson, S. Corbett-Davies, and S. Goel, “Fast threshold tests for detecting discrimination,” in *International Conference on Artificial Intelligence and Statistics*, pp. 96–105, PMLR, 2018.

[103] P. Townsend, P. Phillimore, and A. Beattie, *Health and Deprivation: Inequality and the North*. Routledge, 1988.

[104] A. Zink, Z. Obermeyer, and E. Pierson, “Race corrections in clinical models: Examining family history and cancer risk,” *medRxiv*, pp. 2023–03, 2023.

A Related work

Selective labels problems occur in many high-stakes domains, including hiring, insurance, government inspections, tax auditing, recommender systems, lending, healthcare, education, welfare services, wildlife protection, and criminal justice [1–4, 22, 36–60]. As such, there are related literatures in machine learning and causal inference [4, 37, 51, 52, 61–67], econometrics [3, 9, 16, 68–72], statistics and Bayesian models [73–76], and epidemiology [77, 78]. 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 [79–81], which model observations as a function of both observed features X and unobserved "random effects" Z. 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 on distribution shift shows that generic methods often fail to perform well across all types of distribution shifts [61–63, 82, 83] and that incorporating domain information can improve performance. For example, [84] proposes *targeted augmentations*, which augment the data by randomizing known spurious features while preserving robust ones. [85] presents an example of this strategy in the context of histopathology slide analysis. [63] shows that modeling the data generating process is necessary for generalizing across distribution shifts. [86] proposes a framework for selection bias that can place high-probability bounds on values from the target distribution using constraints in the form of functions whose expectations are known under the target distribution. Motivated by this line of work, 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 [29, 60, 87–89]. 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 [29, 87, 88].

B Proofs

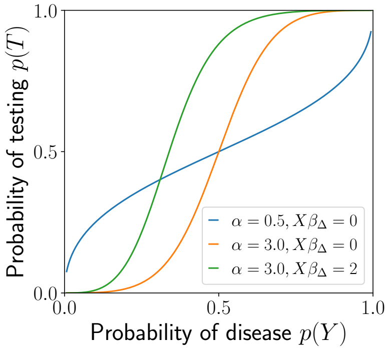


Figure S1: Effect of α and $X\boldsymbol{\beta}_{\boldsymbol{\Delta}}$: α controls how steeply testing probability $p(T)$ increases in disease risk $p(Y)$, while $X\boldsymbol{\beta}_{\boldsymbol{\Delta}}$ captures factors which affect $p(T)$ when controlling for $p(Y)$.

We start by restating the general model:

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

Unobservable Z_i is drawn from a distribution f with scale 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 $\boldsymbol{\beta}_{\boldsymbol{Y}} \in \mathbb{R}^D$) 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 $\boldsymbol{\beta}_{\boldsymbol{\Delta}} \in \mathbb{R}^D$. Because T_i depends on r_i, and r_i is a function of Z_i, T_i depends on Z_i. Figure S1 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 their disease risk. In other words, $\boldsymbol{\beta_\Delta}$ captures deviations from purely risk-based test allocation. Putting things together, the model parameters are $\theta \triangleq (\alpha, \sigma^2, \boldsymbol{\beta_\Delta}, \boldsymbol{\beta_Y})$.

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 [16, 17] is a special case of the general model in eq. (1) (Proposition B.1). It is known that placing constraints on the Heckman model can improve parameter inference [18]. 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 B.2). We then provide conditions under which our constraints strictly improve precision (Proposition B.4).

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^TX 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 [90]:*

$$\begin{aligned} T_i &= \mathbb{1}[X_i^T\tilde{\boldsymbol{\beta}}_{\boldsymbol{T}} + u_i > 0] \\ Y_i &= X_i^T\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}} + Z_i \\ \begin{bmatrix} u_i \\ Z_i \end{bmatrix} &\sim \mathit{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \tilde{\rho} \\ \tilde{\rho} & \tilde{\sigma}^2 \end{bmatrix} \right). \end{aligned} \tag{3}$$

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 eq. (1). We now prove Proposition B.1.

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

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

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 eq. (4) can be rewritten in the form in eq. (3). We first rewrite eq. (4) 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 eq. (3). 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 eq. (4) is equivalent to the Heckman model, where the mapping between the parameters is:

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

□

As described in [18], the Heckman correction model is identified without any further assumptions. It then follows that the special case of our model in eq. (4) is identified without further constraints. One can simply estimate the Heckman model, which by the mapping in eq. (5) immediately yields estimates of $\boldsymbol{\beta}_{\boldsymbol{Y}}$ 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 $\boldsymbol{\beta}_{\boldsymbol{T}}$ (and thus $\boldsymbol{\beta}_{\boldsymbol{\Delta}}$).

While the Heckman model is identified without further constraints, this identification is known to be very weak, relying on functional form assumptions [18]. 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 B.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 parameter 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 B.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:

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

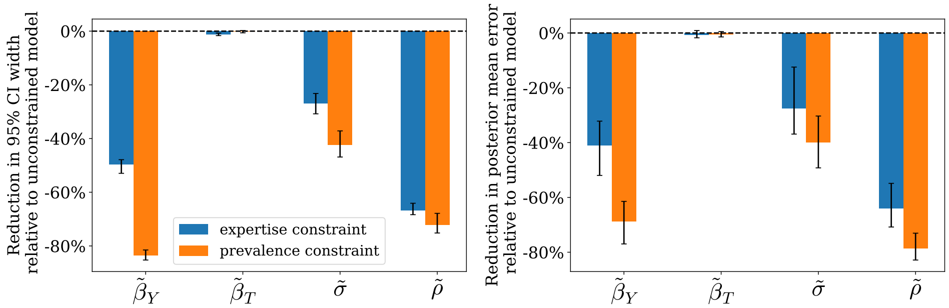


Figure S2: Results using synthetic data from the Heckman model. The prevalence and expertise constraints each produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

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

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

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

We now discuss how Proposition B.2 applies to our proposed constraints and the Heckman model. Both the prevalence and expertise constraints fix the value of at least one parameter. For the Heckman model, the prevalence constraint fixes the value of the intercept $\boldsymbol{\beta_Y}_0$ (assuming the standard condition that columns of X are zero-mean except for an intercept column of ones). The expertise constraint fixes the value of $\boldsymbol{\beta_\Delta}_d$ for some d. Thus by Proposition B.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 [16, 73]. 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 [91]). In the results below, we analyze the conditional mean of Y conditioned on $T = 1$. Thus, we start by defining this value.

Lemma B.3 (Conditional mean of Y conditioned on $T = 1$)**.** *Past work has shown that the expected value of Y_i when $T_i = 1$ is [90]:*

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

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.4. *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.*

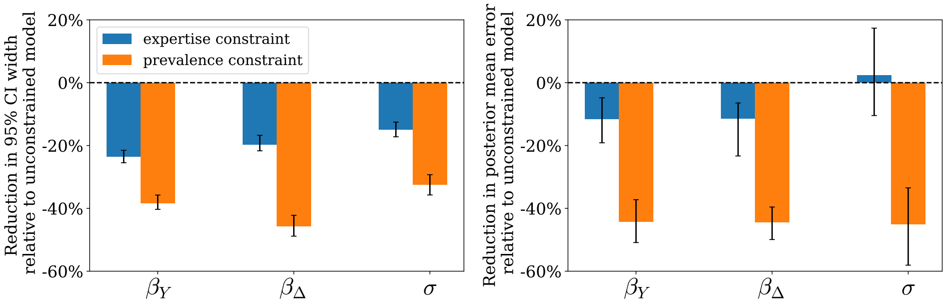


Figure S3: Results using synthetic data from the Bernoulli-sigmoid model with uniform unobservables. The prevalence and expertise constraints each produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95%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

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.3), we get

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

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 [91]:

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

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

When $\tilde{\boldsymbol{\beta}}_{\boldsymbol{Y}_d}$ is constrained to $\tilde{\boldsymbol{\beta}}^*_{\boldsymbol{Y}_d}$ 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 B.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 eq. (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 (Appendix D) and our real-world case study (§4) we validate that our constraints improve parameter inference beyond the special Heckman case.

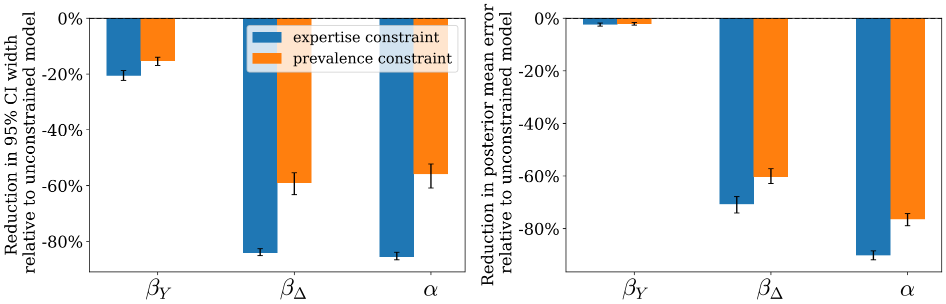


Figure S4: Results using synthetic data from the Bernoulli-sigmoid model with normal unobservables and fixed σ^2. The prevalence and expertise constraints each produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

C Derivation of the closed-form uniform unobservables model

Conducting sampling for our general model described by eq. (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 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.[3] However, we *can* derive closed forms for the special *uniform unobservables* case described by eq. (2).

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:

$$\begin{aligned}
p(Y=1,T=1|X) &= \int_Z p(Y=1,T=1|X,Z)f(Z)dZ \\
p(Y=0,T=1|X) &= \int_Z p(Y=0,T=1|X,Z)f(Z)dZ \\
p(T=0|X) &= \int_Z p(T=0|X,Z)f(Z)dZ\,,
\end{aligned}$$

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 $B = e^{X^T\boldsymbol{\beta_Y}}$:

[3] Specifically, we search over the distributions in [92], combined with logit or probit links, and find that most combinations do not yield closed forms.

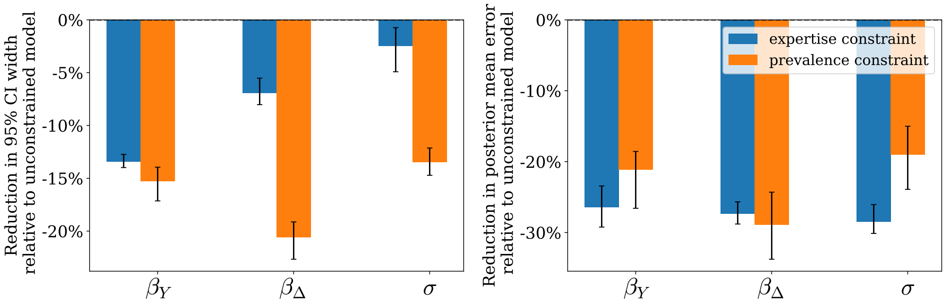


Figure S5: Results using synthetic data from the Bernoulli-sigmoid model with normal unobservables and fixed α. The prevalence and expertise constraints each produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

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

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 Figure S10).

D Synthetic experiments

We validate our 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).

For all experiments, we use the Bayesian inference package Stan [93], which uses the Hamiltonian Monte Carlo algorithm [94]. We first validate that the prevalence and expertise constraints improve the precision and accuracy of parameter inference for the Heckman model described in eq. (3). We then extend beyond this special case and examine various Bernoulli-sigmoid instantiations of our general model in eq. (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 [95–97]. We assess whether our proposed constraints improve model estimation even in this more challenging case. Specifically, we extend beyond the Heckman model to three different data generating settings: (i) uniform unobservables and fixed α, (ii) normal unobservables and fixed σ^2; and (iii) normal unobservables and fixed α. 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.

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).

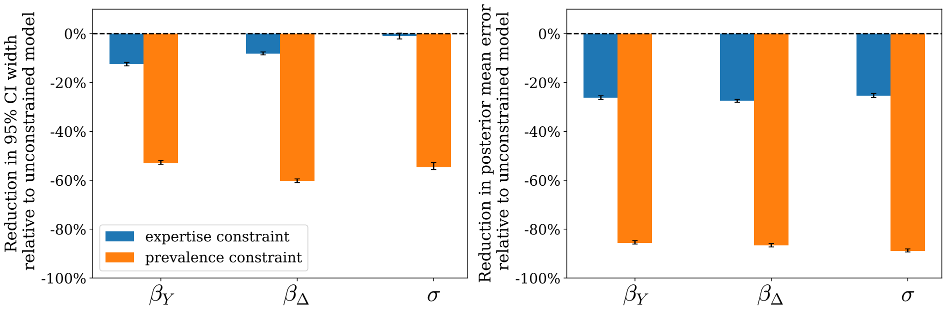


Figure S6: The prevalence and expertise constraints still improve parameter inference when quadrupling the number of features relative to Figure S3. Results are shown using synthetic data from the Bernoulli-sigmoid model with uniform unobservables. Both constraints produce more precise and accurate inferences on this synthetic data. We plot the median across 200 synthetic datasets. Errorbars denote the bootstrapped 95% confidence interval on the median.

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. 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 eq. (3). 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 (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 S2. 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 eq. (2). 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 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 $\boldsymbol{\beta}_{\Delta_2} = \boldsymbol{\beta}_{\Delta_3} = \boldsymbol{\beta}_{\Delta_4} = 0$. 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.

We show results in Figure S3. 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.

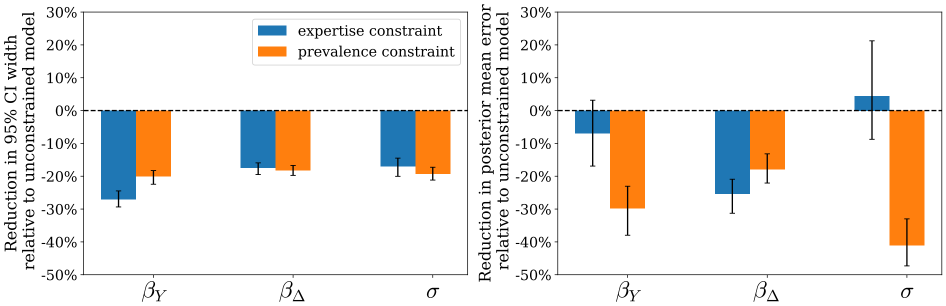


Figure S7: The prevalence and expertise constraints still improve parameter inference even when using pairwise nonlinear interactions between features (rather than only linear terms, as shown in Figure S3). Results are shown using synthetic data from the Bernoulli-sigmoid model with uniform unobservables. 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.

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 S4. 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 S5. 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 S6. 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 S7. 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 [98]. Thus in order for these more complex models to converge, we reduce the prevalence constraint penalty weight to 10,000. Otherwise, we use the same simulation parameters as our standard uniform model experiments.[4]

[4] 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.

E Additional experiments on cancer data

Here we provide additional sets of experiments. We provide additional model validations (Appendix E.1), a comparison to various baseline models (Appendix E.2), and robustness experiments (Appendix E.3).

E.1 Validating the model

In §4.2, we show that the model's inferred risks predict cancer diagnoses. Here we present three more validations.

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 and Y but is not included in the data given to the model.[5] People in the highest inferred unobservables quintile are 2.2× likelier to have a family history of cancer than people in the lowest quintile (15.9% vs 7.4%).

β_Y captures known cancer risk factors: β_Y measures each feature's contribution to risk. The top left plot in Figure 1 shows that the inferred β_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 [25, 26].

β_Δ captures known public health policies: In the UK, all women aged 50-70 are invited for breast cancer testing every 3 years [20]. 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 1 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.

E.2 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.

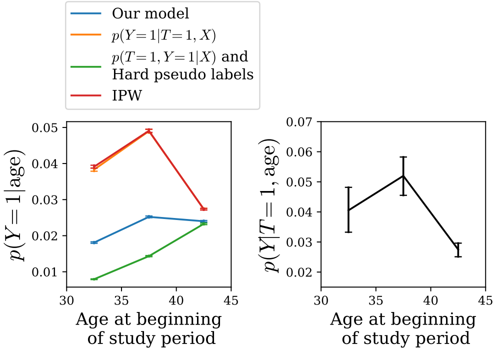


Figure S8: We run three sets of baseline models: (i) models trained solely on the tested population, estimating $p(Y = 1|T = 1, X)$; (ii) models which treat the untested group as negative, estimating $p(T = 1, Y = 1|X)$; and (iii) other selective labels baselines (IPW and hard pseudo labels). Both IPW and the model estimating $p(Y = 1|T = 1, X)$ 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 $p(T = 1, Y = 1|X)$ also learn plausible age trends, but they underperform our Bayesian model in predictive performance.

Comparison to models trained solely on the tested population: The first baseline that we consider is a model which estimates $p(Y_i = 1|T_i = 1, X_i)$ without unobservables: i.e., a model which predicts outcomes using only the tested population.[6] 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

[5] 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 and Y, as is desirable for features in X.

[6] 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.

people who were tested [1, 2]; to predict hypertrophic cardiomyopathy among people who received gold-standard imaging tests [57]; and to predict discharge outcomes among people deemed ready for ICU discharge [42]. It has also been used in the settings of policing [4], government inspections [41], and lending [38].

As shown in Figure S8, we find that all three models trained solely on the tested population learn that cancer risk first increases with age and then falls sharply, contradicting prior epidemiological and physiological evidence [10, 33–35]. We see this same trend for a model fit without a prevalence constraint in §4.4. 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 $p(T = 1, Y = 1|X)$, an approach used in prior selective labels work [89, 99, 100]. We find that, though this baseline no longer learns an implausible age trend, it underperforms our model both 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). We note that this baseline is a special case of our model with the prevalence constraint set such that $p(Y = 1|T = 0) = 0$, 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.

Comparison to other selective labels baselines: We also consider two other common selective labels baselines [100]. First, we predict hard pseudo labels for the untested population [101]: 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 $Y = 0$, 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) [65]: i.e., we train a classifier on the tested population but reweight each sample by the inverse propensity weight $\frac{1}{p(T=1|X)}$.[7] As shown in Figure S8, 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.

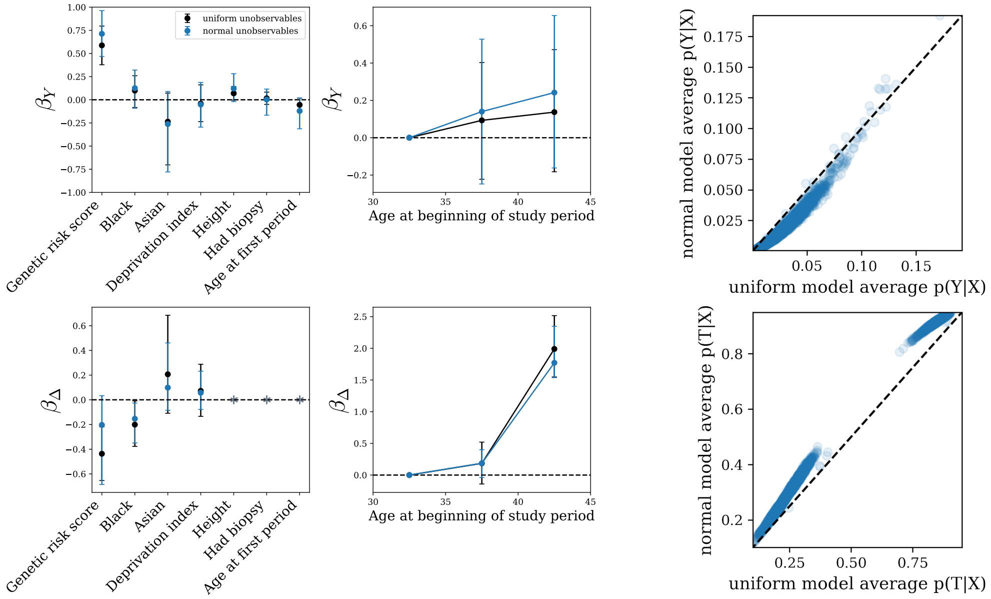
E.3 Robustness checks for the breast cancer case study

Our primary breast cancer results (§4) are computed using the Bernoulli-sigmoid model in eq. (2). 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 [10]. 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 (eq. (2)) to the normal unobservables model (eq. (6)). As described in Appendix D, the normal unobservables model does not allow us to marginalize out unobservables Z and thus converges more slowly. Hence, for computational tractability, we run the model on a random subset of $\frac{1}{8}$ of the breast cancer dataset. In Figure S9a, we compare the $\boldsymbol{\beta_Y}$ and $\boldsymbol{\beta_\Delta}$ coefficients from both models. The estimated coefficients remain similar for both models, with similar trends in the point estimates and overlapping confidence intervals. Figure S9b shows that the inferred values of $p(Y_i|X_i)$ and $p(T_i|X_i)$ for each data point also remain highly correlated across both models, indicating that the models infer similar testing probabilities and disease risks for each person.

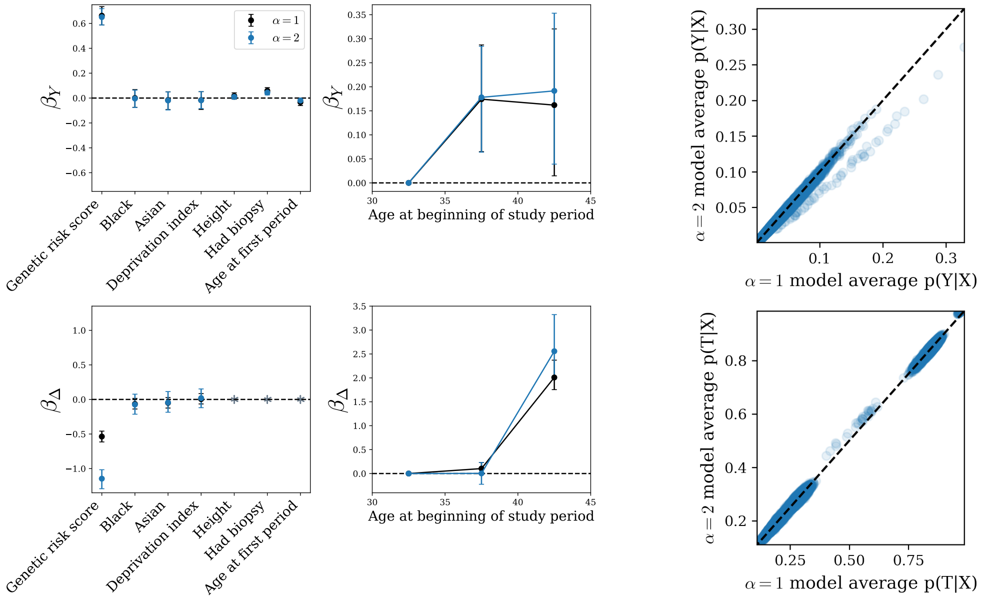
Consistency across different α: We compare the uniform unobservables model with $\alpha = 1$ to a uniform unobservables model with $\alpha = 2$. In Figure S10a, we compare the $\boldsymbol{\beta_Y}$ and $\boldsymbol{\beta_\Delta}$ coefficients from both models. The inferred $\boldsymbol{\beta_Y}$ and $\boldsymbol{\beta_\Delta}$ 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 the genetic risk score. While both the $\alpha = 1$ and $\alpha = 2$ models find a negative coefficient

[7] We clip $p(T = 1|X)$ to be between [0.05, 0.95], consistent with previous work.



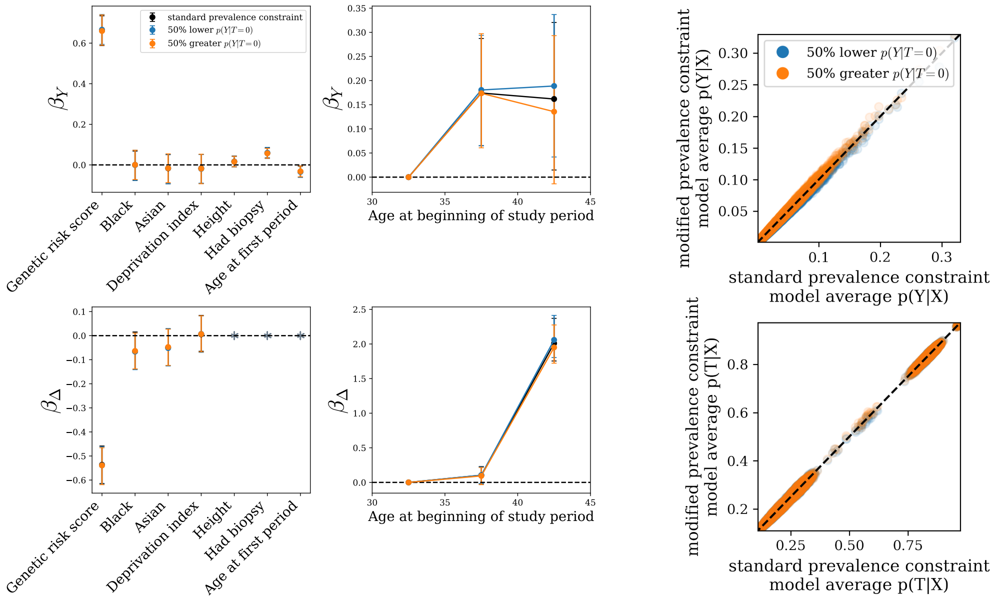
(a) (b)

Figure S9: We compare the results from the uniform unobservable model in eq. (2) (black) and the normal unobservable model in eq. (6) (blue). Figure S9a: 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 S9b: 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 S10: We compare the results from the uniform unobservable model with $\alpha = 1$ (black) and $\alpha = 2$ (blue). Figure S10a: 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 S10b: 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 S11: We compare the results from the uniform unobservables model with the standard prevalence constraint of $\mathbb{E}[Y] = 0.02$ informed by the UK's breast cancer incidence statistics [10] (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 S11a: The predictions for all three models are similar as seen by the similar trends in the point estimates and overlapping confidence intervals. Figure S11b: All three 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.

for $\boldsymbol{\beta_\Delta}$ on the genetic risk score, indicating genetic information is underused, the coefficient is less negative when $\alpha = 1$. These different coefficients occur because altering α changes the assumed parametric 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 parametric relationship between risk and human decision-making [22, 53, 54, 102]. We can restrict the plausible values of α, and thus of $\boldsymbol{\beta_\Delta}$, using any of 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 both α and σ^2 in a model with non-binary Y outcomes (e.g. tumor size or stage) when both parameters can be simultaneously identified.

To systematically confirm model consistency, we again compare the inferred values of $p(Y_i|X_i)$ and $p(T_i|X_i)$ for each data point. As shown in Figure S10b, we confirm that these estimates remain highly correlated across both models, indicating that the models infer very similar testing probabilities and disease risks for each person.

Consistency across different prevalence constraints: The prevalence constraint fixes the model's 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 §4, we set the prevalence constraint to 0.02 based on previously reported breast cancer incidence statistics in the UK for the relevant age groups [10]. However, in general, disease prevalence may not be exactly known [13–15]. Hence, in order to check the robustness of our results to plausible variations in the value of the prevalence constraint, we compare our original results to those with two other prevalence constraints that correspond to 50% lower and 50% higher values of $p(Y = 1|T = 0)$. This yields overall prevalence constraints of $\mathbb{E}[Y] \approx 0.018$ and 0.022, respectively. In Figure S11a, 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 (§4.4). In Figure S11b, 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.

F 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 $T_i = 0$ and $Y_i = 1$ (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 [25, 26, 33]. Second, since features are designed to be used in predicting T and Y, they must be measured prior to T and Y (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[8], 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 missing data for any included features (e.g. responded "do not know"). (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 [104]. (v) We removed everyone above the age of 45 at the beginning of the observation period, since 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.[9]

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 ($T = 0$) at the baseline, we use data from the three UKBB follow-up visits. We only consider the subset of people

[8] The Townsend deprivation index is a measure of material deprivation that incorporates unemployment, non-car ownership, non-home ownership, and household overcrowding [103].

[9] 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, $\geq$55). We find that if anything, predictive performance when using the full cohort is better than when using only the younger cohort from our main analyses in §4.2. Specifically, the model's quintile ratio is 4.6 among the tested population ($T = 1$) and 7.0 among the untested population ($T = 0$) that attended a follow-up visit.

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 and Y. 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.