A Practical Upper Bound on Selection Bias Effects in Medical Prediction Models

Kara Liu
Stanford University
Stanford, California, USA

Maggie Wang
Stanford University
Stanford, California, USA

Russ B. Altman
Stanford University
Stanford, California, USA

Abstract

Selection bias is a common and often unavoidable aspect of real-world data that challenges the generalizability of machine learning models. When models trained on biased data are deployed in the broader target population, poor model generalization may lead to real harm, particularly in high-risk settings such as healthcare. This risk highlights the need for practitioners to reliably assess model generalizability prior to deployment. However, existing methods for predicting model performance rely on unrealistic access to the target distribution or knowledge of the selection mechanism causing bias. To address these limitations, we propose a novel upper bound on the worst-case model performance on the target population under the realistic setting where the selection mechanism and the target population data are only partially observed. We demonstrate the validity and practical utility of our method through experiments on fully synthetic data, semi-synthetic data derived from the All of Us Research Program, and real-world selection bias in MIMIC-IV. Our work offers a principled and practical tool to estimate the impact of selection bias in an otherwise intractable setting, thereby enabling practitioners to build safer and more generalizable models in healthcare and beyond. We release our code for public use at https://github.com/kara-liu/selection_gap_est/.

CCS Concepts

• **Computing methodologies** → *Supervised learning*; Machine learning; **Model verification and validation**; • **Mathematics of computing** → Probability and statistics.

Keywords

Selection bias, generalizability bounds, healthcare, model auditing

ACM Reference Format:
Kara Liu, Maggie Wang, and Russ B. Altman. 2026. A Practical Upper Bound on Selection Bias Effects in Medical Prediction Models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818112

1 Introduction

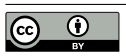
As machine learning models are increasingly deployed in real-world settings, it is imperative to understand how their performance may be affected by selection bias. When models are trained on data that represents only a subset of the population, their failure to generalize to the broader target population could inflict real harm, often in ways that reify discrimination against underrepresented groups [2, 30, 36]. This risk is particularly acute in medical applications, where the data used to support high-stakes decisions are often heavily skewed by selection bias [10, 33, 61]. For instance, selection bias in biobank data, electronic health records, and randomized controlled trials has led to biased estimates in genome-wide association studies [69], analyses of COVID-19 risk factors [61], the prediction of sepsis in hospitals [83], and drug dosage recommendations that are suboptimal for non-Caucasian populations [49].

To support safe deployment under selection bias, machine learning model developers must be able to assess how a model trained on biased data will perform on its intended target population, in a way that is both practical and grounded in the underlying selection mechanism. As a motivating example, the national deployment of the Epic Sepsis Model, which was trained on data from just three healthcare systems, was later criticized for poor generalization [58, 83]. These issues might have been prevented had developers been able to foresee how the model would perform on the broader U.S. population.

Selection bias has been studied in other disciplines, notably in the context of causal effect estimation [3, 57] or under a limited set of selection mechanisms in domain adaptation [52, 75]. However, existing methods that estimate model generalizability often rely on unrealistic assumptions. For instance, density ratio estimation requires access to the full target distribution [50, 52, 75], while inverse probability of participation weighting [11, 15, 67, 80] assumes full observability of the features causal of selection. In practice, however, access to target data is often limited, as collecting fully representative samples – for example, medical records from the entire U.S. population – can be prohibitively expensive, logistically infeasible, and may conflict with privacy regulations. Moreover, model developers rarely observe the explicit underlying selection mechanism, given the complexity in delineating the causes of study participation or health care utilization. As a result, there are currently few practical and principled approaches for practitioners to audit models for selection bias.

Contributions. In this work, we propose a tractable estimate of a prediction model's worst-case performance on an unobserved target population. Our paper offers three important contributions:

(1) We consider a more realistic setting that requires only partial observability of the selection mechanism and target distribution. Target data sources, such as national registries and census microdata, often provide full joint coverage of a few basic sociodemographic variables, along with marginal summary statistics such as first and second moments [25, 29, 67, 69, 80]. Motivated by these real-world constraints,



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818112

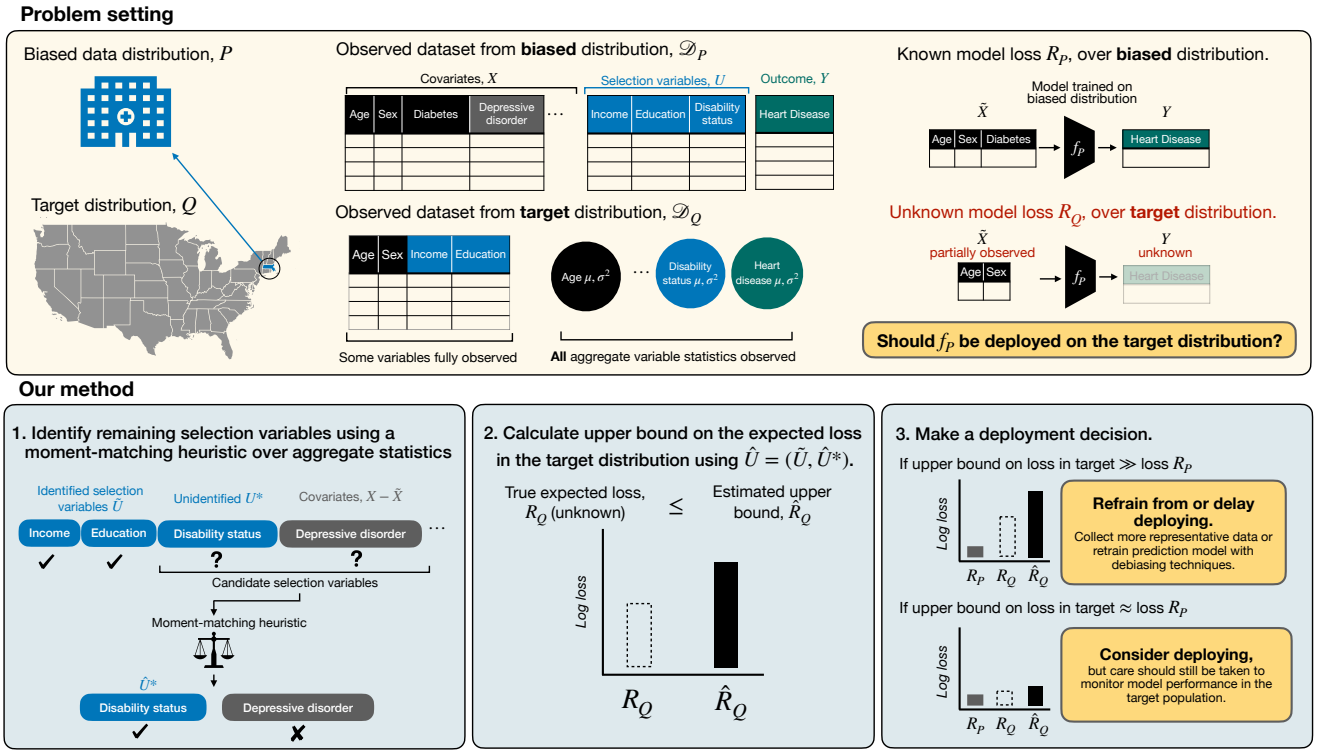


Figure 1: Pipeline illustrating how our method estimates a generalization bound under limited observation of the target-data, thereby enabling better informed decisions of model deployment.

our method assumes: (i) only a subset of variables causing selection has been identified and fully observed and (ii) access to marginal summary statistics for all variables. The feasibility of (i) is further supported by prior work showing that sociodemographic factors such as age, sex, and income, which are often observable in target datasets, are also likely drivers of medical selection bias [68, 69, 71].

(2) We propose a novel upper bound on a prediction model's expected performance on a target dataset. Our bound explicitly models the selection process without requiring full knowledge of the underlying mechanism. To the best of our knowledge, existing methods are incapable of producing such a bound under the realistic constraints outlined in (1).

(3) To render our bound tractable for practitioners auditing models prior to deployment, we propose two heuristics: first, an algorithm that identifies the remaining selection variables, and second, diagnostic techniques for assessing the assumptions underlying our bound.

We evaluate our method in three data settings: (i) simulated selection bias in fully synthetic data, (ii) simulated selection bias in clinical data from the All of Us Research Program [21], and (iii) real selection bias in MIMIC-IV [46]. Across these experiments, we show that our proposed bound is empirically tight and robust, even under realistic data observability constraints. Finally, we provide guidance and diagnostic tools to help practitioners apply and interpret our bound estimate.

2 Related Works

2.1 Selection Bias in Causal Inference

Selection bias has been broadly studied across disciplines including econometrics [38–40, 79], causal inference [3], epidemiology [41, 45, 57, 73], statistics [24, 59], social science [7, 82], and clinical informatics [37, 78]. In these fields, the focus has largely been on causal effect estimation where bias arises when generalizing estimates from a non-representative study sample to a target population. Many debiasing methods have been developed to address this issue, including g-formula adjustment [3, 54], propensity score weighting [42, 54, 64], and Heckman correction [39]. However, these methods are not directly applicable to our setting of assessing the generalizability of prediction models. Furthermore, these methods often rely on unrealistic assumptions, such as full observability of the causes of selection.

Another related line of work has estimated robustness to *unobserved* selection or confounding via sensitivity analysis and partial identification. Motivated by Rosenbaum-style sensitivity analysis [65], Zhao et al. [86] proposed bootstrapped confidence intervals based on a marginal sensitivity model that bounds the unknown selection probability. Other approaches avoid modeling the unobserved confounding mechanism by providing explicit sensitivity parameters to bound confounder-treatment and confounder-outcome associations [23] or by using partial identification to propose worst-case bounds, as with Manski-style bounds [59, 60]. While these

approaches make minimal assumptions on the underlying selection mechanism, they rely on ad hoc specification of the sensitivity parameters which can lead to overly loose bounds in practice.

Selection bias may also be framed as a case of estimation under missingness-not-at-random of the covariates, outcome, or both. When data are only observed conditioned on a positive outcome, [55, 76, 77] proposed methods to identify causal effects by leveraging instrumental variables. However, these works assume full observation of a valid instrument and covariates in the target distribution.

2.2 Domain Adaptation

In machine learning, selection bias has most often been studied through the lens of domain adaptation where a model trained on a source domain (or distribution) must generalize to a target domain [50, 52]. Domain adaptation can be approached by learning domain-invariant representations [28]; reweighting samples from the source domain [50, 52, 75] using weights obtained by probabilistic classification [15], moment matching [20, 43], or density matching [74]; and through distributionally robust optimization [6, 70], which learns a model that minimizes worst-case performance over "uncertainty sets" of the observed data. However, as in the case of causal inference methods, domain adaptation methods typically rely on the variables causing the data shift to be observable from the target distribution, which is an unrealistic assumption in real-world settings. Additionally, these methods are used *during* model training to improve out-of-distribution performance, rather than *after* model training to audit model generalizability.

2.3 Generalization Bounds

Another relevant area of work involves estimating model generalization bounds under distribution shift. Cortes et al. [18] derive generalization bounds that depend on the stability of the associated importance weights, and Ben-David et al. [4, 5] propose an upper bound on the generalization error of a model trained on samples from distribution P when applied to distribution Q, using $\mathcal{H}$-divergence to characterize the distance between P and Q. Similar bounds have also been proposed using Wasserstein distance [19] and f-divergence [1]. As with domain adaptation methods, these worst-case bounds are intractable without strong assumptions on what is observable.

3 Method

In Sections 3.1, we introduce notation, key assumptions, and observability constraints. In Section 3.2, we introduce our method for calculating the generalization upper bound. In Section 3.3, we outline how our method is practically implemented via a heuristic algorithm that nominates the remaining selection variables. We outline our method in Figure 1.

3.1 Problem Formulation

We denote an upper-case letter V as a single or set of random variables, the variable's set space as $\mathcal{V}$, and the observation in that space as the lower-case v. We use the notation $\mathbb{E}_Q[\cdot]$ to describe the expectation with respect to sampling from a distribution Q. We let $V_j^{(i)}$ denote the j-th variable and i-th unit of a multivariate sample.

Similar to the notation of [66, 85], let Q be the target distribution over the variables (X, U, Y, S), where $S \in \{0, 1\}$ is a binary selection indicator, $U \in \mathcal{U} \subset \mathbb{R}^k$ are the variables causal of selection, $Y \in \mathcal{Y} \subset \mathbb{R}$ is the outcome, and $X \in \mathcal{X} \subset \mathbb{R}^d$ are all other covariates. We assume the biased[1] distribution P is generated by selective sampling from Q such that the (X, Y)-marginal of Q conditioned on $S = 1$ is exactly the distribution P. That is, denoting $p(V)$ as the probability distribution of any variable $V \in (X, Y)$ under Q, then $p(V \mid S = 1)$ is the variable's distribution under P, where we assume P and Q have common support. Furthermore, we assume that the selection indicator S is defined as a probabilistic function of selection variables U, and thus S is conditionally independent of (X, Y) given U. We summarize the variables in Table 1.

Table 1: Notation for our problem setting.

Var.	Description	Example
Q	Target distribution over (X, U, Y, S)	U.S. population distribution
P	Biased distribution over (X, U, Y) generated by selective sampling from Q	Single hospital distribution
S	Binary selection indicator	If patient in the U.S. attended single hospital
U	Variables causal of selection into biased distribution P	Income, education, disability status
$\tilde{U}$	Variables $\subseteq U$ observed in both P and Q	Income, education
U^*	Variables $:= U \setminus \tilde{U}$ not observed under Q	Disability status
X	All covariates non-overlapping[2] with U	Age, sex, diabetes, depressive disorder
$\tilde{X}$	Covariates $\subseteq X$ used to predict Y	Age, sex, diabetes
Y	Outcome	Heart disease

Assumption 1 (Conditional Independence). $X, Y \perp\!\!\!\perp S \mid U$

Assumption 2 (Common Support). $\forall X \in \mathcal{X}, U \in \mathcal{U}, Y \in \mathcal{Y}$, *if* $p(X, U, Y) > 0$, *then* $p(X, U, Y \mid S = 1) > 0$.

Let $f_P : \tilde{\mathcal{X}} \to \mathcal{Y}$ denote a prediction model trained under the biased distribution P to minimize the expected loss $R_P = \mathbb{E}_P[\ell(f_P(\tilde{X}), Y)]$, where the subvector $\tilde{X}$ of X are the prediction features and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a non-negative loss function.

Assumption 3 (Non-negative Loss). $\ell(f_P(\tilde{X}), Y) \geq 0$ *for all* $\tilde{X} \in \tilde{\mathcal{X}}$ *and* $Y \in \mathcal{Y}$.

To evaluate model generalization on the target distribution Q, we consider estimating the expected loss $R_Q = \mathbb{E}_Q[\ell(f_P(\tilde{X}), Y)]$ under the relaxed but more challenging setting of limited Q observability. Specifically, we assume identification and full observation of a subset $\tilde{U}$ of the selection variables, as well as basic summary statistics for all candidate selection variables. We denote the unknown selection variables as $U^* := U \setminus \tilde{U}$.

Constraint 1 (Absence of Samples $(\tilde{X}, Y) \sim Q$). *The variables* $(\tilde{X}, Y)$ *needed to measure the performance of* f_P *are unobserved in* Q.

Constraint 2 (Observation of Partial $\tilde{U}$ from Q). *We identify a non-empty subset of selection variables* $\tilde{U} \subseteq U$ *and observe* $\tilde{U}$ *under* Q.

[1] In other domains, P is also called the source, sample, or training distribution.
[2] In practice, $\tilde{X}$ and U may overlap, provided all covariates used for prediction are observed in Q.

Constraint 3 (Observation of μ, σ^2 from Q). *For all variables $V \in (X, U)$, we observe its mean $\mu_Q(V) = \mathbb{E}_Q[V]$ and variance $\sigma^2_Q(V) = \mathbb{E}_Q[(V - \mu_Q(V))^2]$.*

Finally, we define the observed dataset $\mathcal{D}_P = \{(X^{(i)}, U^{(i)}, Y^{(i)}, S^{(i)} = 1)\}_{i=1}^n$ drawn independently from the distribution P, on which the prediction model f_P is trained. From Q, we observe the summary statistics defined in Constraint 3, as well as $\mathcal{D}_Q = \{\tilde{U}^{(i)}\}_{i=1}^m$.

3.2 An Upper Bound of R_Q

Under the Constraints 1 - 3 of limited target distribution observability, *the expected loss R_Q is intractable using existing methods.* To address this gap, we propose an upper bound $\hat{R}_Q$ that bounds the performance of the model f_P on the target distribution Q. The construction of the bound relies on both the fully observed $\tilde{U}$ and identification of the remaining selection variables U^*.

Theorem 3.1 (Upper Bound $\hat{R}_Q$). *Under Assumptions 1, 2, and 3,*

$$R_Q \leq \hat{R}_Q := \mathbb{E}_P\left[w(\tilde{U}) \cdot \phi(\tilde{X}, Y, \mathcal{U}^*) \cdot \ell(f_P(\tilde{X}), Y)\right]$$

$$w(\tilde{U}) := \frac{p(\tilde{U})}{p(\tilde{U} \mid S = 1)}$$

$$\phi(\tilde{X}, Y, \mathcal{U}^*) := \frac{\max_{u^* \in \mathcal{U}^*} p(\tilde{X}, Y \mid u^*, \tilde{U}, S = 1)}{p(\tilde{X}, Y \mid \tilde{U}, S = 1)}$$

where u^* is an observation in the subspace of $\mathcal{U}^*$.

The full proof is in Appendix A.1. To build intuition for the proof, observe that under Assumption 2, R_Q can be expressed as a reweighted expectation over P, where the weights correspond to the density ratio of $(\tilde{X}, Y, \tilde{U})$ in P and Q. Although the marginal $p(\tilde{U})$ under Q is known, the conditional distribution $p(\tilde{X}, Y \mid \tilde{U})$ is not. The key insight is that this conditional distribution can be upper bounded by $p(\tilde{X}, Y \mid U)$ under Q. Then, invoking conditional independence in Assumption 1, we can replace this unknown distribution with the known marginal $p(\tilde{X}, Y \mid U, S = 1)$ under P.

We define the *true generalization gap* $R_Q - R_P$ as the loss increase when evaluating f_P on Q versus P, the *upper bound*[3] *generalization gap* as $\hat{R}_Q - R_P$, and the *bound error* as $\hat{R}_Q - R_Q$.

3.3 Practical Bound Estimation

We next outline how to estimate our proposed bound. In Section 3.3.1, we propose a heuristic algorithm to identify the remaining selection variables U^*. In Sections 3.3.2 and 3.3.3, we outline our approach to density estimation and potential assumption violations in finite-sample settings. The pseudocode of our bound estimation method is presented in Algorithm 1.

3.3.1 Heuristic Identification of the Remaining Selection Variables. Our bound requires identifying the remaining selection variables $U^* := U \setminus \tilde{U}$ from the variables observed in $\mathcal{D}_P$. We propose a simple, calibration-based heuristic for nominating U^*.

Suppose that the probability of selection can be expressed as $p(S = 1 \mid U) = g(U\beta)$ for some link function g and coefficient vector β. Let $C := (X \setminus \tilde{X} \setminus Y, U^*)$ denote the set of possible selection variables U^*, where U^* cannot overlap with Y or $\tilde{X}$. We can write the probability of selection equivalently as $p(S = 1 \mid \tilde{U}, C) = g(\tilde{U}\omega + C\gamma)$, where $\gamma_j = 0$ for all variables $C_j \in X$. The task of determining which variables in C are selection variables thus becomes the simpler task of determining which γ_j are non-zero.

Similar to existing calibration-based methods [53, 84], we use the following moment-matching estimating equation over $\mathcal{D}_P$ to empirically solve for ω and γ:

$$\sum_{i:S^{(i)}=1} \frac{m(\tilde{U}^{(i)}, C^{(i)})}{g(\tilde{U}^{(i)}\omega + C^{(i)}\gamma)} = \mathbb{E}_{\mathcal{D}_Q}[m(\tilde{U}, C)]$$

where m is the user-specified moment map evaluated at each sample and $\mathbb{E}_{\mathcal{D}_Q}[m(\tilde{U}, C)]$ is the corresponding empirical moment vector under the target distribution Q. Given the observation of μ_Q, σ^2 from Constraint 3, the choice of m may match on first moments, second moments, or their concatenation.

We construct confidence intervals for γ_j using the bootstrapped distribution with percentile parameter α. The corresponding C_j whose intervals do not contain zero are selected as the estimated U^*. Further details, including an adaptation for searching in high dimensions, are presented in Appendix B.

3.3.2 Density Estimation. When the variables are fully categorical, the conditional densities $p(\tilde{X}, Y \mid \tilde{U}, \hat{U}^*, S = 1)$ and $p(\tilde{X}, Y \mid \tilde{U}, S = 1)$ can be estimated from table counts. For data involving continuous variables, the density functions may be estimated using kernel density estimators or conditional normalizing flow models [62, 81]. Estimation of the propensity $w(\tilde{U}) = p(\tilde{U}) \,/\, p(\tilde{U} \mid S = 1) = p(S = 1) \,/\, p(S = 1 \mid \tilde{U})$ is even simpler and can be computed by fitting a classifier [15] or table counts if data are discrete. Although our main focus is on categorical data given its omnipresence in medical settings, we discuss continuous density estimation in Appendix C.

Algorithm 1 Upper bound estimation on the generalization performance R_Q

Input: prediction model f_P; biased dataset $\mathcal{D}_P = \{X^{(i)}, Y^{(i)}, U^{(i)}, S^{(i)} = 1\}_i$; target dataset $\mathcal{D}_Q = \{\tilde{U}^{(k)}\}_k$; external means and variances $\mu_Q(V), \sigma^2_Q(V)\ \forall V \in (X, U)$; α-level for heuristic algorithm

Output: $\hat{R}_Q$, as defined in Theorem 3.1

```
Û* ← output selection variables from heuristic search with significance α
Estimate the conditional density functions p(X̃, Y | Ũ, Û*, S = 1) and p(X̃, Y | Ũ, S = 1) using D_P
Estimate the propensity weight w(Ũ) = p(Ũ) / p(Ũ | S = 1) using both D_P, D_Q
w ← empty weight vector of dimension |D_P|
for all samples (X̃^(i), Y^(i), Ũ^(i)) ∈ D_P do
    φ1 ← max_{u* ∈ Û} p(X̃^(i), Y^(i) | Ũ^(i), u*, S = 1)
    φ2 ← p(X̃^(i), Y^(i) | Ũ^(i), S = 1)
    w_i ← w(Ũ^(i)) · (φ1/φ2)
end for
return R̂_Q = E_{D_P}[w · ℓ(f_P(X̃), Y)]
```

[3] Estimating the upper bound is appropriate when higher loss ℓ indicates worse performance (such as in the case of logloss or Brier score); if lower loss ℓ indicates worse performance (such as with precision or accuracy), the lower bound $\hat{R}_Q \leq R_Q$ can be constructed by taking the minimum instead of the maximum over U^*.

3.3.3 Testing for Assumption Violations When R_Q Is Observed. Our upper bound assumes conditional independence of selection given U (Assumption 1) and common support between P and Q (Assumption 2). However, these assumptions can fail empirically in finite-sample settings.

If the true R_Q is known, as in settings of simulated selection bias, we can explicitly test how each assumption violation affects the bound error by decomposing our method's bound error $\hat{R}_Q - R_Q$ into a telescoping sum of three factors:

$$\hat{R}_Q - R_Q = \Delta_{\text{TBE}} + \Delta_{\text{CI}} + \Delta_{\text{CS}} \tag{3.1}$$

where each $\Delta_{(\cdot)}$ term is defined in Appendix A.2. At a high level, Δ_{CI} is zero when the Conditional Independence assumption holds, and Δ_{CS} is zero if the Common Support assumption holds. Finally, Δ_{TBE} measures the Theoretical Bound Error, the gap between our bound estimate $\hat{R}_Q$ and the true R_Q when the two aforementioned assumptions are satisfied. This decomposition therefore quantifies how violations of conditional independence and common support cause the final bound error to deviate from the theoretical bound error.

3.3.4 Testing for Assumption Violations in Practice. However, the telescoping sum in Equation 3.1 is usually intractable as R_Q is often unknown. Therefore, to approximately test for assumption violations, we present three diagnostics that are straightforward, computationally inexpensive, and can be readily applied using our code implementation. We provide additional details on the diagnostics in Appendix C.

Common Support Diagnostics.

(1) **KS Test:** The overlap between the observed propensity distribution $p(S = 1 \mid \tilde{U})$ under P and Q can be easily evaluated using a Kolmogorov-Smirnov (KS) test, or a similar statistical test.
(2) **Weight Design Effect d_{eff}:** Moment-matching methods, such as our proposed heuristic in Section 3.3.1, often exhibit instability or poor convergence behavior if the two distributions lack overlap. The stability of the resulting weights, measured via the design effect d_{eff} [51], can diagnose potential violations of common support.

Conditional Independence Diagnostic.

(3) **Propensity Invariance:** We propose a diagnostic that approximately assesses the conditional independence assumption $p(V \mid U, S = 1) = p(V \mid U, S = 0), \forall V \in (X, Y)$. However, under Constraints 1 - 3, the true U is unknown and we only observe samples where $S = 1$. We instead use the observed propensity $\hat{S} = p(S = 1 \mid \tilde{U})$ and predicted $\hat{U} := (\tilde{U}, \hat{U}^*)$ from our moment-matching method to assess for equality across $p(V \mid \hat{U}, \hat{S} = s_1) = p(V \mid \hat{U}, \hat{S} = s_2)$, $\forall s_1, s_2 \in [0, 1], \forall V \in (X, Y)$.

4 Experimental Setup

4.1 Data

We evaluate the quality of our bound $\hat{R}_Q$ in three data settings: (i) fully synthetic data; (ii) semi-synthetic data, where we simulate an EHR-specific selection mechanism in clinical data from All of Us; and (iii) real-world selection bias in MIMIC-IV. We provide additional details for each dataset, including preprocessing steps, in Appendix C.

4.1.1 Synthetic Data. To generate the fully synthetic target dataset $\mathcal{D}_Q$, we sample binary variables U with bivariate correlation ρ, and then generate binary X and Y as linear logistic functions of U. We then sample our biased dataset $\mathcal{D}_P = \{(X, U, Y, S) \in \mathcal{D}_Q : S = 1\}$ through a logistic selection model:

$$p(S = 1 \mid U) = \frac{1}{1 + \exp(-g(U\beta))}$$
$$S \sim \text{Bernoulli}(p(S = 1 \mid U))$$

where we test both linear and nonlinear link functions, g.

4.1.2 All of Us. The All of Us Research Program [21] is a demographically diverse biobank based in the U.S. with over 600,000 participants. It includes sociodemographic and biomarker information collected at enrollment, along with longitudinal outcomes from linked medical records.

To form the target dataset $\mathcal{D}_Q$, we filter All of Us participants to a cohort of 255,612 participants. We simulate selection of the biased dataset $\mathcal{D}_P$ given the same logistic selection mechanism described in the fully synthetic setting, where we explicitly define EHR-specific selection variables U (e.g., income level or insurance status) [12, 16, 26, 27, 63, 67]. For prediction, we consider 19 health outcomes Y (e.g., Type 2 diabetes mellitus) and 41 binary features X (e.g., blood pressure, lifestyle factors).

4.1.3 MIMIC-IV. MIMIC-IV is a deidentified dataset containing over 200,000 patients admitted to the emergency department at the Beth Israel Deaconess Medical Center in the U.S. [31, 47]. MIMIC-IV is a widely-used benchmark in machine learning, and it is therefore critical to audit whether selection bias leads to performance degradation when models are generalized to broader populations.

We treat MIMIC-IV as the biased dataset $\mathcal{D}_P$ and All of Us as the target dataset $\mathcal{D}_Q$. Because MIMIC-IV contains data from a single U.S. hospital and All of Us provides nationally-representative data, this setup naturally reflects the realistic scenario of a complex and unknown selection mechanism. Furthermore, using All of Us as the target enables method validation against the true R_Q when the variables $\tilde{X}, Y$ are observed.

We conduct three real-world experiments: first, we consider two prediction tasks (Y=hypertension and Y=Type 2 diabetes mellitus) where $\tilde{X}, Y$ variables are observed in both datasets, enabling validation of our bound against the true R_Q; second, we evaluate one task (Y=hospital mortality) that reflects the realistic scenario where R_Q is unknown. Motivated by prior work in EHR-specific biases [12, 16, 26, 63], we select the selection variables $\tilde{U}$ as a subset of age, insurance type, and primary language.

4.2 Prediction Tasks

For the prediction model f_P, we learn $p(Y \mid \tilde{X}, S = 1) = f_P(\tilde{X})$ either using XGBoost [14] or an elastic net regularized logistic regression model with class-balancing weights and regularization parameters chosen via cross-validation. In practice, we do not observe a substantial difference in bound characteristics under different prediction models. For each set of experiments, we run a data-driven

search to identify n_{tasks} prediction tasks such that the resulting generalization gap is sufficiently large, i.e., $R_Q - R_P > t_R$. We outline this search in Algorithm C.

4.3 Evaluating Our Proposed Bound in Simulated Selection Settings

We first evaluate our bound estimation method in the fully synthetic and All of Us data settings. By simulating selection, we can validate if our method, which assumes limited target data observability (Constraints 1 - 3), actually recovers the true expected loss R_Q and selection variables U in practice. We run the following experiments, which are described in more detail in Appendix C.

4.3.1 Correctness of Heuristic Identification of Selection Variables. We test how well our moment-matching heuristic (Section 3.3.1) recovers the remaining selection variables compared to random selection and selection based on maximum correlation (regular and Cohen's d) with $\tilde{U}$. For each selection strategy, we compute the F2 score, precision, recall, and the Jaccard index of the selected $\hat{U}^*$ compared to the true U^*.

4.3.2 Robustness to Assumption Violations. We examine how violations of common support and conditional independence may affect the behavior of our bound estimate. To control the degree of assumption violation, we vary four parameters: sample size $|\mathcal{D}_Q|$, the strength of the selection mechanism, covariate imbalance, and the number of extraneous features $X \setminus \tilde{X}$. For each parameter setting, we decompose the estimated bound error $\hat{R}_Q - R_Q$ into the telescoping sum from Equation 3.1.

4.3.3 Validating Our Bound Estimate. For each prediction task, we run our method on all possible observed subsets $\tilde{U} \subseteq U$ and compute the estimated upper bound $\hat{R}_Q$: first, using the true selection variables, denoted as 'UB (true U^*)'; second, using our heuristic, denoted as 'UB (heuristic U^*)'. We then compare the estimated generalization gap (or bound error) with the true quantity.

We compare against the following baselines: *naive inverse probability of participation weighting* (IPPW) [13, 17, 64, 67, 80], which estimates sample weights using the fully observed $\tilde{U}$; *empirical calibration* [22], which estimates sample weights to balance the first moments of all variables in P and Q; *entropy balancing* [35], a form of calibration that additionally matches second moments; and *raking* (iterative proportional fitting) [22], a form of calibration that aligns categorical sample data to target table counts. We describe these baselines in detail in Appendix C. In Appendix **??**, we also provide results on KLIEP [74], KMM [20, 43], logistic regression classification [15, 75], RuLSIF [56], and uLSIF [48], which assume unrealistic data availability and are excluded from the main analysis.

4.4 Application of Proposed Bound to Real-World Settings

We next validate our method in real-world settings and discuss how to practically use our method for model auditing.

4.4.1 Robustness to Assumption Violations. Assuming R_Q is unknown, we apply the three proposed assumption violation diagnostics from Section 3.3.4 on fully synthetic data and outline how to interpret the results in practice.

4.4.2 Validating Our Bound Estimate. We evaluate our method on three tasks with real selection bias in MIMIC-IV when compared to the more diverse target population in All of Us. In two tasks, we compare our method's predicted $\hat{R}_Q$ to the true R_Q. For the third task, we estimate $\hat{R}_Q$ and provide guidelines for practically validating our method when R_Q is unknown.

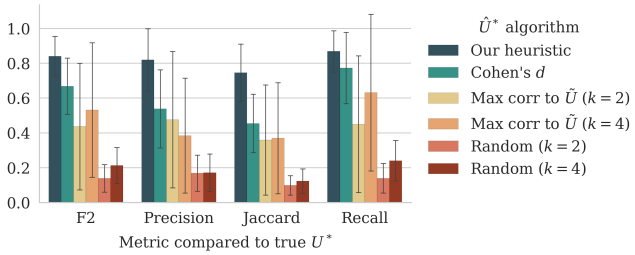


Figure 2: Comparing our heuristic identification algorithm against other baselines in identifying the true selection variables U^* in fully synthetic data. Our heuristic yields high accuracy and thus is a close approximation to the bound under true selection variable identification.

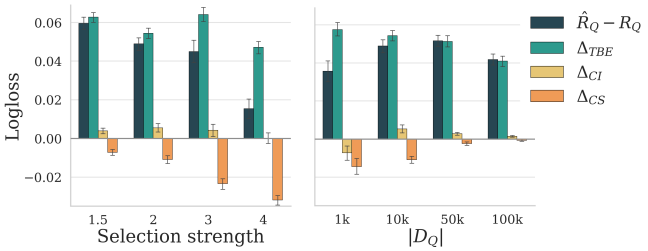


Figure 3: Decomposition of the bound error $\hat{R}_Q - R_Q$ into three terms (see Section 3.3.3) to evaluate assumption violations in fully synthetic data, as sample size and selection strength vary. Extreme violations of common support (i.e., large $|\Delta_{CS}|$) or conditional independence (i.e., large $|\Delta_{CI}|$) may lead to an invalid upper bound.

5 Results

5.1 Evaluating Our Proposed Bound in Simulated Selection Settings

5.1.1 Correctness of Heuristic Identification of Selection Variables. In Figures 2 and C, we demonstrate that our proposed moment-matching heuristic outperforms all baselines and is reasonably able to identify the remaining selection variables U^* with F2 scores of 0.84 and 0.81 for synthetic data and All of Us, respectively. These results suggest that the bound estimated using our heuristic $\hat{U}^*$ closely approximates the bound under the true selection variables U^*.

5.1.2 Robustness to Assumption Violations. In Figure 3, we apply the proposed error bound $\hat{R}_Q - R_Q$ decomposition while varying the strength of the selection mechanism and the sample size $|\mathcal{D}_Q|$. Results are shown for fully synthetic data using a linear selection mechanism and are averaged across $n_{\text{tasks}} = 20$, all $\tilde{U} \subseteq U$, and 5 random seeds. Additional results on All of Us data, as well as varying covariate $\tilde{X}$ imbalance or increasing $\dim(X \setminus \tilde{X})$, are provided in Appendix C. As expected, aggressive selection, small sample size, and high variable imbalance can violate assumptions of conditional independence or common support, as reflected by larger average Δ_{CI} and Δ_{CS} terms. In these scenarios, caution using our method – and perhaps deployment of the model in question – is warranted until more data can be collected. For instance, in tasks where sample size is insufficient, a large negative Δ_{CS} term may dominate due to lack of common support, and $\hat{R}_Q$ may be underestimated. Nonetheless, we find on average our method yields a valid upper bound under reasonable sample sizes and modest selection strength.

5.1.3 Validating Our Bound Estimate. In Table 2, we demonstrate the quality of our bound estimate across $n_{\text{tasks}} = 30$, all subsets $\tilde{U} \subseteq U$, and 20 random seeds for both fully synthetic and All of Us data with a nonlinear selection mechanism. Our results confirm our estimate's empirical validity, with 97% of all tasks in All of Us yielding a valid upper bound versus 82% for the next best baseline. Although our bound in theory could be prohibitively large, in practice it is non-vacuous. For instance, the 95th percentile of the bound error in the All of Us experiments is 0.17 logloss. In Appendix C, we present additional results, including performance on linear selection, synthetic continuous, and synthetic high-dimensional data.

We next examine our method's robustness to which selection variables $\tilde{U}$ are fully observed (Constraint 2). In Figure 4, we plot the bound error $\hat{R}_Q - R_Q$ based on the dimension $\dim(\tilde{U})$ of observed variables out of $\dim(U) = 5$ variables total, using All of Us data and nonlinear selection. As expected, our bound estimate slightly improves as the availability of selection variables $\tilde{U}$ increases. In Figures C and C, we also show that the bound error is correlated with how well $\tilde{U}$ predicts S. However, our method provides reasonable bounds even when $\dim(\tilde{U}) = 1$, highlighting robustness to settings where target observability is highly limited. Additional results are included in Appendix C.

In Figure 5, we examine the generalization gap for two tasks in the All of Us data, where we use a linear selection mechanism designed to simulate EHR-specific selection bias. Five additional tasks are plotted in Figure C. We observe that the baselines severely underestimate generalization performance, risking confident deployment of a model that might perform poorly in practice. On the other hand, our method provides a tight and valid upper bound on the real generalization gap.

5.2 Application of Proposed Bound to Real-World Settings

5.2.1 Robustness to Assumption Violations. In Table 3, we show the application of our three assumption violation diagnostics on fully synthetic data across $n_{\text{tasks}} = 10$, subsets $\tilde{U} \subseteq U$, and 2 seeds; implementation details are provided in Appendix C. As expected, higher levels of assumption violations (as indicated by higher scores for all diagnostics) occur under low sample size and high selection strength. However, even when assumptions were violated, our bound remained largely valid.

Table 2: Bound error $\hat{R}_Q - R_Q$ for both fully synthetic and All of Us data. "Validity" denotes the fraction of tasks that yields a valid upper bound $> \epsilon$, for some small negative ϵ. "(0.05, 0.95)" denotes the 5th and 95th percentiles. "d_{eff}" is the design effect. Our generalization estimate UB (heuristic U^*) provides an non-vacuous upper bound with higher rates of validity than the baselines.

		$\hat{R}_Q - R_Q$			
		$\mu \pm \sigma$	Validity	(0.05, 0.95)	d_{eff}
Synthetic	**UB (true U^*)**	0.05 ± 0.06	0.90	$(-0.02, 0.15)$	4.9
	UB (heuristic U^*)	0.06 ± 0.05	0.90	$(-0.02, 0.15)$	4.7
	Naive IPPW of $\tilde{U}$	-0.03 ± 0.04	0.29	$(-0.11, 0.02)$	2.1
	Raking	-0.01 ± 0.02	0.50	$(-0.05, 0.02)$	3.1
	Calibration	-0.01 ± 0.02	0.50	$(-0.05, 0.02)$	3.1
	Ent. Balancing	-0.01 ± 0.02	0.50	$(-0.05, 0.02)$	3.1
All of Us	**UB (true U^*)**	0.04 ± 0.04	0.97	$(-0.01, 0.12)$	28.9
	UB (heuristic U^*)	0.06 ± 0.05	0.97	$(-0.01, 0.17)$	26.3
	Naive IPPW of $\tilde{U}$	-0.01 ± 0.01	0.82	$(-0.02, 0.01)$	11.8
	Raking	-0.01 ± 0.02	0.79	$(-0.06, 0.00)$	4.6
	Calibration	-0.01 ± 0.02	0.79	$(-0.04, 0.00)$	3.8
	Ent. Balancing	-0.01 ± 0.01	0.74	$(-0.04, 0.00)$	5.3

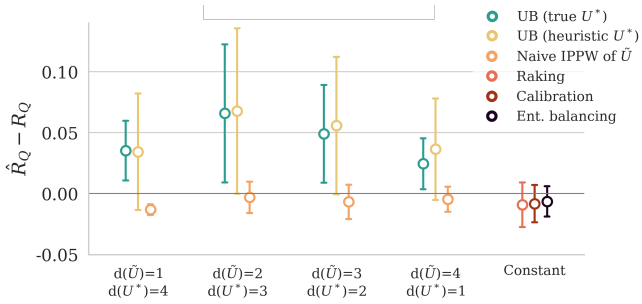


Figure 4: Bound error $\hat{R}_Q - R_Q$ based on the dimension $\mathbf{d}(\tilde{U}) := \dim(\tilde{U})$ of observed selection variables in the All of Us data. While the baselines sometimes underestimate, our method consistently yields a valid upper bound on R_Q.

Table 3: Approximate assumption violation diagnostics on fully synthetic data. 'CS' and 'CI' denote a diagnostic for Common Support and Conditional Independence, respectively. For all diagnostics, a higher score indicates increased violation.

$\lvert\mathcal{D}_Q\rvert$	Selection strength	$\hat{R}_Q - R_Q$ $\mu \pm \sigma$	(CS) KS Test	(CS) Weight d_{eff}	(CI) Propensity Invariance
1000	Low	0.08 ± 0.06	0.18 ± 0.08	50.0 ± 62.1	0.02 ± 0.01
	High	0.07 ± 0.11	0.19 ± 0.10	68.6 ± 89.0	0.03 ± 0.01
100000	Low	0.06 ± 0.06	0.14 ± 0.05	1.4 ± 0.46	0.01 ± 0.00
	High	0.03 ± 0.06	0.15 ± 0.06	3.8 ± 2.85	0.02 ± 0.00

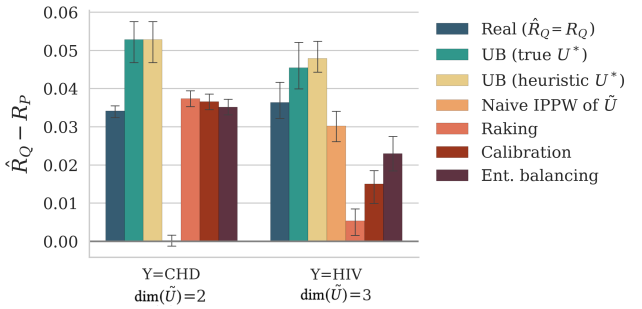
Interpretation of the **KS Test** statistic and p-value should be relative to the user's tolerance for generalization risk. In addition,

because the diagnostic relies on the propensity $p(S = 1 \mid \tilde{U})$, confidence in the diagnostic can be calibrated to the user's confidence in the observed $\tilde{U}$ capturing the drivers of selection. To interpret the **Weight d_{eff}** diagnostic, weights may be deemed unstable when $d_{\text{eff}} \gg 1/\rho$, where $\rho \in [0, 1]$ reflects the user's risk tolerance. For instance, a more conservative user may decide $\rho = 0.1$, i.e., common support is approximately satisfied if $d_{\text{eff}} \ll 10$. Finally, interpreting the **Propensity Invariance** diagnostic similarly depends on the user's confidence in $\tilde{U}$ and the predicted $\hat{U}^*$. Note that while the common support diagnostics do not depend on proper heuristic identification of U^*, the Propensity Invariance test will fail, as intended, under imperfect identification of the set of selection variables $\hat{U} := (\tilde{U}, \hat{U}^*)$. For instance, given one is confident in $\tilde{U}$ (i.e., has determined $\tilde{U}$ are reasonable drivers of selection using domain knowledge), then the failure of the diagnostic indicates an unreliable bound estimate, potentially due to low sample size or inaccurate heuristic identification of U^*. Furthermore, because the diagnostic calculates conditional independence for each variable $V \in (X, Y)$ before averaging across variables, users can examine and potentially reconsider using covariates with high individual diagnostic scores.

5.2.2 Validating Our Bound Estimate. In Figure 6, we show the predicted generalization gap of real-world selection bias in MIMIC-IV relative to the broader All of Us population. In the two leftmost tasks where R_Q is known, we observe that despite the complexity of real-world distribution shifts, our method closely matches the true generalization performance and substantially outperforms baseline approaches.

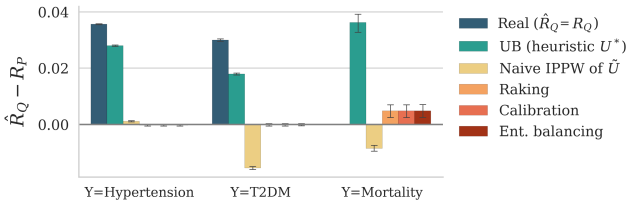
However, in most real-world scenarios, no such target reference exists. For instance, in the MIMIC-IV experiment of predicting Y = Mortality, the true R_Q and variables U are unknown given limited target distribution availability (Constraints 1 - 3). To validate our proposed generalization bound $\hat{R}_Q$ in practice, users may incorporate the assumption diagnostics, as discussed above in Section 5.2.1. In addition, external domain knowledge – such as prior literature, known causal structures, or empirical studies on similar generalization settings – can indicate if a generalization gap is expected. In our setting, for example, prior studies have shown that single-site hospital data often generalize poorly for mortality estimates [44, 72], and Berkson's bias [8, 32] in causal inference implies that hospital-based datasets (e.g., MIMIC-IV) tend to oversample sicker patients relative to the general population. These findings suggest that mortality prediction using MIMIC-IV data may be non-generalizable, thus lending credulity to the nonzero gap predicted by our method.

Together with the above considerations, our proposed bound can help guide deployment decisions. When the assumption violation diagnostics indicate the underlying assumptions are satisfied and the estimated bound $\hat{R}_Q$ is small, practitioners may proceed with model deployment with greater confidence. Conversely, when the bound is large and prior evidence suggests generalization risk, deployment should be delayed in favor of remediation strategies, such as collecting more representative data or applying model-based debiasing techniques.



	Y = Coronary heart disease (CHD)	Y = Human immunodeficiency virus (HIV)
Covariates $\tilde{X}$	Age, BMI, vitals, lifestyle features	Age, BMI, vitals, lifestyle features
Known $\tilde{U}$	Income (+), confidence in medical system (+)	Age (+), has housing (+), education level (+)
Unknown U^*	Employment status (+), general health (-)	Employment status (+)

Figure 5: Generalization gap $\hat{R}_Q - R_P$ (top) and the corresponding prediction task details (bottom) for two specific tasks in the All of Us data. (+) and (-) denotes over- and under-sampling for that selection variable, respectively, guided by EHR-specific selection bias. While baselines risk underestimating generalization performance, our method produces a valid upper bound.



	Y = Hypertension	Y = Type 2 diabetes mellitus (T2DM)	Y = Mortality*
Covariates $\tilde{X}$	Diastolic blood pressure, substance abuse disorder, T2DM, Alzheimer's disease, gender	Marital status, race, primary language, anxiety disorder	Weight, T2DM, endometriosis, hypertension, platelet counts*, insurance type, bicarbonate levels*, systolic blood pressure
Proposed $\tilde{U}$	Age, English as primary language	Age, English as primary language, insurance type	Age, insurance type
Predicted U^*	Comorbidities, diastolic and systolic blood pressure, insurance type, race, marital status	Comorbidities, age, diastolic and systolic blood pressure, insurance type, race, marital status	Comorbidities, age, diastolic blood pressure, insurance type, race, marital status

Figure 6: Real-world generalization gap $\hat{R}_Q - R_P$ (top) and the corresponding prediction task details (bottom) for three tasks using the biased MIMIC-IV data, relative to the target All of Us data. '*'indicates that variable is not observed in All of Us. Note for the task Y=Mortality, the real gap $R_Q - R_P$ is unknown and thus not plotted.

6 Conclusion

In this work, we propose a practical method for estimating the upper bound on the worst-case performance of prediction models under selection bias. Our work has several limitations. First, we focus on

low-dimensional categorical data given its prevalence in medical settings and straightforward density estimation. Extensions to high-dimensional, time-series, and mixed-type data – as we explore in Appendix C – warrant further research. Second, selection bias is closely tied to historical causes of discrimination and underrepresentation. Future work into model generalization should examine subgroup-specific bounds or, even better, root-cause remedies such as deliberate data collection of historically underrepresented populations [9]. Third, we recommend exploring strategies to reduce the high variance of our bound, such as density-free approaches. Fourth, while we limited our study to relatively simple prediction models f_P, it is also important to understand how selection bias affects more complex models, such as LLMs [34]. Finally, general-purpose guidelines for which type of prediction tasks and data structures are most affected by selection bias, as has been done in causal inference [3, 42, 45, 57, 73], would be highly valuable.

Our work makes several important contributions. Unlike existing methods, our method operates under the realistic assumption of limited target data availability. Under these pragmatic data constraints, we propose a novel upper bound on worst-case generalization error and an estimation method using a moment-matching heuristic. We demonstrate through extensive experiments on both synthetic and real-world medical datasets that our method recovers tight and valid bounds even under modest assumption violations. Our experiments on MIMIC-IV highlight the ability of our method to identify real-world cases of selection bias that could lead to generalizability harms if left unaddressed. To this end, we also provide diagnostics for assessing assumption violations, guidance on interpreting the bound, and user-friendly code for running our method. While our work focuses on applications to medical settings, particularly EHR-specific selection bias, the proposed method is applicable to any setting of selection bias. For instance, in Appendix C we apply our bound to detect potential selection bias in political survey data. We hope our work will serve as a practical tool for researchers and practitioners to audit for selection bias, enabling more trustworthy and generalizable prediction models.

Acknowledgments

We thank Vasilis Syrgkanis, Panagiotis Stanitsas, Dominik Rothenhäusler, Roshni Sahoo, and Sanmi Koyejo for their comments and insights.

References

[1] David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. 2021. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*. PMLR, 66–75.

[2] Amar Ahmad, Yvonne Vallès, and Youssef Idaghdour. 2025. Bias in AI systems: integrating formal and socio-technical approaches. *Frontiers in Big Data* 8 (2025).

[3] Elias Bareinboim, Jin Tian, and Judea Pearl. 2014. Recovering from selection bias in causal and statistical inference. In *Probabilistic and causal inference: The works of Judea Pearl*. 433–450.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1 (2010), 151–175.

[5] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems* 19 (2006).

[6] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59, 2 (2013), 341–357.

[7] Richard A Berk. 1983. An introduction to sample selection bias in sociological data. *American sociological review* (1983), 386–398.

[8] Joseph Berkson. 1946. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* 2, 3 (1946), 47–53.

[9] Kirsten Bibbins-Domingo and Alex Helman. 2022. Improving representation in clinical trials and research. *Washington DC: National Academies of Sciences, Engineering, and MedicinePolicy and Global Affairs* (2022).

[10] Kirsten Bibbins-Domingo, Alex Helman, National Academies of Sciences Engineering, Medicine, et al. 2022. Why diverse representation in clinical research matters and the current state of representation within the clinical research ecosystem. *Improving representation in clinical trials and research: Building research equity for women and underrepresented groups* (2022), 23–46.

[11] Carl Bonander, Anton Nilsson, Jonas Björk, Göran ML Bergström, and Ulf Strömberg. 2019. Participation weighting based on sociodemographic register data improved external validity in a population-based cohort study. *Journal of clinical epidemiology* 108 (2019), 54–63.

[12] Andrew D Boyd, Rosa Gonzalez-Guarda, Katharine Lawrence, Crystal L Patil, Miriam O Ezenwa, Emily C O'Brien, Hyung Paek, Jordan M Braciszewski, Oluwaseun Adeyemi, Allison M Cuthel, et al. 2023. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. *Journal of the American Medical Informatics Association* 30, 9 (2023), 1561–1566.

[13] Valerie Bradley and Thomas E Nichols. 2022. Addressing selection bias in the UK Biobank neurological imaging cohort. *MedRxiv* (2022), 2022–01.

[14] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[15] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. 2020. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*. PMLR, 1887–1898.

[16] Rumi Chunara, Jessica Gjonaj, Eileen Immaculate, Iris Wanga, James Alaro, Lori AJ Scott-Sheldon, Judith Mangeni, Ann Mwangi, Rajesh Vedanthan, and Joseph Hogan. 2024. Social determinants of health: the need for data science methods and capacity. *The Lancet Digital Health* 6, 4 (2024), e235–e237.

[17] Andrew Copas, Sarah Burkill, Fred Conrad, Mick P Couper, and Bob Erens. 2020. An evaluation of whether propensity score adjustment can remove the self-selection bias inherent to web panel surveys addressing sensitive health behaviours. *BMC medical research methodology* 20, 1 (2020), 251.

[18] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. 2010. Learning bounds for importance weighting. *Advances in neural information processing systems* 23 (2010).

[19] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. 2016. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* 39, 9 (2016), 1853–1865.

[20] Antoine de Mathelin, Francois Deheeger, Mathilde Mougeot, and Nicolas Vayatis. 2022. Fast and accurate importance weighting for correcting sample bias. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 659–674.

[21] Joshua C. Denny, Joni L Rutter, David B Goldstein, Anthony Philippakis, Jordan W Smoller, Gwynne Jenkins, and Eric Dishman. 2019. The "All of Us" research program. *New England Journal of Medicine* 381, 7 (2019), 668–676.

[22] Jean-Claude Deville and Carl-Erik Särndal. 1992. Calibration estimators in survey sampling. *Journal of the American statistical Association* 87, 418 (1992), 376–382.

[23] Peng Ding and Tyler J VanderWeele. 2016. Sensitivity analysis without assumptions. *Epidemiology* 27, 3 (2016), 368–377.

[24] Jonas H Ellenberg. 1994. Selection bias in observational and experimental studies. *Statistics in medicine* 13, 5-7 (1994), 557–567.

[25] Michael R Elliott and Richard Valliant. 2017. Inference for nonprobability samples. (2017).

[26] Jennifer Lalitha Flaubert, Suzanne Le Menestrel, David R Williams, and Mary K Wakefield. 2021. Social determinants of health and health equity. In *The Future of Nursing 2020-2030: Charting a path to achieve health equity*. National Academies Press (US).

[27] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. 2017. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology* 186, 9 (2017), 1026–1034.

[28] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. arXiv:1505.07818 [stat.ML] https://arxiv.org/abs/1505.07818

[29] Salvatore Giorgi, Veronica E Lynn, Keshav Gupta, Farhan Ahmed, Sandra Matz, Lyle H Ungar, and H Andrew Schwartz. 2022. Correcting sociodemographic selection biases for population prediction from social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 228–240.

[30] Lea Goetz, Nabeel Seedat, Robert Vandersluis, and Mihaela van der Schaar. 2024. Generalization—a key challenge for responsible AI in patient-facing clinical applications. *NPJ Digital Medicine* 7, 1 (2024), 126.

[31] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation [Online]* 101, 23 (2000), e215–e220.

[32] Benjamin A Goldstein, Nrupen A Bhavsar, Matthew Phelan, and Michael J Pencina. 2016. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American journal of epidemiology* 184, 11 (2016), 847–855.

[33] Chris Graham, Jenny King, Clare Lerway, and Alan J Poots. 2025. All the voices we cannot hear: a taxonomy of why some populations' experiences are missing from health and care quality evidence and the Toolkit for Assessing Under Representation in User Surveys (TAURUS). *BMJ open* 15, 2 (2025).

[34] Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915* (2024).

[35] Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis* 20, 1 (2012), 25–46.

[36] Larry Han. 2025. Addressing Distribution Shift for Robust and Trustworthy Prediction and Causal Inference in Clinical AI Settings. *JAMA Network Open* 8, 6 (2025).

[37] Sebastien Haneuse. 2016. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Medical care* 54, 4 (2016), e23–e29.

[38] James Heckman. 1990. Varieties of selection bias. *The American Economic Review* 80, 2 (1990), 313–318.

[39] James J Heckman. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society* (1979), 153–161.

[40] James J Heckman, Hidehiko Ichimura, Jeffrey A Smith, and Petra E Todd. 1998. Characterizing selection bias using experimental data.

[41] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. 2004. A structural approach to selection bias. *Epidemiology* 15, 5 (2004), 615–625.

[42] Miguel A Hernán and James M Robins. 2020. Causal inference: What If.

[43] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. 2006. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems* 19 (2006).

[44] Yousif M Hydoub, Andrew P Walker, Robert W Kirchoff, Hossam M Alzu'bi, Patricia Y Chipi, Danielle J Gerberi, M Caroline Burton, M Hassan Murad, and Sagar B Dugani. 2023. Risk Prediction Models for Hospital Mortality in General Medical Patients: A Systematic Review. *American journal of medicine open* 10 (2023), 100044.

[45] Claire Infante-Rivard and Alexandre Cusson. 2018. Reflection on modern methods: selection bias—a review of recent developments. *International journal of epidemiology* 47, 5 (2018), 1714–1722.

[46] Alistair Johnson, Luca Bulgarelli, Tom Pollard, Benjamin Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). doi:10.13026/kpb9-mt58 RRID:SCR_007345.

[47] Alistair E. W. Johnson, Luca Bulgarelli, Li-wei Shen, and et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* 10, 1 (2023). doi:10.1038/s41597-022-01899-x

[48] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. 2009. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research* 10 (2009), 1391–1445.

[49] Justin B Kaye, Lauren E Schultz, Heidi E Steiner, Rick A Kittles, Larisa H Cavallari, and Jason H Karnes. 2017. Warfarin pharmacogenomics in diverse populations. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 37, 9 (2017), 1150–1163.

[50] Masanari Kimura and Hideitsu Hino. 2024. A short survey on importance weighting for machine learning. *arXiv preprint arXiv:2403.10175* (2024).

[51] Leslie Kish. 1992. Weighting for unequal Pi. *Journal of Official Statistics* 8, 2 (1992), 183.

[52] Wouter M Kouw and Marco Loog. 2019. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence* 43, 3 (2019), 766–785.

[53] Ritoban Kundu, Xu Shi, Jean Morrison, Jessica Barrett, and Bhramar Mukherjee. 2024. A framework for understanding selection bias in real-world healthcare data. *Journal of the Royal Statistical Society Series A: Statistics in Society* 187, 3 (2024), 606–635.

[54] Catherine R Lesko, Ashley L Buchanan, Daniel Westreich, Jessie K Edwards, Michael G Hudgens, and Stephen R Cole. 2017. Generalizing study results: a potential outcomes perspective. *Epidemiology* 28, 4 (2017), 553–561.

[55] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. John Wiley & Sons.

[56] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. 2013. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks* 43 (2013), 72–83.

[57] Haidong Lu, Stephen R Cole, Chanelle J Howe, and Daniel Westreich. 2022. Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology* 33, 5 (2022), 699–706.

[58] Patrick G Lyons, Mackenzie R Hofford, Sean C Yu, Andrew P Michelson, Philip RO Payne, Catherine L Hough, and Karandeep Singh. 2023. Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. *JAMA internal medicine* 183, 6 (2023), 611–612.

[59] Charles F Manski. 1989. Anatomy of the selection problem. *Journal of Human resources* (1989), 343–360.

[60] Charles F Manski. 2003. *Partial identification of probability distributions*. Springer.

[61] Louise AC Millard, Alba Fernández-Sanlés, Alice R Carter, Rachael A Hughes, Kate Tilling, Tim P Morris, Daniel Major-Smith, Gareth J Griffith, Gemma L Clayton, Emily Kawabata, et al. 2023. Exploring the impact of selection bias in observational studies of COVID-19: a simulation study. *International Journal of Epidemiology* 52, 1 (2023), 44–57.

[62] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* 22, 57 (2021), 1–64.

[63] Oriel Perets, Emanuela Stagno, Eyal Ben Yehuda, Megan McNichol, Leo Anthony Celi, Nadav Rappoport, and Matilda Dorotic. 2024. Inherent bias in electronic health records: a scoping review of sources of bias. *ACM Transactions on Intelligent Systems and Technology* (2024).

[64] James M Robins and Dianne M Finkelstein. 2000. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 56, 3 (2000), 779–788.

[65] Paul R Rosenbaum. 2005. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science* 4 (2005), 1809–1814.

[66] Roshni Sahoo, Lihua Lei, and Stefan Wager. 2022. Learning from a biased sample. *arXiv preprint arXiv:2209.01754* (2022).

[67] Maxwell Salvatore, Ritoban Kundu, Xu Shi, Christopher R Friese, Seunggeun Lee, Lars G Fritsche, Alison M Mondul, David Hanauer, Celeste Leigh Pearce, and Bhramar Mukherjee. 2024. To weight or not to weight? The effect of selection bias in 3 large electronic health record-linked biobanks and recommendations for practice. *Journal of the American Medical Informatics Association* 31, 7 (2024), 1479–1492.

[68] Thomas Saphner, Andy Marek, Jennifer K Homa, Lisa Robinson, and Neha Glandt. 2021. Clinical trial participation assessed by age, sex, race, ethnicity, and socioeconomic status. *Contemporary clinical trials* 103 (2021), 106315.

[69] Tabea Schoeler, Doug Speed, Eleonora Porcu, Nicola Pirastu, Jean-Baptiste Pingault, and Zoltán Kutalik. 2023. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour* 7, 7 (2023), 1216–1227.

[70] Alexander Shapiro. 2017. Distributionally robust stochastic programming. *SIAM Journal on Optimization* 27, 4 (2017), 2258–2275.

[71] Sérgio Henrique Almeida da Silva Junior, Simone M Santos, Cláudia Medina Coeli, and Marilia Sá Carvalho. 2015. Assessment of participation bias in cohort studies: systematic review and meta-regression analysis. *Cadernos de saude publica* 31 (2015), 2259–2274.

[72] Harvineet Singh, Vishwali Mhasawade, and Rumi Chunara. 2022. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS digital health* 1, 4 (2022), e0000023.

[73] Louisa H Smith. 2020. Selection mechanisms and their consequences: understanding and addressing selection bias. *Current Epidemiology Reports* 7, 4 (2020), 179–189.

[74] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems* 20 (2007).

[75] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. 2012. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics* 64, 5 (2012), 1009–1044.

[76] BaoLuo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric J Tchetgen Tchetgen. 2018. Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica* 28, 4 (2018), 1965.

[77] Eric J Tchetgen Tchetgen and Kathleen E Wirth. 2017. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* 73, 4 (2017), 1123–1131.

[78] Giovanni Tripepi, Kitty J Jager, Friedo W Dekker, and Carmine Zoccali. 2010. Selection bias and information bias in clinical research. *Nephron Clinical Practice* 115, 2 (2010), c94–c99.

[79] Jenny Wu Tucker. 2010. Selection bias and econometric remedies in accounting and finance research. *Journal of Accounting Literature* 29 (2010), 31–57.

[80] Sjoerd van Alten, Benjamin W Domingue, Jessica Faul, Titus Galama, and Andries T Marees. 2024. Reweighting UK Biobank corrects for pervasive selection bias due to volunteering. *International journal of epidemiology* 53, 3 (2024), dyae054.

[81] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. 2019. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042* (2019).

[82] Christopher Winship and Robert D Mare. 1992. Models for sample selection bias. *Annual review of sociology* 18, 1 (1992), 327–350.

[83] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, et al. 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine* 181, 8 (2021), 1065–1070.

[84] Changbao Wu. 2003. Optimal calibration estimators in survey sampling. *Biometrika* 90, 4 (2003), 937–951.

[85] Bianca Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*. 114.

[86] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. 2019. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J. R. Stat. Soc. Series B Stat. Methodol.* 81, 4 (Sept. 2019), 735–761.

A Proofs

A.1 Upper Bound

Proof. Under Assumption 2 of common support, and by the invariance of the loss term to $\tilde{U}$, we can rewrite R_Q as Let us denote

$$\begin{aligned} R_Q &= \mathbb{E}_P\left[\frac{p(\tilde{X},Y)}{p(\tilde{X},Y \mid S=1)}\cdot \ell(f_P(\tilde{X}),Y)\right] \\ &= \mathbb{E}_P\left[\frac{p(\tilde{X},Y,\tilde{U})}{p(\tilde{X},Y,\tilde{U} \mid S=1)}\cdot \ell(f_P(\tilde{X}),Y)\right] \\ &= \mathbb{E}_P\left[\frac{p(\tilde{X},Y \mid \tilde{U})}{p(\tilde{X},Y \mid \tilde{U},S=1)}\cdot w(\tilde{U})\cdot \ell(f_P(\tilde{X}),Y)\right] \end{aligned}$$

where $w(\tilde{U}) \coloneqq \frac{p(\tilde{U})}{p(\tilde{U}|S=1)}$. By Assumption 1 of conditional independence, we know that $p(\tilde{X},Y \mid U) = p(\tilde{X},Y \mid U,S=1)$ and thus

$$\begin{aligned} p(\tilde{X},Y \mid \tilde{U}) &= \int_{u^*\in\mathcal{U}^*} p(u^*|\tilde{U})p(\tilde{X},Y|\tilde{U},u^*)du^* \\ &= \mathbb{E}_{Q_{U^*|\tilde{U}}}\left[p(\tilde{X},Y \mid \tilde{U},U^*)\right] \\ &= \mathbb{E}_{Q_{U^*|\tilde{U}}}\left[p(\tilde{X},Y \mid \tilde{U},U^*,S=1)\right] \end{aligned}$$

If we substitute this into the above equation for R_Q, we have

$$R_Q = \mathbb{E}_P\left[\frac{\mathbb{E}_{Q_{U^*|\tilde{U}}}\left[p(\tilde{X},Y \mid \tilde{U},U^*,S=1)\right]}{p(\tilde{X},Y \mid \tilde{U},S=1)}\cdot w(\tilde{U})\cdot \ell(f_P(\tilde{X}),Y)\right]$$

By definition of expectation and non-negativity of density functions, there exists some $\varepsilon_{\tilde{X},Y,\tilde{U}} \geq 0$ such that

$$\max_{u^*\in\mathcal{U}^*} p(\tilde{X},Y \mid \tilde{U},u^*,S=1) = \mathbb{E}_{Q_{U^*|\tilde{U}}}\left[p(\tilde{X},Y \mid \tilde{U},u^*,S=1)\right] + \varepsilon_{\tilde{X},Y,\tilde{U}}$$

Thus

$$R_Q = \mathbb{E}_P\left[\frac{\max_{u^*\in\mathcal{U}^*} p(\tilde{X}, Y \mid \tilde{U}, u^*, S=1) - \varepsilon_{\tilde{X},Y,\tilde{U}}}{p(\tilde{X}, Y \mid \tilde{U}, S=1)} \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right]$$

$$= \mathbb{E}_P\left[\frac{\max_{u^*\in\mathcal{U}^*} p(\tilde{X}, Y \mid \tilde{U}, u^*, S=1)}{p(\tilde{X}, Y \mid \tilde{U}, S=1)} \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right] - \mathbb{E}_P\left[\frac{\varepsilon_{\tilde{X},Y,\tilde{U}}}{p(\tilde{X}, Y \mid \tilde{U}, S=1)} \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right]$$

$$\leq \mathbb{E}_P\left[\frac{\max_{u^*\in\mathcal{U}^*} p(\tilde{X}, Y \mid \tilde{U}, u^*, S=1)}{p(\tilde{X}, Y \mid \tilde{U}, S=1)} \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right]$$

since $p(\tilde{X}, Y \mid \tilde{U}, S=1) \geq 0$ and $\ell(f_P(\tilde{X}), Y) \geq 0$ from Assumption 3. □

A.2 Bound Decomposition

In Section 4.3, we proposed that the bound error could be decomposed as follows:

$$\hat{R}_Q - R_Q = \Delta_{\text{TBE}} + \Delta_{\text{CI}} + \Delta_{\text{CS}}$$

where Δ_{CS} is the error contribution from violating the necessary assumption of Common Support between P and Q; and Δ_{CI} is the error contribution from violating the necessary assumption of Conditional Independence given U. In theory (i.e., under infinite samples), these two assumptions hold and $\Delta_{\text{CI}} = \Delta_{\text{cs}} = 0$, meaning $\hat{R}_Q - R_Q = \Delta_{\text{TBE}}$, where Δ_{TBE} is the Theoretical Bound Error when all assumptions hold.

Here, we formally define each of the factors Δ_{TBE}, Δ_{CI}, and Δ_{CS}. Notice that $\hat{R}_Q - R_Q$ can be written as the following telescoping sum:

$$\begin{aligned}
\hat{R}_Q - R_Q &= \mathbb{E}_P\left[\phi(\tilde{X}, Y, \mathcal{U}^*) \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right] \\
&\quad - \mathbb{E}_P\left[\frac{\mathbb{E}_{Q_{U^*|\tilde{U}}}[p(\tilde{X}, Y \mid \tilde{U}, U^*, S=1)]}{p(\tilde{X}, Y \mid \tilde{U}, S=1)} \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right] && (\Delta_{\text{TBE}}) \\
&\quad + \mathbb{E}_P\left[\frac{\mathbb{E}_{Q_{U^*|\tilde{U}}}[p(\tilde{X}, Y \mid \tilde{U}, U^*, S=1)]}{p(\tilde{X}, Y \mid \tilde{U}, S=1)} \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right] \\
&\quad - \mathbb{E}_P\left[\frac{p(\tilde{X}, Y \mid \tilde{U})}{p(\tilde{X}, Y \mid \tilde{U}, S=1)} \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right] && (\Delta_{\text{CI}}) \\
&\quad + \mathbb{E}_P\left[\frac{p(\tilde{X}, Y \mid \tilde{U})}{p(\tilde{X}, Y \mid \tilde{U}, S=1)} \cdot w(\tilde{U}) \cdot \ell(f_P(\tilde{X}), Y)\right] \\
&\quad - \mathbb{E}_Q\left[\ell(f_P(\tilde{X}), Y)\right] && (\Delta_{\text{CS}})
\end{aligned}$$

Defining the first difference as Δ_{TBE}, the second difference as Δ_{CI}, and the third difference as Δ_{CS}, we have the proposed decomposition. These definitions clarify how and why Δ_{CI} and Δ_{CS} reveal assumption violations. First, Δ_{CI} tests if conditional independence is violated by isolating the effect of changing $E_{Q_{U^*|\tilde{U}}}[p(\tilde{X}, Y \mid \tilde{U}, U^*, S=1)]$ to $E_{Q_{U^*|\tilde{U}}}[p(\tilde{X}, Y \mid \tilde{U}, U^*)] = p(\tilde{X}, Y \mid \tilde{U})$, where the two terms should be equal if conditional independence given U holds. Second, Δ_{CS} measures violation of the assumption of common support between P and Q by characterizing the difference between the density ratio re-weighted expectation over P and the unweighted expectation over Q. These expressions are equal provided P and Q share common support.

B Heuristic U^* Selection

Recall the heuristic solves for

$$\sum_{i:S^{(i)}=1} \frac{m(C^{(i)}, \tilde{U}^{(i)})}{g(\tilde{U}^{(i)}\omega + C^{(i)}\gamma)} = \mathbb{E}_{\mathcal{D}_Q}[m(\tilde{U}, C)] \tag{B.1}$$

which computes point estimates for the γ_j coefficients. To derive confidence intervals for the estimated $\hat{\gamma}_j$, we take a bootstrapping approach. Specifically, we resample the dataset $\mathcal{D}_P$ with replacement B times. We then solve the above equation for each dataset sample, yielding the bootstrapped distribution $\hat{\gamma}_j^{(1)}, \hat{\gamma}_j^{(2)}, \ldots, \hat{\gamma}_j^{(B)}$ for each j. We construct a 1-α level confidence interval as $(\hat{\gamma}_{j_{(\alpha/2)}}, \hat{\gamma}_{j_{(1-\alpha/2)}})$, where $\hat{\gamma}_{j_{(\alpha/2)}}$ denotes the $\alpha/2$ percentile of the bootstrapped coefficients and $\hat{\gamma}_{j_{(1-\alpha/2)}}$ denotes the $1-\alpha/2$ percentile. In our experiments, we set $B = 1000$ iterations and $\alpha = 0.005$. We set α to be smaller than typically used for bootstrapped confidence intervals because the resulting conservative confidence intervals for $\hat{\gamma}_j$ help limit the number of false positives (variables mistakenly nominated as U^*).

C Supplementary Materials

We present the full appendix at https://arxiv.org/abs/2606.00563.