Beyond the Training Distribution: Evaluating Predictions Under Distribution Shift & Selection Bias

Annie Ulichney
Department of Statistics
University of California, Berkeley
annie_ulichney@berkeley.edu

Amanda Coston
Department of Statistics
University of California, Berkeley
acoston@berkeley.edu

Abstract

Understanding how a prediction model will perform in a new environment before deployment is essential to preventing harm when algorithms inform decision-making. Two common sources of model performance degradation are (i) *covariate shift*, where the target covariate distribution differs from the source, and (ii) *selective labels*, where the observability of outcomes depends on historical decisions. We study *pre-deployment* model evaluation under the joint presence of covariate shift and labeling of outcomes selectively based on observed features. In particular, we present a double machine learning procedure for estimating the target risk of an arbitrary black-box prediction model under a general loss function. We show identification of this estimand under standard assumptions and derive a bias-corrected estimator based on the influence function of the target risk. Finally, we evaluate our estimator through experiments using the eICU electronic health records database, showing that it tracks the true target risk more accurately than methods that address either selective labels or covariate shift alone, as well as baselines that combine standard plug-in approaches.

1 Introduction

Machine learning models that achieve strong performance on a held-out test set routinely degrade when deployed in new environments. This degradation is well-documented across high-stakes settings: medical imaging models underdiagnose underrepresented demographic groups [60, 61, 31, 64, 45], clinical risk scores exhibit performance drops across hospitals and demographic groups [40, 63, 19], and natural language processing systems underperform on linguistic minorities and dialects [42, 56, 39, 67]. These failures are often attributed to *distribution shift* [48] where the training and deployment populations differ systematically. However, in many decision-making settings, distribution shift is only part of the problem: labels are often observed only for individuals selected by a prior human, institutional, or algorithmic decision, which can make the evaluation data an inaccurate reflection of the deployment population.

We study the intersection of these evaluation challenges. The first is *covariate shift*, where the input distribution of features changes while the relationship between features and outcomes remains stable [62]. Under covariate shift, considering a model's average performance on a held-out test set can obscure deployment failures; if a model's error is higher in feature regions that are more common in the deployment setting, then its target risk may be much worse than indicated by the test error [27].

The second is the *selective labels problem* [30]. In domains such as lending, hiring, and healthcare, outcomes are observed only for individuals selected by a decision policy: approved borrowers, hired applicants, or patients who receive diagnostic medical tests. Such a mechanism induces *selection bias* because the observed labels are drawn from a decision-dependent subset of the population [26]. As a result, evaluating performance only on labeled samples from the training data can be misleading.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The challenges of *covariate shift* and *selective labels* often coexist in precisely the high-stakes settings where reliable pre-deployment evaluation matters most. A salient example is the use of *Clinical Decision-making Instruments (CDIs)*, predictive models that use patient demographics, symptoms, and test results to aid diagnosis, triage, and treatment decisions. Many CDIs are developed using data from patient populations that differ systematically from the populations in which they are ultimately deployed– for example, models trained on predominantly white patient populations are often later used in demographically diverse populations [41]. This practice induces a covariate shift. At the same time, many clinically relevant outcomes are observed only for patients who receive follow-up testing, monitoring, or treatment, and those decisions are shaped by existing clinical policies or the CDI itself. Thus, to accurately determine whether a CDI is suitable for deployment in a new hospital, a practitioner must address not only that the target population may differ systematically from the population in which the model was evaluated but also that outcomes are observed only for patients selected by prior clinical decisions.

Contributions: This work studies *pre-deployment evaluation* of an arbitrary black-box prediction model under the joint presence of covariate shift and selective labels. Our first contribution is to formalize the problem of evaluating black-box predictors under simultaneous covariate shift and selective labeling. We frame evaluation around a *target risk* functional that captures the risk the model would incur in the target population under a fixed general loss function. This estimand answers the practically important question of how a model will perform in its actual deployment population, even when that population looks different from the training data and when outcome labels are systematically missing due to historical decisions.

Second, we establish identification of the target risk using observable quantities in the data under standard assumptions. In particular, we establish identification supposing that the evaluator observes (i) selectively labeled data from the source distribution (used to train the model) where selection is on observables and (ii) unlabeled samples from the target deployment environment.

Third, we derive the efficient influence function of the target risk estimand and use it to construct a double machine learning (DML) estimator that is asymptotically normal and semiparametrically efficient. We empirically validate our framework through controlled synthetic and semi-synthetic experiments in healthcare settings; we show that our estimator tracks deployment performance more reliably than standard plug-in approaches or methods designed for evaluation under *either* covariate shift or selective labels in isolation. Finally, we illustrate our estimator on several real-world clinical outcomes where both label observability and outcomes are taken from observed eICU data. We find that jointly accounting for covariate shift and selective labels can yield meaningfully different risk estimates from methods addressing either mechanism individually in real-world settings.

1.1 Related Work

Our work lies at the intersection of three literatures: evaluation under distribution shift, learning and evaluation under selective labels, and semiparametric doubly robust estimation. We summarize the closest work here and defer broader discussion to Appendix B.

Evaluation under covariate shift: A large literature studies covariate shift in which the marginal covariate distribution differs between the training and deployment environments while the conditional distribution of outcomes given features remains invariant [62, 48]. Classical approaches use importance weighting or density ratio estimation [62, 66, 25]. More directly related to our work are recent *doubly robust* and *semiparametric* estimators for target-domain risk [50, 22, 10, 38, 16]. Also closely related are broader recent methods for estimating or diagnosing performance degradation under covariate shift [8, 5, 3]. These works share our goal of estimating target performance before deployment, but they typically assume that outcome labels are fully observed in the source evaluation data. In contrast, our setting allows source labels to be only partially observed through a covariate-dependent selection mechanism.

Selective labels and counterfactual evaluation: A separate line of work studies this *selective labels* problem [30, 26] where outcome labels are observed only for individuals selected by a historical policy or a human decision-maker. This problem arises in settings such as credit scoring, criminal justice, and clinical medicine, where outcomes are only revealed after approval, release, or testing decisions [40, 12]. The closest work to ours in this literature is counterfactual risk estimation under selective labels, which uses causal inference tools to evaluate a predictive model when labels

are selectively observed [11, 6]. Other related approaches include correcting selection bias using heterogeneity across decision-makers [26, 7] and targeted data acquisition or augmentation for subpopulations underrepresented in the labeled data [13, 28]. Our setting is also connected to questions about feedback loops and *performativity*, where interventions guided by model predictions change the process by which future labeled data are generated [43, 15]. These methods address outcome observability, but generally evaluate performance in the same setting that generated the selectively labeled data. Our setting differs by considering the case where the deployment population itself may also be shifted relative to the source population.

Doubly robust estimation: Methodologically, our estimators build on semiparametric and *doubly robust* approaches for incomplete, biased, or shifted data [53, 57, 2]. These methods combine outcome regression with inverse-probability weighting and use influence function-based corrections to reduce sensitivity to *nuisance function* estimation errors [52, 51, 29, 23]. Such methods have been applied to data missingness, off-policy evaluation, and distribution shift [14, 9, 10, 5, 3, 36, 65]. Our contribution is to use these tools to derive an efficient estimator of the performance of a predictive model under joint covariate shift and selective labels.

2 Problem Setting

We study the problem of evaluating a fixed prediction model under the joint presence of covariate shift and selective labeling. We observe n independent and identically distributed (i.i.d.) draws

$$Z \coloneqq (X, R, R \cdot D, R \cdot D \cdot Y) \tag{1}$$

where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a covariate vector, $R \in \{0, 1\}$ is a domain indicator, and $D \in \{0, 1\}$ indicates whether the outcome of interest is observed. The domain indicator R satisfies $R = 1$ for units from the *source* population and $R = 0$ for units from the *target* population. Due to selective labeling, the latent outcome $Y \in \mathbb{R}$ is observed only when $R = 1$ and $D = 1$.

Let p_S and p_T denote the densities of $X|R = 1$ and $X|R = 0$, respectively. For an integrable random variable Q, we write $\mathbb{E}_S[Q] \coloneqq \mathbb{E}[Q|R = 1], \mathbb{E}_T[Q] \coloneqq \mathbb{E}[Q|R = 0]$. When Q is a function of only X, these reduce to expectations under the covariate distributions P_S and P_T, respectively.

We aim to evaluate how well a fixed prediction model $f : \mathcal{X} \to \mathbb{R}$ predicts Y in the target environment. For a given loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ (e.g., squared loss), our estimand is the *target risk*

$$\psi \coloneqq \mathbb{E}_T \left[\ell \left(f(X), Y \right) \right]. \tag{2}$$

Note that f may be an arbitrary fixed predictor (e.g., a deep network, an ensemble model, an LLM scorer), and ℓ may be an arbitrary fixed loss function. Our procedure treats both as black boxes.

3 Identification and Estimation of the Target Risk

In this section, we identify ψ under the joint presence of covariate shift and selective labels, and we then derive an influence-function-based estimator that accounts for both mechanisms. We use $L \coloneqq \ell(f(X), Y)$ as shorthand notation for the loss under f on observed outcomes. Let $P_X = (1 - \rho)P_S + \rho P_T$ denote the pooled covariate law. Also define the following *nuisance functions*:

$$\rho \coloneqq \mathbb{P}(R = 0), \quad g(X) \coloneqq \mathbb{P}(R = 0 \mid X), \quad \mu(X) \coloneqq \mathbb{E}[\, L \mid X, R = 1, D = 1 \,], \tag{3}$$

where we suppose $\rho \in (0, 1)$. and the pooled distribution and source distribution propensity scores, respectively, as

$$\pi(X) \coloneqq \mathbb{P}(D = 1, R = 1 \mid X), \quad \pi_S(X) \coloneqq \mathbb{P}(D = 1 \mid X, R = 1).$$

Estimation of ψ is complicated both by the fact that Y is unobserved in the target domain ($R = 0$) and by the fact that, in the source domain, Y is selectively observed (when $D = 1, R = 1$). To ensure identifiability, we impose the following assumptions, which are standard in the causal inference and transfer learning literatures.

Assumption 3.1 (Selection on observables in the source)**.** $Y \perp\!\!\!\perp D \mid X, R = 1$.

Assumption 3.2 (Covariate Shift)**.** $P_S(Y \mid X) = P_T(Y \mid X)$.

Assumption 3.3 (Positivity)**.** *There exists $\varepsilon > 0$ such that $\pi_S(X) > \varepsilon$ P_S-almost surely.*

Assumption 3.4 (Bounded likelihood ratio)**.** *There exists $C < \infty$ such that $\frac{dP_T}{dP_S}(x) \leq C \quad \forall x \in \mathcal{X}$.*

Note that Assumptions 3.3 and 3.4 together imply that $1 - g(x)$ and $\pi(x)$ are bounded away from zero (see Appendix A.4.2).

These assumptions identify the target risk from observed data. Intuitively, these conditions require that (i) source label observability is as good as random conditional on X, (ii) the conditional relationship between covariates and outcomes is invariant across the source and target domains, (iii) every covariate profile admits a positive probability of being observed in the source population, and (iv) the target covariate distribution is sufficiently covered by the source distribution. [1]

Proposition 3.5 (Identification of the Target Risk)**.** *Under Assumptions 3.1-3.4 and with nuisance functions π, ρ, g, and μ as defined in* (3)*, the target risk ψ is identifiable from the observed data as*

$$\psi = \mathbb{E}_T\left[\mu(X)\right] = \mathbb{E}_S\left[\frac{p_T(X)}{p_S(X)} \cdot \frac{D}{\pi_S(X)}\, L\right].$$

Proposition 3.5 is the central identification result of the paper. It makes explicit how an evaluator can simultaneously transport across populations *and* correct for a selective labeling process to identify the target risk of a fixed black-box predictor; existing covariate shift identifications presume fully labeled source data, and existing identifications for selectively labeled data presume a single population. This result lays the foundation for our proposed evaluation procedure by connecting the unobservable target risk to quantities that can be learned from observable, pre-deployment data.

The first equality shows that the target risk can be written as an expectation over the *target* covariate distribution of the nuisance function $\mu(X)$, which summarizes the model's conditional expected loss after accounting for selective labels. We note that $\mu(X)$ depends only on covariates, i.e., it does not require observing target outcome labels. This expression demonstrates that pre-deployment evaluation is feasible in our setting; if we could estimate $\mu(\cdot)$ from selectively labeled source data, we could estimate ψ by averaging $\mu(X)$ over the target population.

The second expression provides an alternative identification as an expectation with respect to the source population: the first factor $\frac{p_T(X)}{p_S(X)}$ transports the loss estimate from the source to the target population to account for covariate shift, and the second factor $\frac{D}{\pi_S(X)}$ is an inverse propensity weighted (IPW) correction for selective labels. Together, these equalities formalize the insight that, even when deployment labels are unavailable and source labels are observed only through a selective labeling process, we can still evaluate a model's target risk from pre-deployment data. A proof is provided in Appendix A.1.

3.1 Efficiency Theory for Target Risk Estimation

Next, we characterize the efficient influence function (EIF) of ψ, which is the basis of our estimator and its asymptotic guarantees. In particular, the EIF enables us to construct an estimator that can achieve $\sqrt{n}$ rates while allowing for flexible nuisance estimation. Let $\mathcal{P}$ denote the nonparametric model defined by Assumptions 3.1-3.4. The following result identifies the efficient influence function of the target risk ψ and gives the corresponding von Mises expansion. This result demonstrates the bias-correction step that stabilizes the naïve plug-in estimators.

Proposition 3.6 (Target Risk Influence Function)**.** *For every $\mathbb{P}, \overline{\mathbb{P}} \in \mathcal{P}$, the map $\psi : \mathbb{P} \to \mathbb{R}$ admits the expansion*

$$\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) = \int \varphi(z; \overline{\mathbb{P}})\, d(\overline{\mathbb{P}} - \mathbb{P})(z) + R_2(\mathbb{P}, \overline{\mathbb{P}}), \tag{4}$$

where the efficient influence function is

$$\varphi(Z; \mathbb{P}) = \frac{RD}{\pi(X)} \frac{g(X)}{\rho}(L - \mu(X)) + \frac{1 - R}{\rho}(\mu(X) - \psi(\mathbb{P})) \tag{5}$$

[1] These assumptions are standard, though only some have observable implications; in particular, Assumptions 3.1 and 3.2 are not directly testable from observable data while Assumption 3.3 can be partially tested from observable data [44]. In Appendix F, we empirically examine the sensitivity of our estimators to violations of each assumption. In Appendix C, we discuss real-world implications of these assumptions.

and the remainder is

$$R_2(\mathbb{P}, \overline{\mathbb{P}}) = \left(\frac{\overline{\rho} - \rho}{\overline{\rho}}\right)\left(\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P})\right) + \frac{1}{\overline{\rho}}\int g\left(\frac{\pi}{\overline{\pi}} - 1\right)(\mu - \overline{\mu})\, d\mathbb{P} + \frac{1}{\overline{\rho}}\int (\overline{g} - g)\frac{\pi}{\overline{\pi}}(\mu - \overline{\mu}) d\mathbb{P}.$$

The EIF given in Proposition 3.6 incorporates two complementary signals from the source and target distributions, respectively, in each term. The first term of (5) is a residual correction computed on observed labeled source units ($RD = 1$). Here, $RD/\pi(X)$ is the Horvitz-Thompson (inverse probability) weight for appearing as a labeled source observation in the pooled sample. This factor corrects the selective-labeling mechanism. The additional factor $g(X)/\rho$ transports the correction toward the target covariate law. Indeed, if $p(x) = \rho \cdot p_T(x) + (1 - \rho) \cdot p_S(x)$ denotes the pooled covariate density, then Bayes' rule gives $\frac{g(X)}{\rho} = \frac{p_T(X)}{p(X)}$.

The residual $L - \mu(X)$ centers the labeled source contribution conditional on X, which reduces variance relative to the uncentered representation. The second term is the target domain plug-in component. Since labels (and, therefore, L) are not observed in the target domain ($R = 0$), this term averages the conditional loss regression $\mu(X)$ over unlabeled target covariates. The factor $(1 - R)/\rho$ converts the pooled expectation into an expectation in the target domain, and subtraction of $\psi(\mathbb{P})$ centers the influence function (since $\psi = \mathbb{E}_T[\mu(X)]$). Thus, the EIF combines a transported labeled-source residual correction with a centered target-domain plug-in term.

The proof of Proposition 3.6 applies semiparametric calculus to the first identified representation given in Proposition 3.5. Differentiating this functional yields the two terms of (5). The von Mises remainder is obtained by comparing $\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P})$ to the first-order expansion, then collecting and controlling leftover terms, which are the products of nuisance functions. We then invoke Kennedy [24, Lemma 2] to conclude efficiency. A full proof is provided in Appendix A.2.1.

3.2 A double machine learning estimator of target risk

We use the efficient influence function of ψ (Proposition 3.6) to construct a *double machine learning* (DML) estimator of the target risk [53, 57, 2]. Our construction uses an EIF-based bias correction that makes the DML estimator robust to errors in the nuisance functions μ, π and g. The DML approach combines bias correction with sample splitting (or cross-fitting) to remove the bias that would otherwise emerge due to using the same data for nuisance estimation and evaluation.

Let $\mathcal{S}_n \coloneqq \{Z_i\}_{i=1}^n$ denote an evaluation sample of n i.i.d. draws from P, and let $\widehat{\pi}$, $\widehat{g}$, $\widehat{\mu}$ be nuisance estimators fit on an auxiliary sample $\mathcal{S}'_n$ consisting of n samples and independent of $\mathcal{S}_n$. Motivated by the EIF in (5), we propose the *DML estimator* of ψ as

$$\widehat{\psi} = \frac{1}{n}\frac{1}{\widehat{\rho}}\sum_{i=1}^{n}\left[\frac{R_i D_i}{\widehat{\pi}(X_i)}\widehat{g}(X_i)\left(L_i - \widehat{\mu}(X_i)\right) + (1 - R_i)\widehat{\mu}(X_i)\right] \tag{6}$$

where $\widehat{\rho} = n^{-1}\sum_{i=1}^n \mathbb{1}\{R_i = 0\}$ is the empirical estimator of ρ. The first summand corrects labeled source units ($R_i = 1, D_i = 1$) for the selection and covariate shift mechanisms; the second is a regression-based plug-in estimate for target units ($R_i = 0$).

Theorem 3.7 given below establishes that $\widehat{\psi}$ is $\sqrt{n}$-consistent, asymptotically linear, and semiparametrically efficient under standard regularity conditions and mild convergence-rate requirements for nuisance function estimation. For simplicity, we state the results under single splitting and note that analogous guarantees for cross-fitting follow under analogous foldwise conditions.

Theorem 3.7 (Asymptotic properties of the DML estimator $\widehat{\psi}$)**.** *Let $\widehat{\psi}$ be the estimator in* (6)*, constructed with nuisance estimators $\widehat{\pi}$, $\widehat{g}$, and $\widehat{\mu}$ trained on an auxiliary sample $\mathcal{S}'_n$ independent of $\mathcal{S}_n$. Suppose Assumptions 3.1-3.4 hold. In addition, suppose that (i) the nuisance estimators are $L_2(P)$-consistent, (ii) each of $\widehat{\mu}, \widehat{\pi}$, and $\widehat{g}$ converges at a rate $o_p(n^{-1/4})$ in $L_2(P)$, (iii) the estimated propensity denominators are bounded away from zero with probability tending to one, and (iv) $\mathbb{E}\left[\varphi(Z; P)^2\right] < \infty$. Then, $\widehat{\psi}$ is asymptotically linear with influence function $\varphi(\cdot; P)$:*

$$\widehat{\psi} - \psi(P) = \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i; P) + o_p(n^{-1/2}), \tag{7}$$

and hence

$$\sqrt{n}\left(\widehat{\psi} - \psi\right) \xrightarrow{d} \mathcal{N}\left(0, V_\varphi\right), \quad V_\varphi \coloneqq \mathrm{Var}(\varphi(Z; P)). \tag{8}$$

A more general statement of Theorem 3.7, which formalizes conditions (i)-(iv), is given in Appendix A.4, along with the proof. The main implication of this result is that our proposed estimator supports valid pre-deployment inference, even when the nuisance functions are estimated with flexible machine learning methods. In particular, the conditional loss model μ, the pooled labeled-source propensity π, and the domain propensity g need not be estimated at parametric rates. It suffices that their errors are small enough for the second-order remainder to be negligible; a sufficient condition is $o_p(n^{-1/4})$ convergence in $L_2(P)$ for each nuisance function.

The theorem also highlights why the influence function correction is essential for $\sqrt{n}$ inference in our setting. Naïve plug-in or IPW estimators generally inherit first-order bias from errors in $\widehat{\mu}$, $\widehat{\pi}$, or $\widehat{g}$, making valid inference difficult. By contrast, the DML correction removes first-order sensitivity to nuisance estimation error: after sample splitting, the remaining bias is governed by products of nuisance errors.

The asymptotic linear expansion in (30) also yields a practical uncertainty quantification procedure. Since the leading term is the average of the influence scores $\varphi(Z_i; P)$, the asymptotic variance can be estimated by the empirical variance of the estimated influence scores. This enables the construction of standard Wald-type confidence intervals for the target risk. Therefore, the theorem justifies the use of modern prediction methods to estimate the nuisance components while retaining classical $\sqrt{n}$ inference for the target risk.

4 Experiments

We present two sets of clinical experiments. First, we perform a semi-synthetic experiment with real-world covariates and simulated outcome and label-observability mechanisms. This setting preserves realistic covariate shifts, which are known to be more complex and difficult to address than shifts simulated in controlled, synthetic experiments [27]. Simulating the label and outcome mechanisms enables comparison of each estimator to the *true* target risk. Second, we illustrate the estimator on real-world clinical outcomes, where both label observability and outcomes are taken from observed eICU data. Synthetic experiments are deferred to Appendix D.1.

eICU Data: The eICU Collaborative Research Database [47] includes de-identified individual-level electronic health records from over 200,000 admissions to ICUs across multiple hospitals in the United States. We use admission-level patient demographics, vitals, clinical unit type, and hospital ID.

Real-world Shifts: We leverage the fact that the data include multiple hospitals to capture the dynamics of real-world distribution shifts where a model is trained on a population that differs in demographic makeup from the population on which it is deployed. We identify 3 candidate target hospitals that look systematically different from the general population in age, race, and ethnicity and one baseline source hospital with demographics that are representative of the overall eICU sample. We train a model on the source hospital, then evaluate its performance in each candidate, covariate-shifted target hospital.

In particular, our covariate-shifted hospitals include hospital 443 (which tends to have younger patients and more African American patients), hospital 199 (with a typical age profile and more Caucasian patients), hospital 283 (which skews older with a larger share of patients labeled as unknown or "other" ethnicity). We use hospital 208 as a no-covariate-shift benchmark since it is approximately average across these demographics. We illustrate these covariate shifts in Figure 8.

Semi-Synthetic Design: We generate outcomes from a sparse logistic model in which older patients have a higher outcome risk. We then train a fixed prediction rule f on simulated labels from the source training split and hold this rule fixed throughout. We simulate label observability for source patients. The probability that a source label is observed depends on patient covariates, where the selection mechanism is chosen so that selection strength values increasingly distort label observability away from older patients. We vary feature-dependence in the selection mechanism while holding the overall labeling rate approximately constant. As a result, the experiment varies the degree of feature dependence in the selective labeling mechanism while holding the amount of labeled source data roughly constant. The target hospital outcomes are treated as unobserved by the estimators, which matches the pre-deployment evaluation setting in which the evaluator accesses unlabeled target covariates but not target labels. See Appendix H.3 for details.

Estimators: For each selection strength value, we compare our DML estimator (6) to three baselines. First, a natural benchmark is the *Plug-in Estimator* that reweighs observed labeled source losses by the estimated density ratio $\widehat{w}(x)$ and the inverse propensity weights $1/\widehat{\pi}(x)$ to account for both covariate shift and selective labels. Second, we compare against a recent approach for evaluating models under covariate shift using double machine learning that we denote *Covariate Shift (CS)-only* [38]. Finally, we compare against a *Selective Labels (SL)-only* estimator from Coston et al. [11] that uses double machine learning to address selective labels. The *CS-only* and *SL-only* baselines represent leading approaches for their respective settings: each uses double machine learning to leverage flexible nuisance estimation, but each addresses only one of the two challenges considered here. See Appendix H.1 for more details.

Nuisance Estimation: We estimate three nuisance components using cross-fitting: (i) a domain classifier for covariate shift reweighting, (ii) a label observability propensity for selective labels, and (iii) an outcome regression model for the conditional expected loss.

First, we estimate the domain propensity $g(x)$ for covariate shift reweighting via logistic regression on the pooled source and target covariates. Second, we estimate the source label observability propensity $\pi_S(x)$ via logistic regression. Then, noting that $\pi(X) = (1 - g(X)) \cdot \pi_S(X)$, we estimate the source outcome-labeling propensity via $\widehat{\pi}(x) = (1 - \widehat{g}(x))\widehat{\pi}_S(x)$. Finally, we estimate the conditional loss $\mu(x)$ by noting that for absolute loss with binary outcomes, we can write $\mu(x) = \eta(x)|1 - f(x)| + (1 - \eta(x))|f(x)|$ where $\eta(x) \coloneqq \mathbb{P}(Y = 1|X = x, R = 1, D = 1)$. We use a cross-fitted random forest regression model to estimate $\eta(x)$ with labeled source units, then use this estimate to construct $\widehat{\mu}(x)$.

To evaluate each estimator, we compute a Monte Carlo estimate of the ground truth target risk by simulating an oracle dataset of target samples. For each target covariate vector, we generate an outcome using the same outcome model as the experiment. The ground-truth risk estimate is then computed as the mean absolute error (MAE) between the model prediction and the oracle outcomes. We evaluate each estimator by comparing its estimated target risk to the oracle estimate of ground truth. We report uncertainty via 95% bootstrap confidence intervals. While our proposed DML estimator is an asymptotically linear estimator of the target risk under the conditions of Theorem 3.7, the same need not be true in general for the other estimators. Consequently, we acknowledge that the bootstrap intervals may be inaccurate for such estimators; we nevertheless use the same uncertainty quantification procedure across estimators to provide a consistent comparison.

Semi-Synthetic Results: Fig. 1 reports the bias and RMSE of each estimator relative to the ground-truth target risk as the feature-dependence of the selective labeling mechanism increases on the x-axis. Across the shifted target hospitals, our DML estimator exhibits the least bias. The CS-only estimator is positively biased in all target hospitals, with bias increasing as selection becomes more feature-dependent. As expected, the SL-only estimator is approximately unbiased in the no-shift case (hospital 208), but exhibits bias in the other shifted settings. The plug-in estimator also exhibits significant bias. Together, these results show that methods addressing only one source of bias, as well as naïve attempts to combine corrections, can systematically misrepresent deployment performance when covariate shift and selective labels occur simultaneously.

The RMSE results show that our DML method consistently has among the lowest RMSE. In some settings (hospital 443), CS-only performs comparably to our method, which has significantly lower error than the plug-in and SL-only estimators. In others (hospital 199), SL-only yields similar error to our method whereas the CS-only and plug-in estimators exhibit significantly higher error. For hospital 283, due to wide uncertainty intervals, we cannot determine which method has the lowest error. The RMSE results suggest that our DML estimator provides a reliable estimate of performance under covariate shift and selective labels. Overall, the combined bias and RMSE results illustrate that, when covariate shift and selective labeling are both present, correcting for either source of bias in isolation can give a misleading estimate of deployment risk.

Real-world eICU Experiments: We next illustrate our estimator in a fully real-world setting using eICU data. As above, we use hospital membership to observe naturally occurring covariate shift. However, now the labeling mechanism and the outcomes are taken directly from the observed clinical data.

We define binary outcomes from laboratory measurements obtained during the first 12–24 hours of ICU admission. Each outcome corresponds to a clinically meaningful threshold value that reflects an underlying complication, such as a heart condition. The prediction model uses patient demographics,

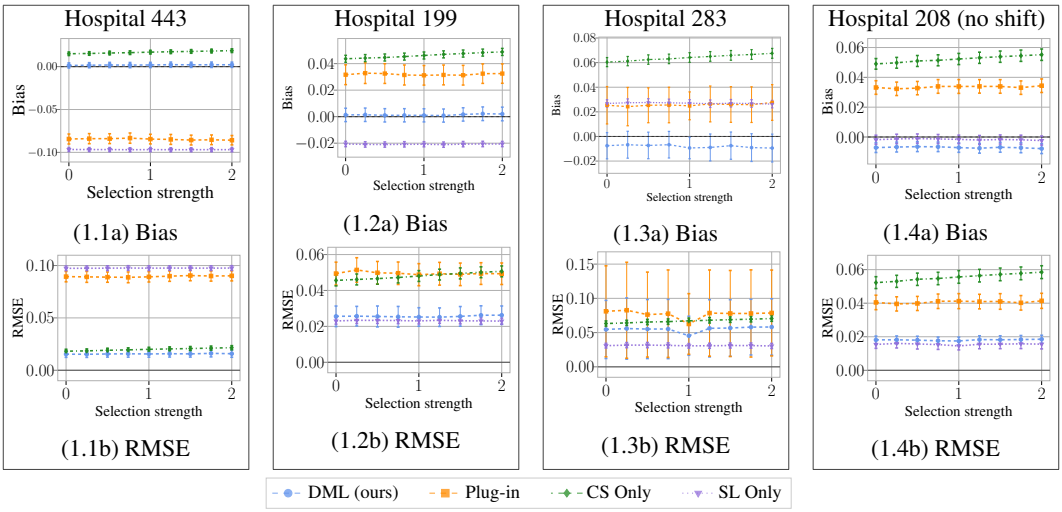


Figure 1: **Semi-synthetic eICU selection strength experiment.** Each column corresponds to a target hospital: 443, 199, and 283 represent covariate-shifted deployment populations, while 208 is a no-shift benchmark. Colors denote different evaluation methods that estimate the risk of a fixed prediction rule trained on source hospital 208. Outcomes and labeling are simulated such that risk is higher and label observability lower for older patients as the selection strength parameter on the horizontal axis increases. The top row reports bias and the bottom row RMSE relative to ground-truth target risk. Our DML estimator is the least biased (closest to zero) across the shifted hospitals (443, 199, and 283). Our estimator also achieves among the lowest RMSE across hospitals. Methods addressing only one source of bias (SL-only and CS-only), as well as naïve attempts to combine corrections (plug-in), can misrepresent deployment performance when covariate shift and selective labels jointly occur. Points denote means, and vertical bars denote 95% bootstrapped confidence intervals. Fig. 11 provides the same experiment with log loss.

admission characteristics, and ICU unit types as covariates. We train a fixed logistic regression prediction rule on labeled observations from hospital 208. We estimate its target risk in three covariate-shifted hospitals, 443, 199, and 283.

The key distinction from the semi-synthetic experiments is that the selective labeling is no longer simulated. Instead, label observability is based on real clinical decision-making: we only observe a given outcome for patients whose clinicians ordered the relevant lab in a specified admission window. Thus, the labeling indicator captures real variation in diagnostic and monitoring practices across hospitals, and, hence, heterogeneity in the selective labeling mechanism. At the same time, hospital membership captures naturally occurring covariate shifts in patient severity, characteristics, and demographics.

We do not observe labels for patients whose labs are not ordered. Thus, we cannot report bias or RMSE against an oracle risk as in the semi-synthetic setting. Instead, we report estimated target risks and interpret the implications of using each estimator in a clinical context.

We illustrate our estimator on four lab-dependent clinical outcomes in the eICU data: indicators of elevated B-type natriuretic peptide (BNP), elevated troponin-I, elevated creatine kinase (CK), and hypoxemia. For each outcome, the label is observed only among ICU patients for whom the relevant laboratory test was ordered within a given time window (24 hours for BNP, troponin-I, CK, and 12 hours for hypoxemia). This observability mechanism exactly reflects the real-world selective labels problem where the outcome is only observed when a clinician orders the corresponding diagnostic test. See Appendix H.4 for further discussion on these outcomes.

Real-world Results: Fig. 2 reports target-risk estimates for the selectively labeled clinical outcomes across target hospitals and a benchmark no-shift hospital. First, we observe that for some outcomes (BNP, hypoxemia), the estimated target risk changes substantially across hospitals while, for others (Troponin-I, CK), the estimates are more stable.

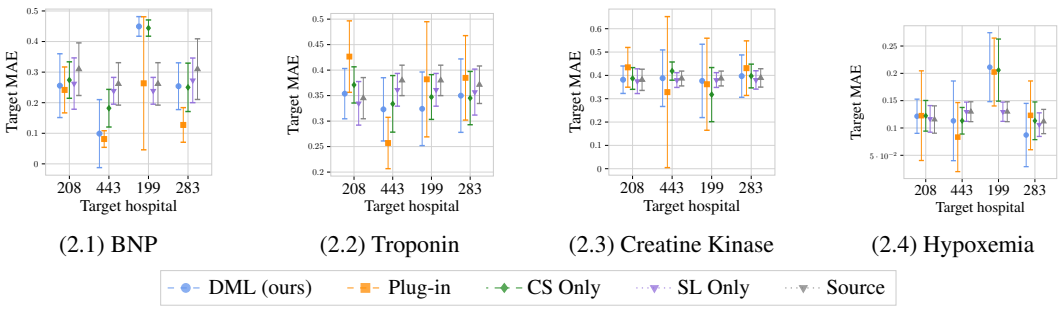


Figure 2: **Real-world eICU target-risk estimates.** Each panel reports the estimated target MAE for one clinical outcome. Hospital 208 is the no-shift benchmark; hospitals 443, 199, and 283 are covariate-shifted targets. We evaluate a fixed model trained on labeled data from hospital 208. Labels are observed only when the corresponding tests are ordered. The source baseline (▲) reports the unadjusted observed source risk. Points denote point estimates and vertical bars denote 95% bootstrap confidence intervals computed by resampling the source and target evaluation samples. Across tasks and hospitals, our DML estimator yields different point estimates and uncertainty intervals than baseline methods. The source and SL-only estimators yield significantly lower point estimates of MAE for BNP on hospitals 443 and 199 and for hypoxemia on hospital 199 relative to our DML estimate. The CS-only estimate reports lower uncertainty than our DML estimate for creatine kinase and hypoxemia on hospital 443. These discrepancies underscore the practical significance of a method that jointly addresses covariate shift and selective labels. Fig. 13 provides the same experiment with log loss.

In some instances, the DML, CS-only, and SL-only estimators approximately agree (e.g., BNP 208, troponin-I 199, CK 208 and 283, hypoxemia 283). In others, however, the DML aligns with the CS-only estimator and diverges from the SL-only estimator (e.g., hypoxemia 199), indicating that the covariate shift dominates the risk adjustment. At the same time, the instances of disagreement between the CS-only and DML estimators (e.g., CK 443, hypoxemia 443 and 283) reflect that accounting for selection can meaningfully change the practitioner's expectation of and confidence about the risk of deploying a clinical model in a new hospital.

We also note the relatively wide intervals for the plug-in estimator across hospitals and outcomes. This underscores the efficiency gains offered by our DML estimator over the naive direct inverse weighting approach, as reflected in both Theorem 3.7 and the semi-synthetic outcome experiments.

5 Limitations and Conclusion

This paper proposes a new framework for pre-deployment evaluation that, to our knowledge, is the first to account jointly for covariate shift and selective labels. Our approach formalizes target risk as the model's expected deployment performance, establishes conditions for its identification under observed data, and derives an influence-function-based double machine learning estimator. We prove that the proposed estimator is asymptotically linear, which yields asymptotic normality and supports Wald-type inference. In semi-synthetic eICU experiments, our estimator more accurately tracks target risk than standard plug-in procedures and evaluation procedures designed for each problem in isolation.

These results highlight the importance of developing tools that can account for multiple coexisting data challenges. In particular, the combination of covariate shift and selective labels, each of which has been studied extensively in isolation, poses compounded difficulties and is likely to arise in high-stakes domains such as healthcare.

These results come with important qualifications. Identification relies on substantive assumptions, including selection on observables and sufficient source-target overlap. These may fail, for instance, when clinical testing decisions depend on unmeasured severity indicators, or when target patients lie in regions of the covariate space that are impossible to observe in labeled source samples. Our robustness checks in Appendix F suggest some tolerance to small violations, but future work could develop methods that relax these assumptions or use adaptive data collection to improve overlap.

Accounting for co-occurring data challenges is especially important in fairness and safety assessments. By applying our method to subgroup risks, evaluators can assess fairness or safety properties under deployments characterized by covariate shift and selective labeling.

Finally, our framework has broader applications beyond clinical prediction. Many foundation model and LLM evaluation settings involve selectively observed feedback while the deployment distribution of prompts, users, and downstream tasks often differs substantially from that of the training and testing. Extending our approach to these settings could support more principled evaluation of modern ML systems under the real-world distortions that often arise together.

Acknowledgments

We are grateful to Avi Feller, Keegan Harris, Edward Kennedy, and several anonymous reviewers for insightful comments and discussion. Annie Ulichney's work is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752 and the Cooperative AI Foundation. Amanda Coston's work is supported by Schmidt Sciences. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the National Science Foundation.

References

[1] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[2] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[3] Jakub Białek, Wojtek Kuberski, Nikolaos Perrakis, and Albert Bifet. Estimating model performance under covariate shift without labels. *arXiv preprint arXiv:2401.08348*, 2024.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[5] Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*, 2023.

[6] Trenton Chang and Jenna Wiens. From biased selective labels to pseudo-labels: an expectation-maximization framework for learning from biased decisions. *arXiv preprint arXiv:2406.18865*, 2024.

[7] Jian Chen, Zhehao Li, and Xiaojie Mao. Learning under selective labels with data from heterogeneous decision-makers: An instrumental variable approach. *arXiv preprint arXiv:2306.07566*, 2023.

[8] Lingjiao Chen, Matei Zaharia, and James Y Zou. Estimating and explaining model performance when both covariates and labels shift. *Advances in Neural Information Processing Systems*, 35: 11467–11479, 2022.

[9] Victor Chernozhukov, Whitney Newey, Vıctor M Quintas-Martınez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022.

[10] Victor Chernozhukov, Michael Newey, Whitney K Newey, Rahul Singh, and Vasilis Srygkanis. Automatic debiased machine learning for covariate shifts. *arXiv preprint arXiv:2307.04527*, 2023.

[11] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 582–593, 2020.

[12] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, pages 2144–2155. PMLR, 2021.

[13] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *arXiv preprint arXiv:1807.00905*, 2018.

[14] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[15] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Decision making with limited feedback: Error bounds for recidivism prediction and predictive policing. *Proceedings of FAT/ML 2017*, 2017.

[16] Luke Guerdan, Justin Whitehouse, Kimberly Truong, Kenneth Holstein, and Zhiwei Steven Wu. Doubly-robust LLM-as-a-judge: Externally valid estimation with imperfect personas. *arXiv preprint arXiv:2509.22957*, 2025.

[17] Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[19] Zongliang Ji, Yifei Sun, Andre Carlos Kajdacsy-Balla Amaral, Anna Goldenberg, and Rahul G Krishnan. Can we generate portable representations for clinical time series data using LLMs? In *The Fourteenth International Conference on Learning Representations*, 2026.

[20] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pages 2439–2448. PMLR, 2018.

[21] Louis Kamulegeya, John Bwanika, Mark Okello, Davis Rusoke, Faith Nassiwa, William Lubega, Davis Musinguzi, and Alexander Börve. Using artificial intelligence on dermatology conditions in uganda: A case for diversity in training data sets for machine learning. *African Health Sciences*, 23(2):753–63, 2023.

[22] Masahiro Kato, Kota Matsui, and Ryo Inokuchi. Double debiased covariate shift adaptation robust to density-ratio estimation. *arXiv preprint arXiv:2310.16638*, 2023.

[23] Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

[24] Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

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

[26] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

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

[28] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4):1166–1173, 2024.

[29] Mark J Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

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

[31] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[32] Seong-ho Lee, Yanyuan Ma, and Jiwei Zhao. Doubly flexible estimation under label shift. *Journal of the American Statistical Association*, 120(549):278–290, 2025.

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

[34] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[35] Neil Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. *arXiv preprint arXiv:2404.00522*, 2024.

[36] Wang Miao and Eric J Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.

[37] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[38] Samantha Morrison, Constantine Gatsonis, Issa J Dahabreh, Bing Li, and Jon A Steingrimsson. Robust estimation of loss-based measures of model performance under covariate shift. *Canadian Journal of Statistics*, 52(4):e11815, 2024.

[39] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A BERT-based transfer learning approach for hate speech detection in online social media. In *International conference on complex networks and their applications*, pages 928–940. Springer, 2019.

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

[41] Jed Keenan Obra, Chandan Singh, Kenshata Watkins, Jean Feng, Ziad Obermeyer, and Aaron Kornblith. Potential for algorithmic bias in clinical decision instrument development. *npj Digital Medicine*, 2025.

[42] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

[43] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[44] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

[45] Thai-Hoang Pham, Jiayuan Chen, Seungyeon Lee, Yuanlong Wang, Sayoko E. Moroi, Xueru Zhang, and Ping Zhang. The boundaries of fair AI in medical image prognosis: A causal perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[46] Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. *arXiv preprint arXiv:2110.06177*, 2021.

[47] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[48] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2022.

[49] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

[50] Sashank Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[51] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[52] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[53] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.

[54] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[55] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[56] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019.

[57] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[58] Jakob Schoeffer, Maria De-Arteaga, and Jonathan Elmer. Perils of label indeterminacy: A case study on prediction of neurological recovery after cardiac arrest. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1080–1094, 2025.

[59] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

[60] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[61] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[62] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[63] Vallijah Subasri, Amrit Krishnan, Ali Kore, Azra Dhalla, Deval Pandya, Bo Wang, David Malkin, Fahad Razak, Amol A Verma, Anna Goldenberg, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Network Open*, 8(6):e2513685, 2025.

[64] Anurag Vaidya, Richard J Chen, Drew FK Williamson, Andrew H Song, Guillaume Jaume, Yuzhe Yang, Thomas Hartvigsen, Emma C Dyer, Ming Y Lu, Jana Lipkova, et al. Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine*, 30(4):1174–1190, 2024.

[65] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, pages 6638–6647. PMLR, 2019.

[66] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5): 1324–1370, 2013.

[67] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.

[68] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. Pmlr, 2013.

A Proofs

A.1 Proof of Proposition 3.5

Recall the notation $L \coloneqq \ell(f(X), Y)$. By law of iterated expectations,

$$\psi = \mathbb{E}_T [L] = \mathbb{E}_T [\mathbb{E} [L|R = 0, X]] . \tag{9}$$

By Assumption 3.2,

$$\mathbb{E}_T [\mathbb{E} [L|R = 0, X]] = \mathbb{E}_T [\mathbb{E} [L|R = 1, X]] . \tag{10}$$

By Assumption 3.1,

$$\mathbb{E}_T [\mathbb{E} [L|R = 1, X]] = \mathbb{E}_T [\mathbb{E} [L|R = 1, D = 1, X]] . \tag{11}$$

Combining (9)-(11) and recalling the definition of μ yields the first representation

$$\psi = \mathbb{E}_T [\mu(X)] . \tag{12}$$

To show the second representation, we start from (12). Using the change of measure justified by Assumption 3.4, we obtain:

$$\psi = \int_{x\in\mathcal{X}} \mu(x) p_T(x) dx = \int_{x\in\mathcal{X}} \mu(x) \frac{p_T(x)}{p_S(x)} p_S(x) dx = \mathbb{E}_S \left[\frac{p_T(X)}{p_S(X)} \mu(X) \right] . \tag{13}$$

Under the source law, conditioning on $R = 1$ is implicit by construction. Hence,

$$\mu(X) = \mathbb{E}[L \mid X, R = 1, D = 1] = \mathbb{E}_S[L \mid X, D = 1]. \tag{14}$$

Since $D \in \{0, 1\}$, we have

$$\mathbb{E}_S [DL|X] = \mathbb{P}(D = 1|X, R = 1)\mathbb{E}_S[L|X, D = 1] = \pi_S(X)\mu(X). \tag{15}$$

By Assumption 3.3, rearrange (15) to get

$$\mu(X) = \frac{\mathbb{E}_S [DL|X]}{\pi_S(X)}. \tag{16}$$

Substituting (16) into (13) yields

$$\psi = \mathbb{E}_S \left[\frac{p_T(X)}{p_S(X)} \frac{\mathbb{E}_S [DL|X]}{\pi_S(X)} \right] .$$

Applying the tower property yields the representation

$$\psi = \mathbb{E}_S \left[\frac{p_T(X)}{p_S(X)} \frac{DL}{\pi_S(X)} \right]$$

which proves the claim. □

A.2 Candidate Influence Function Derivation

The following lemma recalls well-known results characterizing the influence functions of conditional expectation and density functions. See, e.g., [24].

Lemma A.1 (Auxiliary Influence Functions)**.** *For the conditional loss function $\mu(x)$, its influence function $\mathbb{IF}\{\mu(X)\}$ is given by:*

$$\mathbb{IF}\{\mu(x)\} = \frac{D \cdot R \cdot \mathbb{1}\{X = x\}}{\mathbb{P}(X = x, R = 1, D = 1)} \left(L - \mu(x)\right). \tag{17}$$

Similarly, for the target covariate density $p_T(x)$, its influence function is given by:

$$\mathbb{IF}\{p_T(x)\} = \frac{1}{\rho}(1 - R)\left(\mathbb{1}\{X = x\} - p_T(x)\right). \tag{18}$$

Lemma A.2 (Target Risk Influence Function)**.** *Define*

$$\varphi(Z; \mathbb{P}) = \frac{RD}{\pi(X)}\frac{g(X)}{\rho}(L - \mu(X)) + \frac{1 - R}{\rho}(\mu(X) - \psi(\mathbb{P})). \tag{19}$$

Then $\mathbb{E}_{\mathbb{P}}\left[\varphi(Z; \mathbb{P})\right] = 0$ and, for every one-dimensional parametric sub-model $\mathbb{P}_\varepsilon = (1 - \varepsilon) \cdot \mathbb{P} + \varepsilon\overline{\mathbb{P}}$ with score function s_ε,

$$\frac{\partial}{\partial\varepsilon}\psi(\mathbb{P}_\varepsilon)\big|_{\varepsilon=0} = \mathbb{E}_{\mathbb{P}}\left[\varphi(Z; \mathbb{P})s_\varepsilon(Z)\right].$$

That is, $\varphi(\cdot; \mathbb{P})$ is an influence function for ψ.

A.2.1 Proof of Proposition 3.6

We separate the argument into two steps. First, we use the heuristic of [24] of treating the covariate space as discrete to demonstrate how the candidate function is obtained. Then, we verify the resulting candidate under the full model and a general covariate space.

Candidate influence function derivation: Following the semiparametric calculus of [24], we treat $\mathcal{X}$ as a discrete set, apply Gateaux differentiation separately to each of $\mu(\cdot)$ and $p_T(\cdot)$, and invoke the product rule for influence functions:

$$\mathbb{IF}\{\psi\} = \sum_{x\in\mathcal{X}} \mathbb{IF}\{\mu(x)\}\, p_T(x) + \sum_{x\in\mathcal{X}} \mu(x)\mathbb{IF}\{p_T(x)\}.$$

Applying the building block influence functions (17) and (18) given in Lemma A.1 together with Bayes' Rule, we obtain:

$$\mathbb{IF}\{\psi\} = \frac{R \cdot D}{\pi(X)}\frac{g(X)}{\rho}\left(L - \mu(X)\right) + \frac{(1 - R)}{\rho}\left(\mu(X) - \psi\right)$$

as claimed.

Verification of the candidate influence function: By Lemma A.3 detailed below, we see that the von Mises expansion in (21) has a linear term $\varphi(\cdot; \mathbb{P})$ and remainder $R_2(\mathbb{P}, \overline{\mathbb{P}})$ that is the product of nuisance functions. Therefore, by Kennedy [24, Lemma 2], we conclude that φ is the efficient influence function. □

A.3 von Mises Expansion

Lemma A.3 (von Mises expansion)**.** *For any two candidate laws $\mathbb{P}$ and $\overline{\mathbb{P}} \in \mathcal{P}$, the mapping $\psi : \mathcal{P} \to \mathbb{R}$ admits the expansion*

$$\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) = \int \varphi(z; \overline{\mathbb{P}})\, d(\overline{\mathbb{P}} - \mathbb{P})(z) + R_2(\mathbb{P}, \overline{\mathbb{P}}) \tag{20}$$

where φ is as defined in (19) *and the remainder term $R_2(\mathbb{P}, \overline{\mathbb{P}})$ is given by*

$$R_2(\mathbb{P}, \overline{\mathbb{P}}) = \left(\frac{\overline{\rho} - \rho}{\overline{\rho}}\right)\left(\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P})\right) + \frac{1}{\overline{\rho}}\int g\left(\frac{\pi}{\overline{\pi}} - 1\right)(\mu - \overline{\mu})\, d\mathbb{P} + \frac{1}{\overline{\rho}}\int (\overline{g} - g)\frac{\pi}{\overline{\pi}}(\mu - \overline{\mu}) d\mathbb{P}$$

where we have suppressed the arguments of functions in each term for brevity.

Proof of Lemma A.3. For any two candidate laws $\mathbb{P}$ and $\overline{\mathbb{P}}$ on $Z = (X, R, RD, RDY)$, the von Mises expansion of the estimand ψ around $\mathbb{P}$ is given by:

$$\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) = \int \varphi(z; \overline{\mathbb{P}})\, d(\overline{\mathbb{P}} - \mathbb{P})(z) + R(\mathbb{P}, \overline{\mathbb{P}}) \tag{21}$$

where $\varphi(z; \mathbb{P})$ is a candidate influence function of ψ under $\mathbb{P}$ and $R(\mathbb{P}, \overline{\mathbb{P}})$ is the remainder term which we will show is second-order. Since $\varphi(z; \overline{\mathbb{P}})$ is centered under $\overline{\mathbb{P}}$, (21) can be rearranged to express the remainder term as:

$$R_2(\mathbb{P}, \overline{\mathbb{P}}) = \psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) + \int \varphi(z; \overline{\mathbb{P}})\, d\mathbb{P}(z). \tag{22}$$

To evaluate the remainder, we express the influence function in terms of the nuisance terms $\mu(X)$, $\pi(X)$, and $g(X)$ defined with respect to $\mathbb{P}$ together with their counterparts $\overline{\mu}(X)$, $\overline{\pi}(X)$ and $\overline{g}(X)$ defined with respect to $\overline{\mathbb{P}}$.

By the definition of $\varphi(\cdot; \overline{\mathbb{P}})$,

$$\int \varphi(Z; \overline{\mathbb{P}}) d\mathbb{P} = \mathbb{E}_{\mathbb{P}}\left[\frac{RD}{\overline{\pi}(X)} \frac{\overline{g}(X)}{\overline{\rho}} \left(L - \overline{\mu}(X)\right)\right] + \mathbb{E}_{\mathbb{P}}\left[\frac{1-R}{\overline{\rho}} \left(\overline{\mu}(X) - \psi(\overline{\mathbb{P}})\right)\right].$$

Next, we make use of the following two identities which hold for any measurable h:

$$\mathbb{E}_{\mathbb{P}}[RDh(X,Y)] = \mathbb{E}_{\mathbb{P}}\left[\pi(X)\mathbb{E}_{\mathbb{P}}[h(X,Y) \mid X, R=1, D=1]\right],$$
$$\mathbb{E}_{\mathbb{P}}[(1-R)h(X)] = \mathbb{E}_{\mathbb{P}}[g(X)h(X)] = \rho\mathbb{E}_{\mathbb{P}}[h(X) \mid R=0],$$

to obtain

$$\int \varphi(Z; \overline{\mathbb{P}}) d\mathbb{P} = \mathbb{E}_{\mathbb{P}}\left[\frac{\pi}{\overline{\pi}} \frac{\overline{g}}{\overline{\rho}} \left(\mu - \overline{\mu}\right)\right] + \frac{1}{\overline{\rho}} \left(\mathbb{E}_{\mathbb{P}}[(1-R)\overline{\mu}] - \rho\psi(\overline{\mathbb{P}})\right). \tag{23}$$

Next, write

$$\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) = \left(\frac{\overline{\rho} - \rho}{\overline{\rho}}\right) \left(\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P})\right) + \frac{\rho}{\overline{\rho}} \left(\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P})\right). \tag{24}$$

Substitute (23), (24) into (22) and apply the definition of ψ to obtain

$$R(\mathbb{P}, \overline{\mathbb{P}}) = \left(\frac{\overline{\rho} - \rho}{\overline{\rho}}\right) \left(\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P})\right) + \frac{1}{\overline{\rho}} \int \left(\frac{\pi}{\overline{\pi}} \overline{g} \left(\mu - \overline{\mu}\right)\right) d\mathbb{P} + \frac{1}{\overline{\rho}} \int \left(g\left(\overline{\mu} - \mu\right)\right) d\mathbb{P}.$$

Add and subtract $\frac{g}{\overline{\rho}} \frac{\pi}{\overline{\pi}} \left(\mu - \overline{\mu}\right)$ inside the expectation:

$$\frac{\pi}{\overline{\pi}} \frac{\overline{g}}{\overline{\rho}} (\mu - \overline{\mu}) + \frac{g}{\overline{\rho}} (\overline{\mu} - \mu) = \frac{g}{\overline{\rho}} \left(\frac{\pi}{\overline{\pi}} - 1\right) (\mu - \overline{\mu}) + \frac{\overline{g} - g}{\overline{\rho}} \cdot \frac{\pi}{\overline{\pi}} (\mu - \overline{\mu})$$

and substitute to yield the desired result. □

A.4 Formal Statement and Proof of Theorem 3.7

First, we provide a formal general statement of Theorem 3.7. For a measurable function h, write $\|h\|_{L_2(P)} \coloneqq \left(\int h(x)^2 dP(x)\right)^{1/2}$.

Theorem A.4 (Asymptotic properties of the DML estimator $\widehat{\psi}$)**.** *Let $\widehat{\psi}$ be the estimator in* (6) *with nuisance estimators $\widehat{\pi}$, $\widehat{g}$, and $\widehat{\mu}$ trained on an auxiliary sample $\mathcal{S}_n'$ independent of $\mathcal{S}_n$. Suppose Assumptions 3.1-3.4 hold and that the following regularity conditions are satisfied:*

1. ***Nuisance consistency:***

$$\|\widehat{\mu} - \mu\|_{L_2(P)} = o_p(1), \quad \|\widehat{\pi} - \pi\|_{L_2(P)} = o_p(1), \quad \|\widehat{g} - g\|_{L_2(P)} = o_p(1). \tag{25}$$

2. ***Overlap of nuisance estimates:*** *There exists $\varepsilon > 0$ such that*

$$\mathbb{P}\left(\inf_{x \in \mathcal{X}} \widehat{\pi}(x) \geq \varepsilon/2, \quad \widehat{\rho} \geq \rho/2\right) \to 1, \quad \mathbb{P}\left(\widehat{g}(x) \in (0,1)\ \forall x \in \mathcal{X}\right) \to 1. \tag{26}$$

3. ***Moment condition:*** *There exists $M < \infty$ such that*

$$\mathbb{E}\left[(L - \mu(X))^2 \mid X, R = 1, D = 1\right] \leq M, \quad \mathbb{E}\left[(\mu(X) - \psi(P))^2\right] \leq M. \tag{27}$$

4. ***Product condition:***

$$\|\widehat{\mu} - \mu\|_{L_2(P)} \left(\|\widehat{\pi} - \pi\|_{L_2(P)} + \|\widehat{g} - g\|_{L_2(P)}\right) = o_p(n^{-1/2}). \tag{28}$$

Note that a sufficient condition is that each of $\widehat{\mu}, \widehat{\pi}$, and $\widehat{g}$ converges at a rate $o_p(n^{-1/4})$ in $L_2(P)$.

5. ***Finite variance:***

$$\mathbb{E}\left[\varphi(Z; P)^2\right] < \infty \tag{29}$$

Then, $\widehat{\psi}$ is asymptotically linear with influence function $\varphi(\cdot; P)$:

$$\widehat{\psi} - \psi(P) = \frac{1}{n}\sum_{i=1}^{n} \varphi(Z_i; P) + o_p(n^{-1/2}), \tag{30}$$

and

$$\sqrt{n}\left(\widehat{\psi} - \psi\right) \xrightarrow{d} \mathcal{N}\left(0, V_\varphi\right), \quad V_\varphi \coloneqq \mathrm{Var}(\varphi(Z; P)). \tag{31}$$

A.4.1 Proof of Theorem A.4

Throughout, we suppress the arguments of all nuisance functions for notational convenience. We adopt the notation

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(Z_i), \quad Pf = \int f(z) dP(z) = \mathbb{E}_P[f(Z)]$$

where P denotes the pooled law of Z. We assume throughout that $\widehat{\pi}, \widehat{g}, \widehat{\mu}$ are estimated on an auxiliary sample $\mathcal{S}'_n = \{Z'_i\}_{i=1}^n$, sampled independently from the evaluation sample $\mathcal{S}_n = \{Z_i\}_{i=1}^n$. Also define

$$\widehat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{R_i = 0\}.$$

Recall the DML estimator (6):

$$\widehat{\psi} = \frac{1}{\widehat{\rho}}\mathbb{P}_n\left(\frac{RD}{\widehat{\pi}}\widehat{g}(L - \widehat{\mu}) + (1 - R)\widehat{\mu}\right).$$

Subtracting $\psi(P)$ from both sides and using $\widehat{\rho} = \mathbb{P}_n(1 - R)$ gives

$$\widehat{\psi} - \psi(P) = \frac{\rho}{\widehat{\rho}}\mathbb{P}_n\varphi(\cdot; P) + \frac{1}{\widehat{\rho}}\mathbb{P}_n B.$$

where

$$B \coloneqq RD\left(\frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi}\right)(L - \mu) - RD\frac{\widehat{g}}{\widehat{\pi}}(\widehat{\mu} - \mu) + (1 - R)(\widehat{\mu} - \mu).$$

Rearranging,

$$\widehat{\psi} - \psi(P) = \mathbb{P}_n\varphi(\cdot; P) + \left(\frac{\rho}{\widehat{\rho}} - 1\right)\mathbb{P}_n\varphi(\cdot; P) + \frac{1}{\widehat{\rho}}(\mathbb{P}_n - P)B + \frac{1}{\widehat{\rho}}PB. \tag{32}$$

We proceed by controlling each term of (32) separately.

Preliminary bounds: Define the event

$$\mathcal{E} \coloneqq \left\{ \inf_{x \in \mathcal{X}} \widehat{\pi}(x) \geq \varepsilon/2, \widehat{\rho} \geq \rho/2, \widehat{g}(x) \in (0,1) \forall x \in \mathcal{X} \right\}.$$

By Assumption (26), $\mathbb{P}(\mathcal{E}) \to 1$. On $\mathcal{E}$, we have

$$\frac{1}{\widehat{\rho}} \leq \frac{2}{\rho}, \quad \left| \frac{\rho}{\widehat{\rho}} - 1 \right| \leq \frac{2}{\rho} |\widehat{\rho} - \rho|, \quad \left\| \frac{1}{\widehat{\pi}} \right\|_\infty \leq \frac{2}{\varepsilon}, \quad \left\| \frac{\widehat{g}}{\widehat{\pi}} \right\|_\infty \leq \frac{2}{\varepsilon}. \tag{33}$$

Moreover, since $\widehat{\rho} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{R_i = 0\}$ is the sample mean of i.i.d. Bernoulli(ρ) random variables, we have that

$$|\widehat{\rho} - \rho| = O_p(n^{-1/2}). \tag{34}$$

Also, on $\mathcal{E}$,

$$\left| 1 - \frac{\pi}{\widehat{\pi}} \right| = \frac{|\pi - \widehat{\pi}|}{\widehat{\pi}} \leq \frac{2}{\varepsilon} |\pi - \widehat{\pi}| \tag{35}$$

First term: First, we note that we can write

$$P\varphi = \frac{1}{\rho} \mathbb{E}\left[g(X) \mathbb{E}\left[\frac{RD}{\pi}(L - \mu) | X \right] \right] + \frac{1}{\rho} \mathbb{E}\left[(1 - R)(\mu - \psi) \right].$$

For the first term, by Law of Iterated Expectation and the definition of μ,

$$\mathbb{E}\left[\frac{RD}{\pi}(L - \mu) | X \right] = \mathbb{E}[L - \mu | X, R = 1, D = 1] = 0. \tag{36}$$

For the second term, by Law of Iterated Expectation and Law of Total Probability

$$\mathbb{E}\left[(1 - R)(\mu - \psi) \right] = \mathbb{E}\left[(\mu(X) - \psi) \mathbb{E}\left[1 - R | X \right] \right] = \mathbb{E}\left[g(X)(\mu - \psi) \right] = \rho \mathbb{E}_T\left[\mu - \psi \right] = 0.$$

Hence, $P\varphi = 0$ and it follows that

$$\mathbb{P}_n \varphi(\cdot; P) = (\mathbb{P}_n - P)\varphi(\cdot; P).$$

By (29), $\mathbb{E}[\varphi(Z; P)^2] < \infty$, so, by the central limit theorem,

$$\mathbb{P}_n \varphi(\cdot; P) = O_p(n^{-1/2}). \tag{37}$$

Second term: By (37) and (33), we get

$$\left(\frac{\rho}{\widehat{\rho}} - 1 \right) \mathbb{P}_n \varphi(\cdot; P) = O_p(n^{-1/2}) O_p(n^{-1/2}) = o_p(n^{-1/2}).$$

Third term: First note that, since the nuisance functions are constructed on the independent sample $\mathcal{S}_n'$,

$$\mathbb{E}\left[(\mathbb{P}_n - P) B \mid \mathcal{S}_n' \right] = 0.$$

Next, on $\mathcal{E}$, by (33) and triangle inequality,

$$|B| \leq RD \left| \frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi} \right| |L - \mu| + RD \left| \frac{\widehat{g}}{\widehat{\pi}} \right| |\widehat{\mu} - \mu| + (1 - R)|\widehat{\mu} - \mu| \leq \left| \frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi} \right| |L - \mu| + \left(1 + \frac{2}{\varepsilon} \right) |\widehat{\mu} - \mu|.$$

Then, by Cauchy-Schwarz,

$$P(B^2) \leq 2P\left(\left(\frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi} \right)^2 RD(L - \mu)^2 \right) + 2\left(1 + \frac{2}{\varepsilon} \right)^2 P\left((\widehat{\mu} - \mu)^2 \right). \tag{38}$$

For the first term of (38), by iterated expectation and (29),

$$\begin{aligned} P\left(\left(\frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi} \right)^2 RD(L - \mu)^2 \right) &= \mathbb{E}\left[\left(\frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi} \right)^2 \mathbb{E}\left[RD(L - \mu)^2 | X \right] \right] \\ &= \mathbb{E}\left[\left(\frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi} \right)^2 \pi(X) \mathbb{E}\left[(L - \mu)^2 | X, R = 1, D = 1 \right] \right] \\ &\leq M \left\| \frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi} \right\|_{L_2(P)}^2 \end{aligned}$$

Moreover, by (33) and since $\pi > \varepsilon$ under Assumptions 3.3 and 3.4, on $\mathcal{E}$,

$$\left|\frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi}\right| \leq \frac{2}{\varepsilon}|\widehat{g} - g| + |g|\left|\frac{1}{\widehat{\pi}} - \frac{1}{\pi}\right| \leq \frac{2}{\varepsilon}|\widehat{g} - g| + \frac{2}{\varepsilon^2}|\widehat{\pi} - \pi|.$$

Therefore,

$$\left\|\frac{\widehat{g}}{\widehat{\pi}} - \frac{g}{\pi}\right\|_{L_2(P)} \leq \frac{2}{\varepsilon}\left\|\widehat{g} - g\right\|_{L_2(P)} + \frac{2}{\varepsilon^2}\left\|\widehat{\pi} - \pi\right\|_{L_2(P)} = o_p(1)$$

where the final equality applies (25). We use (25) once again to evaluate the second term of (38):

$$P[(\widehat{\mu} - \mu)^2] = o_p(1).$$

Altogether, we have

$$P(B^2) = o_p(1).$$

Hence, the conditional variance of the empirical process term on $\mathcal{E}$ is bounded as

$$\operatorname{Var}\left((\mathbb{P}_n - P)B \mid \mathcal{S}_n'\right) = \frac{1}{n}\operatorname{Var}(B(Z)|\mathcal{S}_n') \leq \frac{1}{n}\mathbb{E}[B^2|\mathcal{S}_n'] = \frac{1}{n}P[B^2] = o_p(n^{-1}).$$

It follows that, by Chebyshev's inequality,

$$(\mathbb{P}_n - P)B = o_p(n^{-1/2}).$$

By (33) we conclude

$$\frac{1}{\widehat{\rho}}(\mathbb{P}_n - P)B = o_p(n^{-1/2})$$

Fourth term: As in (36),

$$\mathbb{E}\left[\frac{RD}{\pi}(L - \mu)|X\right] = 0.$$

Thus,

$$PB = -\mathbb{E}\left[RD\frac{\widehat{g}}{\widehat{\pi}}(\widehat{\mu} - \mu)\right] + \mathbb{E}\left[(1 - R)(\widehat{\mu} - \mu)\right].$$

Using law of iterated expectation combined with $\mathbb{E}[RD|X] = \pi(X)$ and $\mathbb{E}[1 - R|X] = g(X)$, we obtain

$$PB = -\mathbb{E}\left[\frac{\widehat{g}}{\widehat{\pi}}\pi(\widehat{\mu} - \mu)\right] + \mathbb{E}\left[g(\widehat{\mu} - \mu)\right] = \mathbb{E}\left[\left(g - \widehat{g}\frac{\pi}{\widehat{\pi}}\right)(\widehat{\mu} - \mu)\right].$$

Therefore, on the event $\mathcal{E}$,

$$\begin{aligned}
|PB| &\leq \left\|\widehat{\mu} - \mu\right\|_{L_2(P)}\left\|g - \widehat{g}\frac{\pi}{\widehat{\pi}}\right\|_{L_2(P)} \\
&\leq \left\|\widehat{\mu} - \mu\right\|_{L_2(P)}\left(\left\|g - \widehat{g}\right\|_{L_2(P)} + \left\|\widehat{g}\left(1 - \frac{\pi}{\widehat{\pi}}\right)\right\|_{L_2(P)}\right) \\
&\leq \left\|\widehat{\mu} - \mu\right\|_{L_2(P)}\left(\left\|g - \widehat{g}\right\|_{L_2(P)} + \frac{2}{\varepsilon}\left\|\widehat{\pi} - \pi\right\|_{L_2(P)}\right)
\end{aligned}$$

where the first inequality uses Cauchy-Schwarz, the second line follows by triangle inequality, and the last line uses (26), (33), and (35). By (28) and (33), we reach

$$\frac{1}{\widehat{\rho}}|PB| = o_p(n^{-1/2}).$$

Collecting bounds: Returning to (32), we have established so far that

$$\left(\frac{\rho}{\widehat{\rho}} - 1\right)\mathbb{P}_n\varphi(\cdot; P) = o_p(n^{-1/2}), \quad \frac{1}{\widehat{\rho}}(\mathbb{P}_n - P)B = o_p(n^{-1/2}), \quad \frac{1}{\widehat{\rho}}|PB| = o_p(n^{-1/2}).$$

Hence,

$$\widehat{\psi} - \psi(P) = \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i; P) + o_p(n^{-1/2})$$

which establishes asymptotic linearity (30). Finally, since $\mathbb{E}[\varphi(Z; P)^2] < \infty$, applying the central limit theorem with Slutsky's theorem gives (31) as claimed. □

A.4.2 Implications of labeled-source positivity and bounded likelihood ratio conditions

Here, we demonstrate two joint implications of the labeled-source positivity assumption Assumption 3.3 and source-target bounded likelihood condition Assumption 3.4. In particular, we illustrate that these two conditions imply (i) an analogous positivity condition in the pooled distribution $\pi(x) > 0$ and (ii) that $g(x) < 1$.

First, write $\pi(x) = \pi_S(x)\mathbb{P}(R = 1|X = x)$. By Bayes' rule,

$$\pi(x) = \pi_S(x)\frac{1-\rho}{1-\rho+\rho\cdot dP_T/dP_S(x)}. \tag{39}$$

The second term on the right hand side is strictly positive under Assumption 3.4 when $\rho \in (0, 1)$.

Also by Bayes' rule,

$$g(x) = \frac{\rho dP_T/dP_S(x)}{1-\rho+\rho\cdot dP_T/dP_S(x)} < 1 \tag{40}$$

where the inequality follows from Assumption 3.4. In other words, Assumption 3.3 and Assumption 3.4 together imply that the conditional probability that an individual with covariates $X = x$ is from the target distribution is bounded away from 1.

B Additional Related Works

In this section, we include additional discussion of related works.

Covariate shift: Our work addresses *covariate shift* [62], in which the marginal feature distribution $P(X)$ differs between the training and deployment environments while the conditional distribution of outcomes given features $P(Y|X)$ remains stable (see, e.g., [48, 34, 37] for broader surveys). Classical approaches to addressing covariate shift rely on importance weighting or density ratio estimation [62, 66, 25]. More recent work develops doubly robust procedures for estimation and evaluation under covariate shift [50, 22, 10, 38, 16].

Beyond methods for correcting covariate shift, a growing body of work addresses the problem of evaluating models under covariate shift [8, 5, 3] and assessing whether a given shift will degrade performance [49, 46, 35, 17]. However, even seemingly benign covariate shifts at the population level among labeled samples may mask significant performance degradation for subpopulations whose labels are systematically unobserved in training data [28]. Our work addresses these evaluation and equity concerns that emerge when covariate shift and selective labels occur simultaneously.

Multiple forms of distribution shift: In contrast to covariate shift, *conditional shift* occurs when the relationship between features and outputs $P(Y|X)$ changes across environments [54]. We note that the selective labels problem can induce an *apparent* conditional shift by systematically hiding certain labels, causing the observed distribution to differ from the true conditional law $P(Y|X)$ [30, 26]. However, our framework imposes additional structure that distinguishes our setting from the joint presence of covariate and conditional shifts [68, 5]: we assume selection on observables, which implies that the unlabeled samples remain informative about both the source covariate distribution and the selection process. This assumption enables a key methodological distinction relative to settings that treat multiple simultaneous shifts.

Selective labels and sample selection: The *selective labels* problem [30, 26] arises when outcome labels are observed only for individuals selected by a historical policy or a human decision-maker. This occurs in settings such as credit scoring, criminal justice, and clinical medicine, where outcomes are only revealed after approval, release, or testing decisions [40, 12]. In such settings, the labeled data are not representative of the full population on which the model will ultimately be used, which complicates both training and evaluation [20].

Existing approaches include counterfactual risk estimation using causal inference [11, 6], correcting selection bias using heterogeneity across decision-makers [26, 7], and targeted data acquisition or augmentation for subpopulations underrepresented in the labeled data [13, 28]. Our setting is also connected to questions about feedback loops and *performativity*, where interventions guided by model predictions change the process by which future labeled data are generated [43, 15].

Missing data: The selective labels problem is a special case of the broader challenge of *missing data* [33]. Under the assumption of selection on observables, selective labeling can be viewed as a missing at random (MAR) mechanism [55]. Classical approaches to estimation under MAR include inverse probability weighting [52, 59], outcome regression and augmentation [53, 51], and doubly robust estimation [57, 2]. Our approach is similar to the semiparametric missing-data literature in that it combines outcome regression with inverse-probability weighting to construct a doubly robust estimator. However, our setting departs from the standard MAR formulation by aiming to evaluate model performance in a target population with no observed outcomes. This distinction requires a procedure that simultaneously corrects for the selection and covariate transport mechanisms.

A recent related development is prediction-powered inference (PPI) [1], which provides a framework for statistical inference that combines a small labeled dataset with predictions for a large unlabeled set. PPI shares with our setting the goal of making use of unlabeled target data, though it does not address outcome missingness induced by historical decision policies or covariate shift. More broadly, this challenge also connects to concerns about outcome measurement quality, where the ground truth outcome is contested or incomplete [51, 18, 58].

Double Machine Learning: Our evaluation framework builds on the literature on *doubly robust* (DR) and *double machine learning* (DML) estimators [53, 57, 2]. These methods are well-suited to settings with incomplete or biased data due to the fact that the resulting estimators remain consistent in parametric settings if either the propensity score or outcome model are correctly specified [52, 51, 29]. They enjoy fast rates of convergence in nonparametric settings [23] and have been extended to debiased estimation of more general functionals using deep learning methods [9]. These methods have also been used for estimation in settings closely related to selective labels and dataset shift, e.g., policy learning [14], distribution shift [8, 10, 5, 3, 32], and data missingness [36, 65].

Algorithmic Fairness: Our setting is also connected to the literature on algorithmic fairness, since both covariate shift and selective outcome observability can systematically misrepresent a model's performance for subpopulations which are underrepresented in the labeled training data [4, 40, 27]. For instance, models trained on populations from well-resourced institutions and deployed in under-resourced settings may systematically underperform for patient populations with different characteristics [60, 31, 61, 21].

Moreover, when outcome observability is shaped by historical decisions, labeled data may systematically exclude the subpopulations for whom fairness harms are most significant [30, 20, 26, 11]. This gap presents an important fairness concern: without pre-deployment evaluation methods, practitioners cannot assess *a priori* whether their models will perpetuate or exacerbate inequities when deployed.

C Additional Discussion

C.1 Interpretation of Assumptions

In this section, we provide additional discussion of the real-world implications of our assumptions in our motivating setting of clinical decision making.

The assumption of selection on observables (Assumption 3.1) holds when X contains all features that jointly affect both label observability and the outcome. For instance, in clinical settings, the assumption is more plausible if the decision to order labs depends only on recorded vitals, severity, and prior lab results that are all contained in X; it may fail if clinicians also use unrecorded symptoms. This latter case corresponds to our robustness experiment in which an unobserved variable jointly affects D and Y (Fig. 6.4a). In practice, evaluators can review clinical workflows to understand what features influence clinical decision-making and assess whether X includes the recorded information known to influence labeling decisions. We note that, while these diagnostics can aid in revealing violations of this condition, they cannot exclude the use of unrecorded clinical information.

The covariate shift assumption (Assumption 3.2) is more credible when outcomes are defined and measured consistently, and when X captures major prognostic factors. It is more plausible in clinical settings to the extent that hospitals use the same clinical equipment and outcome measurement timing, and when there is a known biological process governing the relationship between features X and the outcome. It may fail, for instance, if hospitals vary in the timing at which they take measurements

of clinically-relevant features after administering treatment while also omitting treatment and measurement timing from X. Our conditional shift robustness experiments probe this sort of violation (Fig. 6.2a).

D Synthetic Experiments

D.1 Synthetic Experiments

We demonstrate the performance of our proposed DML estimator and compare it to the existing doubly robust pre-deployment evaluation methods designed to evaluate model performance under *either* covariate shift or selective labels.

Synthetic Data Generation: In each Monte Carlo iteration, we draw n_S source covariates and n_T target Gaussian covariates, each m-dimensional: $X_i^{(S)} \sim \mathcal{N}(\mu_S, \Sigma_S), \quad i \in \{1, \dots n_S\}$, $X_j^{(T)} \sim \mathcal{N}(\mu_T, \Sigma_T), \quad j \in \{1, \dots n_T\}$. We simulate covariate shift by varying $\mu_S \neq \mu_T$ and/or $\Sigma_S \neq \Sigma_T$. We pool the covariates into a single dataset $\{(X_k, R_k)\}_{k \in [n]}$ where $n \coloneqq n_S + n_T$, $R_k = 1$ denotes a source unit, and $R_k = 0$ denotes a target unit.

Labeling and Outcome Models: For source units $(R = 1)$, label observability follows a covariate-dependent Bernoulli mechanism with an overlap floor:

$$D|(X, R = 1) \sim \text{Bern}(\pi_S(X)), \quad \pi_S(X) = \pi_{\min} + (1 - 2\pi_{\min})\sigma(-\alpha^\top X). \tag{41}$$

The vector α controls the direction and strength of the selective labels mechanism, where the sign convention in (41) makes labels less likely to be observed in the direction α. In the fixed-selection experiments, we set $\alpha = c \cdot \mathbb{1}_m$ for a fixed, experiment-dependent scalar c. In experiments that vary selection strength, we sweep over values of c, with larger values corresponding to stronger dependence of label observability on the covariates. The constant $\pi_{\min} \in (0, 1/2)$ ensures that Assumption 3.3 is satisfied with $\varepsilon = \pi_{\min}$.

We generate continuous latent outcomes according to

$$Y = m(X) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{42}$$

where the conditional mean is defined relative to the fixed prediction model f. In particular, we construct the conditional mean as $m(x) = f(x) + \eta(x)$, such that, under squared loss, $L = (f(X) - Y)^2$, the conditional expected loss satisfies $\mu(x) = \mathbb{E}[L|X = x] = \eta(x)^2 + \sigma_\varepsilon^2$. In other words, this construction allows us to specify where the prediction model f performs poorly through $\eta(x)$. To understand the robustness of our framework to misspecification, we take $\eta(x)$ to be nonlinear throughout our experiments and refer the reader to Appendix H.2 for details.

D.1.1 Synthetic Experiment Results

The bias and Root Mean Squared Error (RMSE) of each estimator across four experiments, each varying either the magnitude of covariate shift, selection strength, or the ratio of samples from the source and target distributions are reported in Figure 3.

In the mean-shift experiment (Figure 3.1a, Figure 3.1b), the target distribution mean is shifted increasingly from the source mean. We construct this mean shift so that it is in the direction of increasing prediction error, i.e., in the direction of $\eta(x)$. This means that, as the mean shift magnitude grows, the target risk increases. We observe that SL-only and CS-only estimators grow increasingly biased for the target risk. The plug-in and DML estimators remain approximately unbiased. However, the DML estimator has lower RMSE than the other estimators, especially for larger shifts.

The covariance-shift experiments yield the same qualitative conclusion, except that the SL-only estimator decreases in bias as the shift increases. However, we note that, as reported in Figure 9, this is merely a product of the fact that the SL-only estimate of risk is a constant, initially upward biased estimate of the increasing true target risk.

As selection increasingly depends on covariates, the CS-only estimator deteriorates because it treats the selectively labeled source sample as representative after covariate transport. The SL-only estimator

remains biased because it corrects selective labels only in the source distribution. Both DML and the plug-in remain approximately unbiased, but DML once again achieves uniformly smaller RMSE.

Finally, the target/source ratio experiment varies the amount of unlabeled target data while holding both covariate shift and selective labeling fixed. The CS-only and SL-only estimators exhibit non-vanishing bias as the sample size increases. The DML and plug-in estimators remain approximately unbiased over the interval, while the DML maintains lower RMSE than the plug-in.

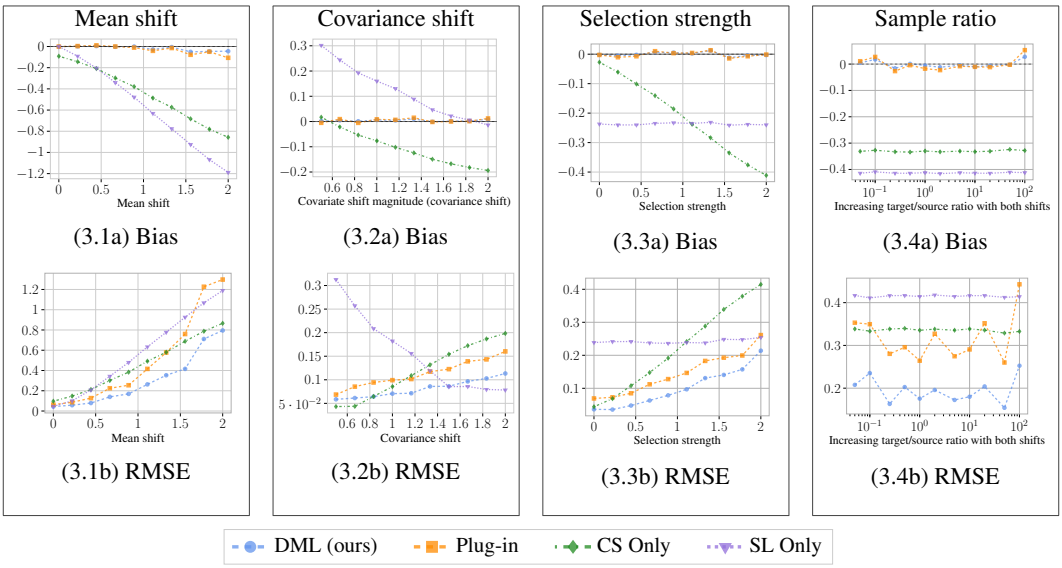


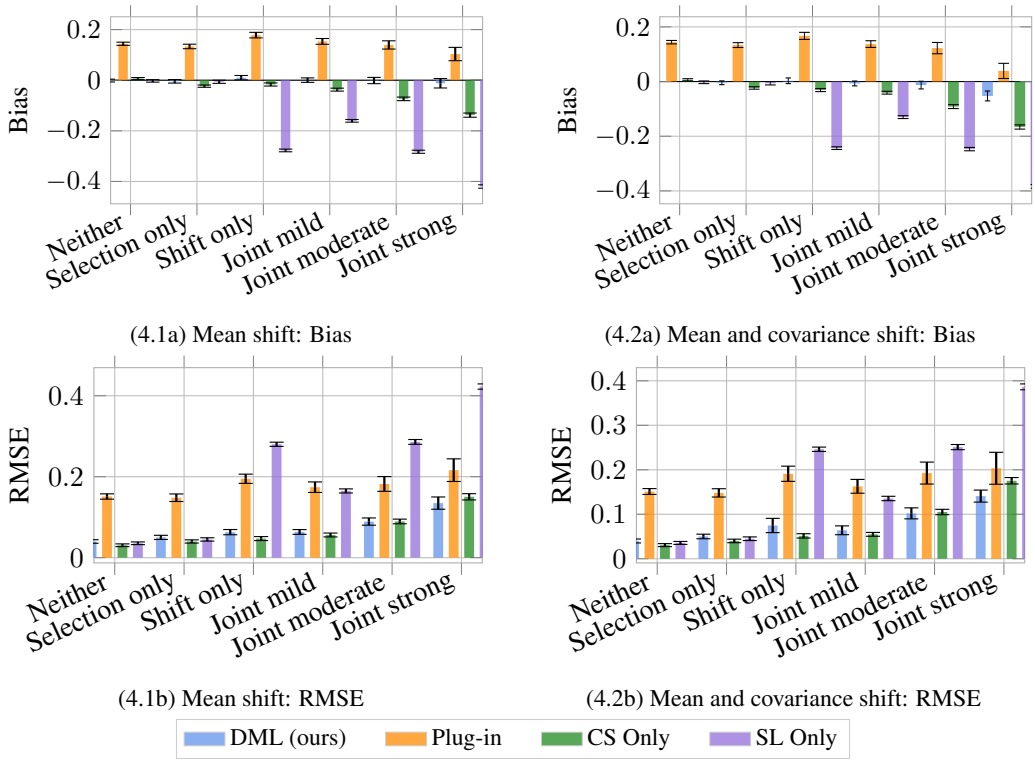
Figure 3: **Synthetic experiments under covariate shift and selective labels.** Each column summarizes one experiment and reports estimator bias in the top row and Root Mean Squared Error (RMSE) in the bottom row. The four panels sweep on the x-axis, respectively: (3.1) the magnitude of the target mean shift, (3.2) the magnitude of the target covariance shift, (3.3) the strength of the covariate-dependent labeling mechanism, and (3.4) the target/source sample size ratio under fixed covariate shift and selection bias. Note that the x-axes in (3.1)–(3.3) are linear and (3.4) is on a log scale. See Appendix H.2 for further details on the experimentation setup.

E Joint shift experiments

In this section, we provide additional synthetic experiments in which we progressively and jointly increase the amount of covariate shift and selectivity in our synthetic data generating process. These experiments motivate the importance of jointly considering both challenges simultaneously in evaluation and the improved performance of our DML estimator in such settings in comparison with single-issue methods.

Fig. 4 begins with the setting where neither selective labels nor covariate shift are present, considers each problem in isolation, then progressively strengthens their joint presence. Fig. 5 depicts combinations of three increasing degrees of covariate shift and selective labeling. We observe that the DML estimator has lower absolute bias than both the CS-only and SL-only estimator across all combinations in Fig. 5. In Fig. 4, we see that the CS-only and SL-only estimators exhibit comparable RMSE in the single-issue settings, but the DML has either (i) lower bias and comparable RMSE or (ii) both lower bias and lower RMSE. The plug-in estimator exhibits both milt consistent positive bias across all experiments and higher RMSE.

These results support the conclusion that, while specialized estimators can perform comparably when the mechanism they omit is absent or weak, the DML estimator achieves lower bias, RMSE, or, often, both once selective labeling and covariate shift are significant.



(4.1a) Mean shift: Bias

(4.2a) Mean and covariance shift: Bias

(4.1b) Mean shift: RMSE

(4.2b) Mean and covariance shift: RMSE

Figure 4: Progressive synthetic joint covariate shift and selection experiment. Bias (top) and RMSE (bottom) of the DML, plug-in, CS-only, and SL-only estimators. The left column varies target mean shift, whereas the right column jointly varies target mean and covariance shift. In each column, the settings proceed from neither covariate shift nor selective labeling to selective labeling only, covariate shift only, then mild, moderate, and strong joint violations. Vertical bars show 95% Monte Carlo confidence intervals.

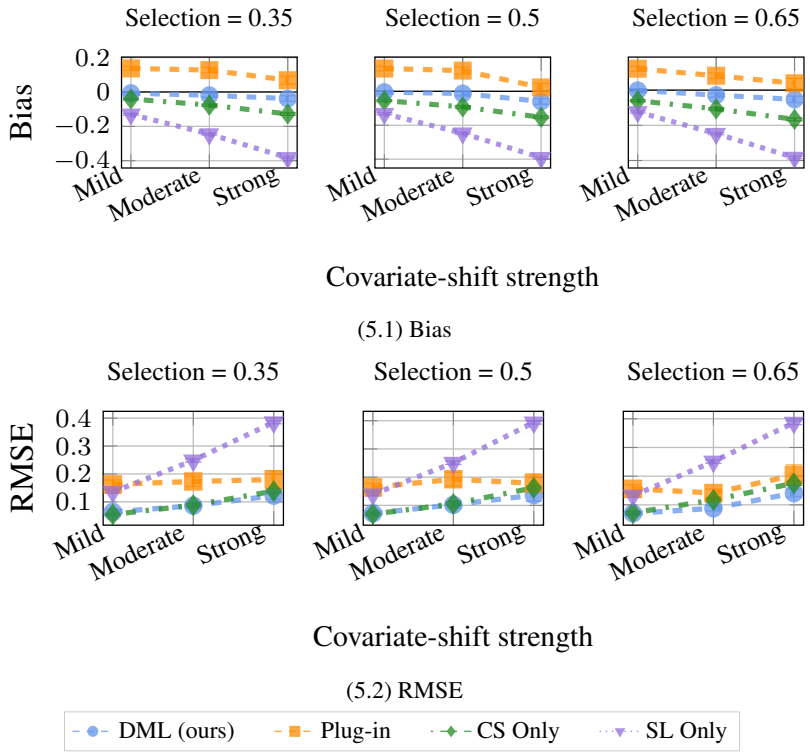


Figure 5: **Synthetic mean and covariance shift stress test.** Bias (top) and RMSE (bottom) of the DML, plug-in, CS-only, and SL-only estimators under simultaneous mean and covariance shifts in the target distribution at three levels of selective-labeling strength. Vertical bars show 95% Monte Carlo confidence intervals.

F Robustness Checks

We conduct an additional set of synthetic experiments to evaluate the sensitivity of each estimator to violations of the identifying assumptions (Assumption 3.1- Assumption 3.4). These experiments illuminate whether and how the performance of our DML estimator and the comparison estimators degrades as each assumption is increasingly violated.

The experiments use the same notation as the synthetic experiments. Here, the fixed prediction rule to be evaluated is $f(x) = 0.30x^\top u + 0.20x_1$ where $u = \mathbb{1}_m \in \mathbb{R}^m$. We take $\eta(x) = 1.90\sigma(1.50x^\top u) + 0.50\sin(x_1) - 0.60$.

To stress Assumption 3.4, we modify the target distribution to include a mixture component from a distant region in the covariate space. In particular, we take $X^{(T)} \sim (1-\lambda)\mathcal{N}(\mu_T, \Sigma_T) + \lambda\mathcal{N}(\mu_T + au, \Sigma_T)$, where u is the unit vector in the direction of increasing loss, a is a fixed offset parameter, and λ is the robustness parameter that we sweep. Note that $\lambda = 0$ corresponds to the synthetic experiment setting, and, as λ increases, more target mass is placed in a region with low source density. This increases the likelihood ratio.

We observe in Fig. 6.1a, Fig. 6.1b and Fig. 7.1 that the DML estimator is the least biased on the interval, but begins to incur a slight and relatively constant bias as the likelihood ratio grows. The plug-in estimator is the next-least biased, though its bias appears consistently monotone in the likelihood ratio magnitude. Interestingly, the CS-only estimator appears to become less biased as the ratio grows, and the SL-only has constant negative bias. DML enjoys the lowest RMSE on the interval.

To evaluate robustness to Assumption 3.2, we incorporate a target-only conditional mean shift. In particular, we suppose that target outcomes experience a mean shift such that the target conditional

expected loss is $\mu(x|R=0) = \left(\eta(x) + \left(\gamma x^\top u\right)\right)^2 + \sigma_\varepsilon^2$. The scalar γ is the parameter we sweep, where $\gamma = 0$ corresponds to the covariate shift assumption where $P_S(Y|X) = P_T(Y|X)$.

We observe in Fig. 6.2a, Fig. 6.2b, and Fig. 7.2 that all estimators become increasingly biased as the covariate shift assumption is increasingly violated. The DML and plug-in estimators perform similarly on the interval with respect to both bias and RMSE, and outperform all other estimators.

To stress Assumption 3.3, we enforce that a tail region of the source distribution has no observed labels. For a fraction q of tail, we set $D = 0$ in an extreme lower-tail region of the source law. We sweep q, where higher values cause more instability for the inverse-propensity weights.

We observe in Fig. 6.3a, Fig. 6.3b, and Fig. 7.3 that all four estimators remain approximately unbiased under increasing positivity stress. The DML, CS-only, and plug-in demonstrate low RMSE, but the SL-only demonstrates increasing RMSE under increasing positivity stress.

To evaluate the robustness to Assumption 3.1, we introduce a latent variable $U \sim \mathcal{N}(0,1)$ that affects both source label observability D and the outcome Y. In particular, we draw $D|(X,U,R=1) \sim \text{Bernoulli}\left(\pi_{\min} + (1-2\pi_{\min})\sigma(-\alpha^\top X - cU)\right)$ and $Y = f(x) + \eta(x) + cU + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,\sigma_\varepsilon^2)$. The scalar c is the parameter we sweep, with $c = 0$ corresponding to selection on observables. As c increases, label observability and outcomes jointly increasingly depend on U.

We observe in Fig. 6.4a, Fig. 6.4b, and Fig. 7.4 that all estimators become increasingly biased under increasing violations of the selection on observables assumption. The DML and plug-in estimators perform comparably with respect to bias, and outperform the other estimators for small to moderate violations. For larger violations, the SL-only estimator becomes less biased. A similar pattern is observed for RMSE.

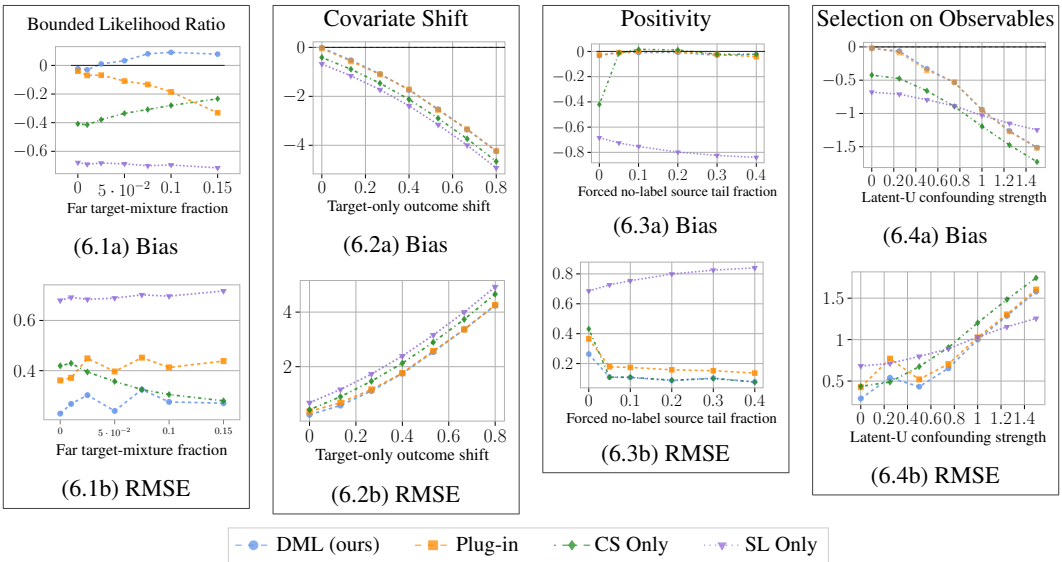
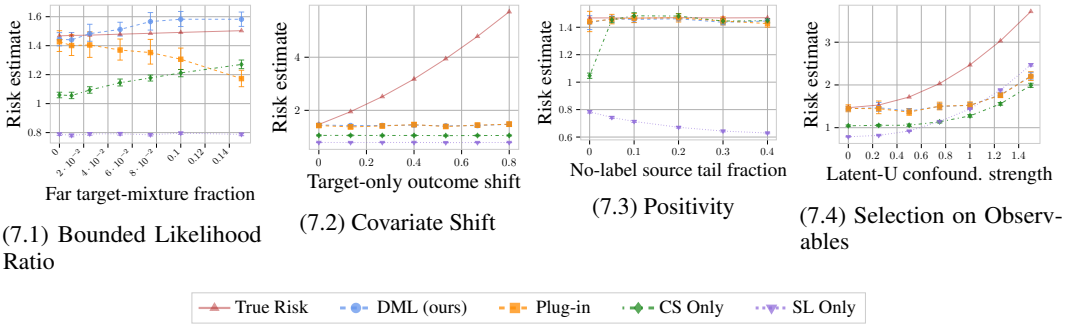


Figure 6: **Robustness checks to violations of identifying assumptions.** Each column reports the results of a controlled violation of an identifying assumption. Within each column, the top panel reports Monte Carlo bias relative to the oracle target risk and the bottom panel reports Root Mean Squared Error (RMSE). The horizontal axis in each column is the corresponding violation parameter for that assumption, with larger values corresponding to more severe violations.



(7.1) Bounded Likelihood Ratio

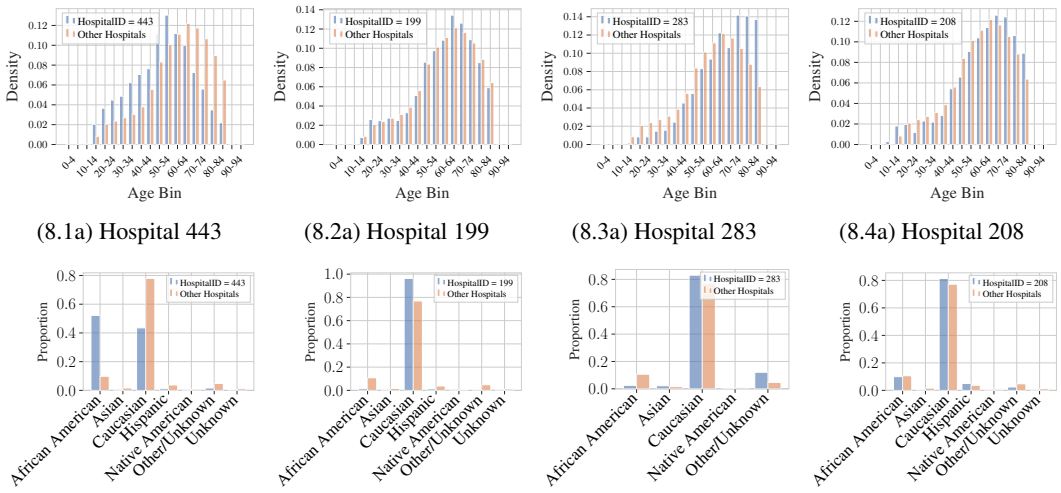
(7.2) Covariate Shift

(7.3) Positivity

(7.4) Selection on Observables

Figure 7: **Risk estimates under violations of identifying assumptions.** Each panel shows the estimated target risk as the corresponding assumption violation becomes more severe. The red curve denotes the oracle target risk computed from the known data-generating process.

G Additional Figures



(8.1a) Hospital 443

(8.2a) Hospital 199

(8.3a) Hospital 283

(8.4a) Hospital 208

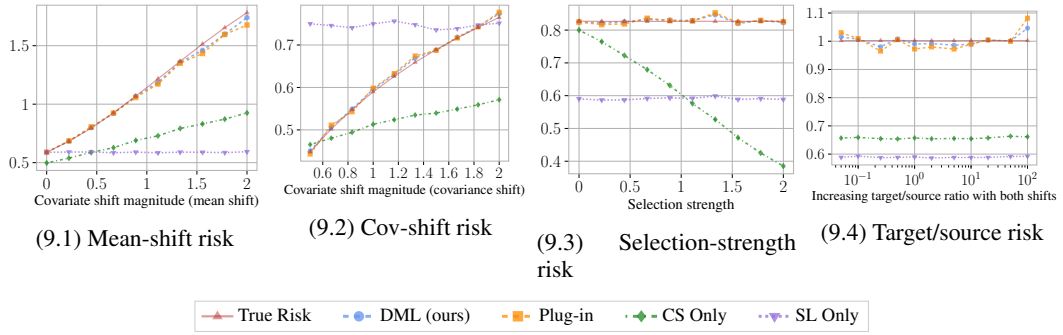
(8.1b) Hospital 443

(8.2b) Hospital 199

(8.3b) Hospital 283

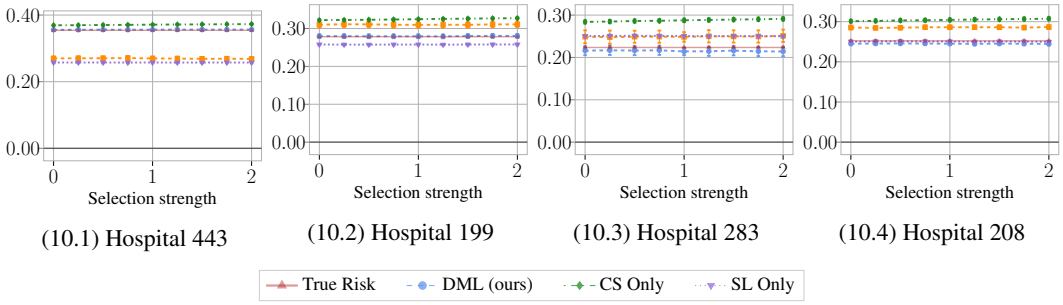
(8.4b) Hospital 208

Figure 8: **eICU Inter-Hospital Covariate Shifts.** Age (top row) and ethnicity (bottom row) distributions across hospitals in the eICU data. Hospital 443 tends to have younger patients and more African American patients; Hospital 199 exhibits a typical age profile and more Caucasian patients; Hospital 283 skews older with a larger share of patients labeled as unknown or "other" ethnicity; Hospital 208 is approximately average on both.



(9.1) Mean-shift risk (9.2) Cov-shift risk (9.3) Selection-strength risk (9.4) Target/source risk

Figure 9: **Synthetic experiments under covariate shift and selective labels risk values.** Each column summarizes one experiment and reports estimated risk. The four sweeps vary (i) the magnitude of the target mean shift, (ii) the magnitude of the target covariance shift, (iii) the strength of the covariate-dependent labeling mechanism, and (iv) the target/source sample size ratio under fixed covariate shift and selection bias. Note that the x-axes in (i)–(iii) are linear, while (iv) is on a log scale. See Appendix H.2 for further details on the experimental setup.



(10.1) Hospital 443 (10.2) Hospital 199 (10.3) Hospital 283 (10.4) Hospital 208

Figure 10: **Semisynthetic eICU experiments under increasing selection strength risk values.** Each column summarizes one experiment and reports estimated risk. The horizontal axis corresponds to increasing feature dependence in the selection mechanism. The experiment settings correspond to those of Fig. 1.

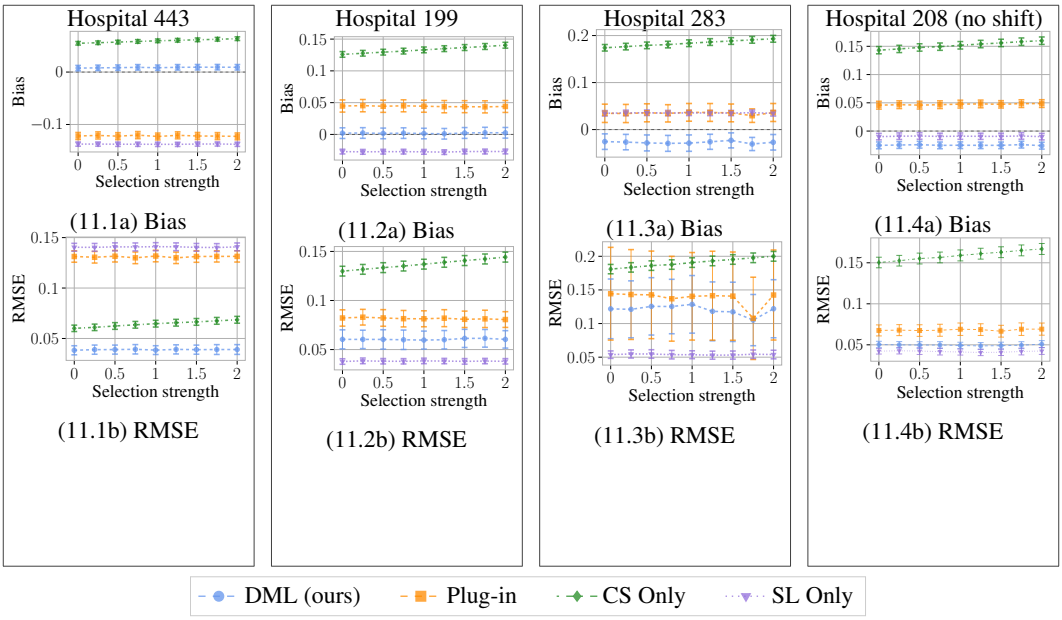


Figure 11: **Semi-synthetic eICU selection strength experiment under log loss.** Bias (top) and RMSE (bottom) of the DML, plug-in, CS-only, and SL-only estimators as selection strength increases. Columns correspond to target hospitals 443, 199, 283, and 208; hospital 208 serves as the no covariate shift benchmark. The design matches the corresponding absolute loss experiments in Fig. 1 with log loss replacing absolute loss. We note that the scale of the Bias and RMSE change under the new loss function, but the qualitative patterns of agreement and disagreement of estimators under the original experiments and the new log loss are consistent with our original findings

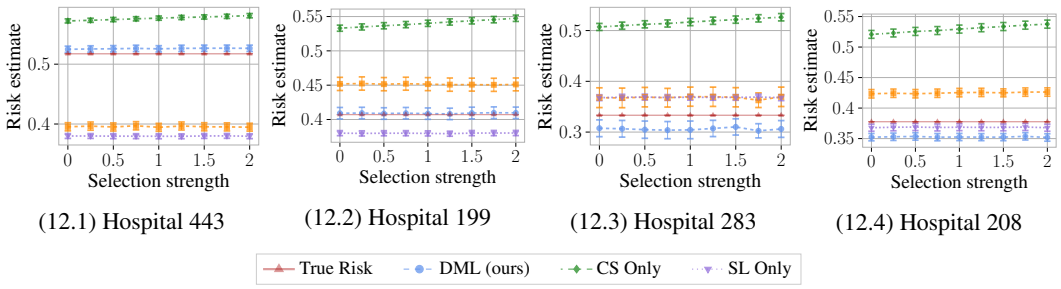


Figure 12: **Semi-synthetic eICU experiments under increasing selection strength risk values with log loss.** Each column summarizes the risk corresponding to Fig. 11.

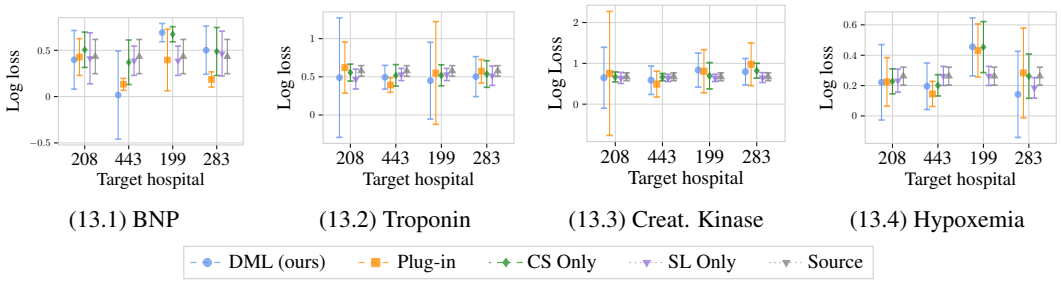


Figure 13: **Real-world eICU target-risk estimates under log loss.** Estimated target risks for BNP, troponin, creatine kinase, and hypoxemia from left to right. The design matches the corresponding absolute loss experiment that is the setting of Fig. 2, but with log loss instead of absolute loss.

H Experimentation Details

H.1 Estimator Details

Here we provide precise formulae for the estimators with which we compare the DML estimator.

DML Estimator: We also compute the DML estimator for the target risk by implementing (6). :

$$\widehat{\psi}_{\mathrm{DML}} = \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{R_k \cdot D_k}{\widehat{\pi}(X_k)} \cdot \frac{\widehat{g}(X_k)}{\widehat{\rho}} \cdot (L_k - \widehat{\mu}(X_k)) + \frac{1-R_k}{\widehat{\rho}} \cdot \widehat{\mu}(X_k)\right\}.$$

Naïve (Plug-in) Estimator: A natural benchmark is the plug-in estimator that directly combines observed labeled source losses. The observed losses are reweighted by the estimated density ratio $\widehat{w}(x)$ and the inverse propensity weights $1/\widehat{\pi}(x)$ to account for both the covariate shift and selective labels. We compute:

$$\widehat{\psi}_{\text{plug-in}} = \frac{1}{n}\sum_{k=1}^{n}\frac{R_k \cdot D_k}{\widehat{\pi}(X_k)} \cdot \frac{\widehat{g}(X_k)}{\widehat{\rho}} \cdot L_k.$$

Covariate Shift Benchmark: To isolate the importance of correcting for selective labels, we implement the doubly robust estimator for evaluation under covariate shift from [38]. Their estimator addresses the setting where the source and target covariate distributions differ according to Assumption 3.2; however, it assumes that outcomes are observed for all source units. We adapt their estimator to our setting where labels are only observed for treated source units (i.e., where $R_i = 1, D_i = 1$):

$$\widehat{\psi}_{\mathrm{CS}} = \frac{1}{n_T}\sum_{i=1}^{n}\left\{\frac{\widehat{g}(X_i)}{1-\widehat{g}(X_i)} \cdot R_i D_i \cdot (L_i - \widehat{\mu}(X_i)) + (1-R_i)\widehat{\mu}(X_i)\right\}.$$

Their estimator corrects for covariate shift via the density ratio $\frac{\widehat{g}(X_i)}{1-\widehat{g}(X_i)}$, but it does not account for selective labels.

Selective Labels Benchmark: To isolate the importance of correcting for covariate shift, we implement the doubly robust counterfactual risk estimator from [11]. Their estimator addresses the setting where outcome labels are only observed for treated units and uses inverse propensity weighting to account for the selective labeling mechanism. However, their setting assumes that the training and deployment populations share the same covariate distribution. We use their estimator to evaluate the risk on the source distribution using selectively labeled source samples:

$$\widehat{\psi}_{\mathrm{SL}} = \frac{1}{n_S}\sum_{i=1}^{n_S}\left\{\frac{D_i}{\widehat{\pi}_S(X_i)}(L_i - \widehat{\mu}(X_i)) + \widehat{\mu}(X_i)\right\}.$$

H.2 Synthetic Experiments

This appendix provides the full specification of the synthetic experiments. In all synthetic experiments, we fix the dimension as $m = 5$. In the mean-shift, covariance shift, and selection-strength shift experiments reported in Figures 3.1a, 3.1b, 3.2a, 3.2b, 3.3a, 3.3b, we fix $n_S = 1000, n_T = 4000$. We use $n_{\mathrm{oracle}} = 50,000$ Monte Carlo iterations to estimate the target risk and implement $N_{\mathrm{it}} = 200$ iterations for each setting in the grid. Across all experiments, we evaluate the fixed, non-linear prediction model $f(x) = 0.30x^\top u + 0.20x_1$ where x_1 denotes the first coordinate of x and $u := \left(\frac{\mathbb{1}_m}{\|\mathbb{1}_m\|_2}\right)$. First, we take $\eta(x) = 1.80\sigma(1.50x^\top u) + 0.50\sin(x_1) - 0.60$. We generate outcomes according to $Y = m(x) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $m(x) = f(x) + \eta(x)$. Thus, the gap between the prediction rule and the conditional mean outcome given $X = x$ is specified by the nonlinear choice $\eta(x)$. Also note that we take $\sigma_\varepsilon = 0.15$ and $\pi_{\min} = 0.005$ across all synthetic experiments. We use 3-fold cross-fitting to fit nuisance parameters across all synthetic experiments.

Increasing mean-shift: Fix $\mu_S = (0, \ldots, 0) \in \mathbb{R}^m$, $\Sigma_T = \Sigma_S = I_m$, $\alpha = 1 \cdot u$ (i.e., $c = 1$), $\beta = u$. For $s \in [0.0, 0.222, 0.444, 0.667, 0.889, 1.111, 1.333, 1.556, 1.778, 2.0]$, we take $\mu_T = \mu_S + s \cdot u$.

Increasing covariance-shift: Fix $\mu_S = \mu_T = (0, \ldots, 0) \in \mathbb{R}^m$, $\Sigma_S = 3I_m$, $\alpha = 1 \cdot u$. For $c \in [0.5, 0.667, 0.833, 1.0, 1.167, 1.333, 1.5, 1.667, 1.833, 2.0]$, we take $\Sigma_T(x) = I_m + (c-1)uu^\top$.

Increasing selection strength: Fix $\mu_S = (0, \ldots, 0) \in \mathbb{R}^m$, $\Sigma_S = I_m$, $\Sigma_T = I_m$, and $\mu_T = \mu_S + 0.5u$. For $a \in [0.0, 0.222, 0.444, 0.667, 0.889, 1.111, 1.333, 1.556, 1.778, 2.0]$, take $\alpha = a \cdot u$.

Increasing Target/source sample size ratio: Fix $n_S = 1000$ and vary $\frac{n_T}{n_S} \in [0.05, 0.1, 0.25, 0.5, 1, 2, 5, 10, 20, 50, 100]$. For each ratio r, we set $n_T = \mathrm{round}(1000r)$ and fix $\mu_S = (0, \ldots, 0) \in \mathbb{R}^m$, $\mu_T = \mu_S + 0.75u$, $\Sigma_S = I_m$, $\Sigma_T = I_m + (1.25 - 1)uu^\top$, and $\alpha = a \cdot u$.

H.3 Semi-synthetic Experiments

Data: We construct a covariate vector with $m = 26$ patient-level demographics and admission variables. Specifically, we include indicators for gender, ethnicity, age bin, and ICU unit type, together with admission height, weight, and body mass index (BMI). Age is binned into categories 0–19, 20–39, 40–59, 60–79, 80–89, and 90+. The continuous variables are median imputed to remove non-finite values and then standardized.

The fixed prediction rule f is trained once on a training split from hospital 208 using 1095 observations and evaluated on 730 test observations from a test split. The oracle risk is computed over a held-out target evaluation sample of sizes 1829, 2120, 1049, 1301 for hospitals 443, 199, 283, and 208, respectively.

Target and source hospitals: Though eICU hospitals are de-identified, the data report hospital-level metadata which we give here for additional context on the real-world shifts of our experiments. All hospitals are in the United States. Hospital 208 is a large non-teaching hospital in the South, hospital 443 is a large teaching hospital in the South, hospital 199 is a large teaching hospital in the Northeast, and hospital 283 is a mid-sized non-teaching hospital in the Midwest.

Setup: Across semi-synthetic experiments, we generate binary outcomes according to the logistic model $Y|X = x \sim \mathrm{Bernoulli}(\sigma(x^\top \beta))$ where σ denotes the sigmoid function. The coefficient vector $\beta \in \mathbb{R}^{26}$ is sparse. In particular, for the coordinates corresponding to the covariates age 60–79, age 80–89, age 90+, BMI, and weight, we set the coordinates of β to be $2.5, 4.0, 5.0, 0.5, 0.3$, respectively, with all other entries 0. This makes it so that older patients are higher risk with some additional variation by weight and BMI.

To train a fixed f, we draw binary labels on the training split held out for the prediction model and fit a random forest classifier with maximum depth 6 and minimum leaf size 10. This classifier is held fixed across all Monte Carlo iterations and all values of the selection strength sweep. In the shifted hospital experiments, the fixed model is the one trained on Hospital 208, which we take to be the source hospital.

For the labels, we construct $\alpha \in \mathbb{R}^{26}$ as follows. First, set the coordinates corresponding to age 80–89 and 90+ of α' to be $-2, -2.5$, respectively, with all other entries 0, then set α to be the unit-norm vector in the same direction as α'. Then, set $s(x) = x^\top \alpha + (x^\top \alpha)^2 - \frac{1}{n_S} \sum_{i:R_i=1} (X_i^\top \alpha)^2$. For selection strength a, the labeling probability for $X = x$ is $e_a(x) = \pi_{\min} + (1 - 2\pi_{\min})\sigma(C_a + as(x))$ where $\pi_{\min} = 0.05$ and C_a is an intercept calibrated for each a such that the labeling probability is approximately held constant at 0.3. Then, we draw $D_i|(X_i, R_i = 1) \sim \mathrm{Bernoulli}(e_a(x))$.

We sweep $a \in [0, 0.214,\ 0.429,\ 0.643,\ 0.857, 1.071, 1.286, 1.500]$. We run 20 Monte Carlo iterations per selection strength value. The target estimand is mean absolute prediction error. Since the outcome model is known in the semi-synthetic setting, we evaluate the oracle target risk via $\lambda_{\text{true}} = \frac{1}{N_{\text{oracle}}} \sum_{i \in \mathcal{S}_{\text{oracle}}} \left(\sigma(x^\top \beta)|f(X_i) - 1| + (1 - \sigma(x^\top \beta))|f(X_i)|\right)$ where $\mathcal{S}_{\text{oracle}}$ denotes the oracle hold-out set. All nuisance functions are fit with 5-fold cross-fitting. The domain and source classifiers are trained via logistic regression.

For the loss regression, we estimate the conditional outcome probability $\widehat{p}(x)$ via a random forest regression with 300 trees, max depth 4, minimum leaf size 50. This is trained with labeled source observations and we then plug the estimate into the conditional absolute loss formula $\widehat{p}(x)|f(x) - 1| + (1 - \widehat{p}(x))|f(x)|$.

Bias and RMSE: For a fixed setting across all synthetic and semi-synthetic experiments, let $\widehat{\psi}_{e,i}$ denote the target risk estimate from estimator e in Monte Carlo iteration $i \in \{1, \ldots, N_{\text{it}}\}$. Let ψ^{oracle} denote the corresponding oracle estimator of the target risk. We define Bias and RMSE for estimator e as:

$$\text{Bias}_e = \frac{1}{N_{\text{it}}} \sum_{i=1}^{N_{\text{it}}} (\widehat{\psi}_{e,i} - \psi^{\text{oracle}}), \quad \text{RMSE}_e = \sqrt{\frac{1}{N_{\text{it}}} \sum_{i=1}^{N_{\text{it}}} (\widehat{\psi}_{e,i} - \psi^{\text{oracle}})^2}.$$

H.4 Real-world Outcome Experiments

Outcome clinical significance: Here we provide clinical context for outcomes of interest in our real-world outcome experiments. We select these outcomes because they are clinically meaningful, yet not universally observed for patients in the ICU: the outcomes are unobserved unless, for instance, a clinician suspects a condition and subsequently orders further tests. Elevated BNP is an indication of cardiac stress. Troponin-I is a marker of damage to the heart muscle, and is an indicator of myocardial injury, sepsis, or shock, among other conditions. CK is an enzyme released in skeletal muscle injury, heart damage, infections, rhabdomyolysis, and other serious conditions. PaO2 measures the pressure of oxygen dissolved in the plasma. Low PaO2 characterizes hypoxemia which can reflect underlying conditions such as lung conditions and heart issues.

Outcome details: We define the elevated BNP outcome as $Y = \mathbb{1}\{\max(\text{BNP}) > 500$ pg/mL within 24 hours of ICU admission$\}$. We define the elevated troponin-I outcome as $Y = \mathbb{1}\{\max(\text{troponin}) > 1.0$ ng/mL within 24 hours of ICU admission$\}$. We define the elevated CK outcome as $Y = \mathbb{1}\{\max(\text{CK}) > 500$ U/L within 24 hours of ICU admission$\}$. Finally, we define hypoxemia as $Y = \mathbb{1}\{\min(\text{PaO}_2) < 60$ mmHg within 12 hours of ICU admission$\}$.

H.5 Resources

All experiments were run on a Linux server running Ubuntu 24.04.4 LTS with an Intel Xeon E5-2609 CPU at 2.40GHz. The machine has 8 CPU cores across two sockets and 125GiB of RAM. Experiments were run using Python 3.13.2 with NumPy 2.1.3, pandas 2.2.3, scikit-learn 1.6.1, SciPy 1.15.2, and Matplotlib 3.10.1. No GPU acceleration was used.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of framework and analysis explicitly.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each theoretical results, the paper provides a full set of assumptions and a complete and correct proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed descriptions of our experiments as well as experiment code. We also provide the code, though the data requires additional access since it contained individual-level health information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We provide open access to the code used to reproduce the main experimental results as described in supplemental material. We provide open access to all code necessary to reproduce our synthetic data and synthetic data experiments. We do not, however, release the data for the real-world eICU experiments, because the Physionet eICU Collaborative Research Database consists of individual-level healthcare data is is not freely accessible without restriction.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper presents a level of detail that is necessary to appreciate the results and make sense of them in the body, and full details are provided in the code and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results are accompanied by error bars or confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper indicates the relevant memory, compute, and storage information in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the harms of improper prediction model evaluation generally.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Assets are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor does it conduct research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor does it conduct research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.