Counterfactual Prediction Under Outcome Measurement Error

Luke Guerdan
lguerdan@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Amanda Coston
acoston@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Kenneth Holstein
kjholste@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Zhiwei Steven Wu
zstevenwu@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Across domains such as medicine, employment, and criminal justice, predictive models often target labels that imperfectly reflect the outcomes of interest to experts and policymakers. For example, clinical risk assessments deployed to inform physician decision-making often predict measures of healthcare utilization (e.g., costs, hospitalization) as a proxy for patient medical need. These proxies can be subject to outcome measurement error when they systematically differ from the target outcome they are intended to measure. However, prior modeling efforts to characterize and mitigate outcome measurement error overlook the fact that the decision being informed by a model often serves as a risk-mitigating intervention that impacts the target outcome of interest and its recorded proxy. Thus, in these settings, addressing measurement error requires counterfactual modeling of treatment effects on outcomes. In this work, we study intersectional threats to model reliability introduced by outcome measurement error, treatment effects, and selection bias from historical decision-making policies. We develop an unbiased risk minimization method which, given knowledge of proxy measurement error properties, corrects for the combined effects of these challenges. We also develop a method for estimating treatment-dependent measurement error parameters when these are unknown in advance. We demonstrate the utility of our approach theoretically and via experiments on real-world data from randomized controlled trials conducted in healthcare and employment domains. As importantly, we demonstrate that models correcting for outcome measurement error or treatment effects alone suffer from considerable reliability limitations. Our work underscores the importance of considering intersectional threats to model validity during the design and evaluation of predictive models for decision support.

CCS CONCEPTS

• **Computing methodologies → Machine learning approaches**; **Model verification and validation**.

FAccT '23, June 12–15, 2023, Chicago, IL, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594101

KEYWORDS

algorithmic decision support, measurement, validity, causal inference, model evaluation

ACM Reference Format:
Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. 2023. Counterfactual Prediction Under Outcome Measurement Error. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3593013.3594101

1 INTRODUCTION

Algorithmic risk assessment instruments (RAIs) often target labels that imperfectly reflect the goals of experts and policymakers. For example, clinical risk assessments used to inform physician treatment decisions target future utilization of medical resources (e.g., cost, medical diagnoses) as a proxy for patient medical need [45, 46, 50]. Predictive models used to inform personalized learning interventions target student test scores as a proxy for learning outcomes [29]. Yet, these proxies are subject to *outcome measurement error* (OME) when they systematically differ from the target outcome of interest to domain experts. Unaddressed, OME can be highly consequential: models targeting poor proxies have been linked to misallocation of medical resources [50], unwarranted teacher firings [72], and over-policing of minority communities [7]. Given its prevalence and implications, increasing research focus has shifted to understanding and mitigating sources of statistical bias impacting proxy outcomes [15, 23, 24, 44, 47, 77].

However, prior work modeling outcome measurement error makes a critical assumption that the decision informed by the algorithm does not impact downstream outcomes. Yet this assumption is often unreasonable in decision support applications, where decisions constitute *interventions* that impact the policy-relevant target outcome *and its recorded proxy* [13]. For example, in clinical decision support, medical treatments act as risk-mitigating interventions designed to avert adverse health outcomes. However, in the process of selecting a treatment option, a physician will *also* influence measured proxies (e.g., medical cost, disease diagnoses) [45, 46, 50]. As a result, the measurement error characteristics of proxies can vary across the treatment options informed by an algorithm.

In this work, we develop a counterfactual prediction method that corrects for outcome measurement error, treatment effects, and selection bias in parallel. Our method builds upon *unbiased risk minimization* techniques developed in the label noise literature [11, 47, 52, 73]. Given knowledge of measurement error parameters, unbiased risk minimization methods recover an estimator for target outcomes by minimizing a surrogate loss over proxy outcomes. However, existing methods are not designed for *interventional settings* whereby decisions impact outcomes – a limitation that we show severely limits model reliability. Therefore, we develop an unbiased risk minimization technique designed for learning counterfactual models from observational data. We compare our approach against models that correct for OME or treatment effects in isolation by conducting experiments on semi-synthetic data from healthcare and employment domains [21, 40, 71]. Results validate the efficacy of our risk minimization approach and underscore the need to carefully vet measurement-related assumptions in consultation with domain experts. Our empirical results also surface systematic model failures introduced by correcting for OME or treatment effects in isolation. To our knowledge, our holistic evaluation is the first to examine how outcome measurement error, treatment effects, and selection bias interact to impact model reliability under controlled conditions.

We provide the following contributions: 1) We derive a problem formulation that models interactions between OME, treatment effects, and selection bias (§ 3); 2) We develop a novel approach for learning counterfactual models in the presence of OME (§ 4.1). We provide a flexible approach for estimating measurement error rates when these are unknown in advance (§ 4.2); 3) We conduct synthetic and semi-synthetic experiments to validate our approach and highlight reliability issues introduced by modeling OME or treatment effects in isolation (§ 5).

2 BACKGROUND AND RELATED WORK

2.1 AI functionality and validity concerns

Prior work has conducted detailed assessments of specific modeling issues [13, 15, 32, 35, 39, 77], which have been synthesized into broader critiques of AI validity and functionality [14, 55, 76]. Raji et al. [55] surface AI functionality harms in which models fail to achieve their purported goal due to systematic design, engineering, deployment, and communication failures. Coston et al. [14] highlight challenges related to value alignment, reliability, and validity that may draw the justifiability of RAIs into question in some contexts. We build upon this literature by studying *intersectional threats to model reliability* arising from outcome measurement error [30, 77], treatment effects [13, 54], and selection bias [32] in parallel.

2.2 Outcome measurement error

Modeling outcome measurement error is challenging because it introduces two sources of uncertainty: which error model is reasonable for a given proxy, and which specific error parameters govern the relationship between target and proxy outcomes under the *assumed* measurement model [30]. Popular error models studied in the machine learning literature include uniform [4, 74], class-conditional [44, 65], and instance-dependent [9, 78] structures of outcome misclassification. Work in algorithmic fairness has also studied settings in which measurement error varies across levels of a protected attribute [77], and proposed sensitivity analysis frameworks that are model agnostic[23].

Numerous statistical approaches have been developed for measurement error parameter estimation in the quantitative social sciences literature [6, 58]. Application of these approaches is tightly coupled with domain knowledge of the phenomena under study, as in biostatistics [28] or psychometrics [69]. To date, data-driven techniques for error parameter estimation have primarily been applied in the machine learning literature, which rely on key assumptions relating the target outcome of interest and its proxy [41, 44, 49, 64, 65, 79]. In this work, we build upon an existing *"anchor assumptions"* framework that estimates error parameters by linking the proxy and target outcome probabilities at specific instances [79]. In contrast to prior work, we provide a range of anchoring assumptions, which can be flexibly combined depending on which are reasonable in a specific algorithmic decision support (ADS) domain.

A Motivating Example. We illustrate the importance of considering interactions between OME and treatment effects by revisiting a widely known audit of an algorithm used to inform screening decisions for a high-risk medical care program [50]. This audit surfaced measurement error in a *"cost of medical care"* outcome targeted as a proxy for patient medical need. *Critically, the measurement error analysis performed by Obermeyer et al. [50] assumes that program enrollment status is independent of downstream cost and medical outcomes.*

Sample	FPR	FNR
Full population	0.37	0.38
Unenrolled	0.37	0.39
Enrolled	0.64	0.13

Yet our re-analysis shows that the *"cost of medical care"* proxy has a substantially higher false positive rate and lower false negative rate among program enrollees as compared to the full population (see Appendix A.1). This error rate discrepancy is consistent with enrollees receiving closer medical supervision (and as a result, greater costs), even after accounting for their underlying medical need. In this work, we show that failing to model the interactions between OME and treatment effects can introduce substantial model reliability challenges.

Natarajan et al. [47] propose a widely-adopted *unbiased risk minimization* approach for learning under noisy labels given knowledge of measurement error parameters [11, 52, 73]. This method constructs a surrogate loss $\tilde{\ell}$ such that the $\tilde{\ell}$-risk over proxy outcomes is equivalent to the ℓ-risk over target outcomes *in expectation*. Additionally, Natarajan et al. [47] show that the minimizer of $\tilde{\ell}$-risk over proxy outcomes is optimal with respect to target outcomes if ℓ is symmetric (e.g., Huber, logistic, and squared losses). In this work, we develop a novel variant of this unbiased risk minimization approach designed for settings with *treatment-conditional* OME.

2.3 Counterfactual prediction

Recent work has shown that counterfactual modeling is necessary when the decision informed by a predictive model serves as a risk-mitigating intervention [13]. Building off of this result, we argue

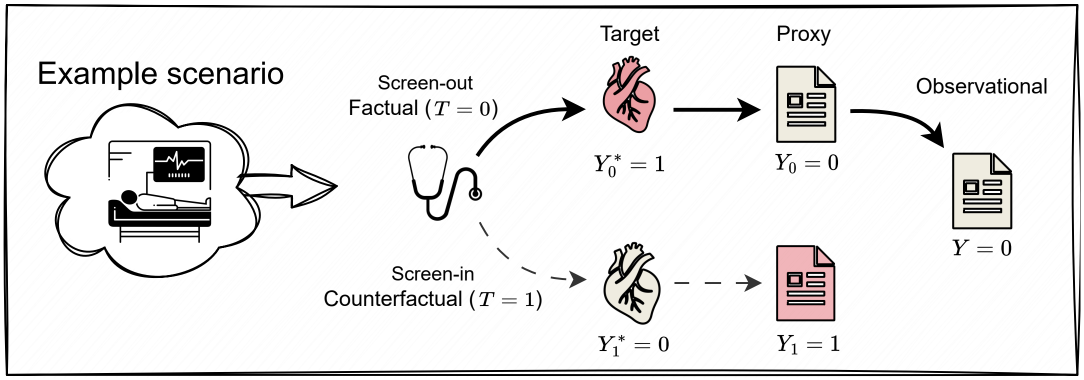

Figure 1: An illustration of treatment-conditional OME in heart attack prediction. Under the factual decision to screen-out from a high-risk care management program ($T = 0$), heart attack occurred ($Y_0^* = 1$) but went undiagnosed ($Y_0 = 0$). Under the counterfactual decision to screen in ($T = 1$), heart attack *would have* been averted ($Y_1^* = 0$) but would have been incorrectly diagnosed ($Y_1 = 1$). The observed outcome in medical records reflects the proxy value under factual decision to screen-out ($Y = 0$).

that it is necessary to account for treatment effects on *target outcomes of interest and their observed proxy* while modeling OME. Our methods build upon conditional average treatment effect (CATE) estimation techniques from the causal inference literature [1, 31, 66]. Subject to identification conditions [53, 62], these approaches predict the difference between the expected outcome under treatment (e.g., high-risk program enrollment) versus control (e.g., no program enrollment) conditional on covariates. One family of *outcome regression estimators* predicts the CATE by directly estimating the expected outcome under treatment or control conditional on covariates [10, 27, 38]. However, these methods suffer from statistical bias when prior decisions were non-randomized (i.e., due to distribution shift induced by selection bias) [4, 68]. Therefore, we leverage a re-weighting strategy proposed by [31] to correct for this selection bias during risk minimization. Our re-weighting method performs a similar bias correction as inverse probability weighting (IPW) methods [60, 68].

Outcome measurement error has also been studied in causal inference literature. Finkelstein et al. [22] bound the average treatment effect (ATE) under multiple plausible OME models. Shu and Yi [70] propose a doubly robust method which accounts for measurement error during ATE estimation, while Díaz and van der Laan [18] provide a sensitivity analysis framework for examining robustness of ATE estimates to OME. This work is primarily concerned with estimating *population statistics* rather than predicting outcomes conditional on measured covariates (i.e., the CATE).

3 PRELIMINARIES

Let $p^*(X, T, Y_0^*, Y_1^*, Y_0, Y_1)$ be a fixed joint distribution over covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$, past decisions[1] $T \in \{0, 1\}$, *target* potential outcomes $\{Y_0^*, Y_1^*\} \in \mathcal{Y} \subseteq \{0, 1\}$, and *proxy* potential outcomes $\{Y_0, Y_1\} \in \mathcal{Y} \subseteq \{0, 1\}$. Under the potential outcomes framework [62], $\{Y_0^*, Y_0\}$ and $\{Y_1^*, Y_1\}$ are the target and proxy outcomes that *would occur* under $T = 0$ and $T = 1$, respectively (Figure 1). Building upon the class-conditional model studied in observational settings [44, 47], we propose a treatment-conditional outcome measurement error model, whereby the class probability of the proxy potential outcome is given by

$$\eta_t(x) = (1 - \beta_t) \cdot \eta_t^*(x) + \alpha_t \cdot (1 - \eta_t^*(x)), \quad \forall x \in X \tag{1}$$

where $\alpha_t := p(Y_t = 1 \mid Y_t^* = 0)$, $\beta_t := p(Y_t = 0 \mid Y_t^* = 1)$ are the proxy false positive and false negative rates under treatment $t \in \{0, 1\}$ such that $\alpha_t + \beta_t < 1$. This model imposes the following assumption on the structure of measurement error.

Assumption 1 (Measurement error). Measurement error rates are fixed across covariates: $Y \perp\!\!\!\perp X \mid Y^*, T$.

While we make this assumption to foreground study of treatment effects, our methods are also compatible with approaches designed for error rates that vary across covariates [77] (see § 6.1 for discussion). Given the joint p^*, we would like to estimate $\eta_t^*(x) := p(Y_t^* = 1 \mid X = x)$, for any target covariates $x \in X$, which is the probability of the target potential outcome under intervention $t \in \{0, 1\}$. However, rather than observing Y_t^* directly, we sample from an *observational distribution* $p(X, T, Y)$, where $Y \in \mathcal{Y} \subseteq \{0, 1\}$ is an observed *proxy outcome*. By consistency, the unobserved target potential outcome and observed proxy potential outcome is determined by the treatment assignment.

Assumption 2 (Consistency). $Y^* = T \cdot Y_1^* + (1 - T) \cdot Y_0^*$; $Y = T \cdot Y_1 + (1 - T) \cdot Y_0$.

This assumption holds that the target and proxy potential outcomes Y_t^*, Y_t are observed among instances assigned to treatment t [53, 61, 62]. To identify observational proxy outcomes Y, we also require the following additional causal assumptions.

Assumption 3 (Ignorability). $\{Y_0^*, Y_1^*, Y_0, Y_1\} \perp\!\!\!\perp T \mid X$. This holds that target and proxy potential outcomes are unconfounded given measured covariates X.

Ignorability can be violated in decision support applications when unobservables impact both the treatment and outcome [15,

[1] We also use the word *treatments* to refer to binary decisions. This draws upon historical applications of causal inference to medical settings.

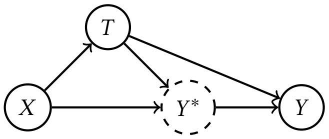

Figure 2: A causal diagram of treatment-conditional outcome measurement error.

35, 39]. Understanding and addressing limitations introduced by ignorability is a major ongoing research focus [13, 18, 56]. We provide follow-up discussion of this assumption in § 6.2.

Assumption 4 (Positivity)**.** $\forall x \in X,\ 0 > p(T = 1|X = x) > 1$. This holds that each instance $x \in X$ has some chance of receiving each decision $t \in \{0, 1\}$.

Positivity is often reasonable in decision support applications because instances $x \in X$ that require support from predictive models are subject to discretionary judgement due to uncertainty. Instances that are certain to receive a given treatment (i.e., $p(T = 1|X = x) = 0$ or $p(T = 1|X = x) = 1$) would normally be routed via a different administrative procedure. Figure 2 shows a causal diagram representing the data generating process we study in this work.

4 METHODOLOGY

We begin by developing an unbiased risk minimization approach which recovers an estimator for η_t^* given knowledge of error parameters (§ 4.1). We then provide a method for estimating α_t and β_t when error parameters are unknown in advance (§ 4.2).

4.1 Unbiased risk minimization

In this section, we develop an approach for estimating η_t^* given observational data drawn from $p(X, T, Y)$ and measurement error parameters α_t, β_t. Let $f_t \in \mathcal{H}$ for $\mathcal{H} \subset \{f_t : \mathcal{X} \rightarrow [0, 1]\}$ be a probabilistic decision function targeting Y_t^* and let $\ell : \mathcal{Y} \times [0, 1] \rightarrow \mathbb{R}_+$ be a loss function. If we observed target potential outcomes $Y_t^* \sim p^*$, we could directly apply supervised learning techniques to minimize the expected ℓ-risk of f_t over target potential outcomes

$$R_\ell^*(f_t) := \mathbb{E}_{p^*}[\ell(f_t(X), Y_t^*)] \tag{2}$$

and learn an estimator for η_t^* via standard empirical risk minimization approaches. Given a *strongly proper composite* loss such that $\arg\min_{f_t} R_\ell^*(f_t)$ is a monotone transform ψ of η_t^* (e.g., the logistic and exponential loss), this would enable recovering class probabilities from the optimal prediction via the link function ψ [2, 44]. However, directly minimizing (2) is not possible in our setting because we sample observational proxies instead of target potential outcomes. We address this challenge by constructing a *re-weighted surrogate risk* $R^w_{t,\tilde{\ell}}$ such that evaluating this risk over observed proxy outcomes is equivalent to R_ℓ^* in expectation.

In particular, let $w : \mathcal{X} \rightarrow \mathbb{R}_+$ be a weighting function satisfying $\mathbb{E}_X[w(X)|T = t] = 1$ and let $\ell : \mathcal{Y} \times [0, 1] \rightarrow \mathbb{R}_+$ be a surrogate loss function. We construct a *re-weighted surrogate risk*

$$R^w_{t,\tilde{\ell}}(f_t) := \mathbb{E}_p\left[w(X)\tilde{\ell}(f_t(X), Y) \mid T = t\right] \tag{3}$$

such that $R_\ell^*(f_t) = R_{t,\tilde{\ell}}(f_t)$ in expectation. Theorem 4.1 shows that we can recover a surrogate risk satisfying this property by constructing $w(x)$ as in (4) and $\tilde{\ell}$ as in (5). Note that this surrogate risk requires knowledge of α_t, β_t.

Theorem 4.1. *Assume treatment-conditional error (1), consistency (2), ignorability (3) and positivity (4). Then under target intervention $t \in \{0, 1\}$, $R_\ell^*(f_t) = R^w_{t,\tilde{\ell}}(f_t)$ for the weighting function $w : \mathcal{X} \rightarrow \mathbb{R}_+$ given by*

$$w(x) := \frac{p(T = t)}{(2t - 1) \cdot \pi(x) + 1 - t} \tag{4}$$

and surrogate loss $\tilde{\ell} : \mathcal{Y} \times [0, 1] \rightarrow \mathbb{R}_+$ given by

$$\begin{aligned} \tilde{\ell}(f_t(x), 1) &:= \frac{(1 - \alpha_t) \cdot \ell(f_t(x), 1) - \beta_t \cdot \ell(f_t(x), 0)}{1 - \beta_t - \alpha_t} \\ \tilde{\ell}(f_t(x), 0) &:= \frac{(1 - \beta_t) \cdot \ell(f_t(x), 0) - \alpha_t \cdot \ell(f_t(x), 1)}{1 - \beta_t - \alpha_t} \end{aligned} \tag{5}$$

where in (4), $\pi(x) := p(T = 1|X = x)$ is the propensity score function.

We prove Theorem 4.1 in Appendix A.2. Intuitively, $R^w_{t,\tilde{\ell}}(f_t)$ applies a *joint bias correction* for OME and distribution shift introduced by historical decision-making policies (i.e., selection bias). The unbiased risk minimization framework dating back to Natarajan et al. [47] corrects for OME by minimizing a surrogate loss $\tilde{\ell}$ on proxies Y observed *over the full population unconditional on treatment.* Yet this approach is untenable when decisions impact outcomes ($T \not\perp\!\!\!\perp \{Y^*, Y\}$) and error rates differ across treatments. One possible extension of unbiased risk minimizers to the treatment-conditional setting involves minimizing $\tilde{\ell}$ over the treatment population $p(X|T = t)$

$$R_{t,\tilde{\ell}}(f_t) := \mathbb{E}_p\left[\tilde{\ell}(f_t(X), Y) \mid T = t\right]. \tag{6}$$

However, $R_{t,\tilde{\ell}} \neq R_\ell^*$ in observational settings because the treatment population $p(X|T = t)$ can differ from the marginal population $p(X)$ under historical selection policies when $X \not\perp\!\!\!\perp T$. Therefore, our re-weighting procedure applies a second bias correction that adjusts $p(X|T = t)$ to resemble $p(X)$.

Learning algorithm. As a result of Theorem 4.1, we can learn a predictor $\hat{\eta}_t^*$ by minimizing the re-weighted surrogate risk over *observed samples* $(X_1, T_1, Y_1), ..., (X_n, T_n, Y_n) \sim p$. First, we estimate the weighting function $\hat{w}(x)$ through a finite sample, which boils down to learning propensity scores $\hat{\pi}(x)$ (as shown in (4)). Estimating the propensity scores can be done by applying supervised learning algorithms to learn a predictor from X to T. Then for any treatment t, weighting function $\hat{w}$, and predictor f_t, we can approximate $R^w_{t,\tilde{\ell}}(f_t)$ by taking the sample average over the treatment population

$$\hat{R}^{\hat{w}}_{t,\tilde{\ell}}(f_t) := \frac{1}{n_t} \sum_{i:T_i = t} \hat{w}(X_i)\tilde{\ell}(f_t(X_i), Y_i) \tag{7}$$

for $n_t = \sum_{i=1}^n \mathbb{1}[T_i = t]$. Therefore, given $\hat{w}$ we can learn a predictor from observational data by minimizing the empirical risk

$$\hat{f}_t \leftarrow \arg\min_{f_t \in \mathcal{H}} \hat{R}^{\hat{w}}_{t,\tilde{\ell}}(f_t). \tag{8}$$

We refer to solving (8) as *re-weighted risk minimization with a surrogate loss* (Algorithm 1).

Algorithm 1: Re-weighted risk minimization with surrogate loss (RW-SL)

```
Input: Data $\mathcal{W} = \{(X_i, T_i, Y_i)\}_{i=1}^n \sim p$
Output: Learned estimator $\hat{\eta}^*_t(x)$
Partition $\mathcal{W}$ into $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$
On $\mathcal{W}_1$, estimate parameters $\hat{\alpha}_t, \hat{\beta}_t \leftarrow \text{CCPE}(\mathcal{W}_1)$
On $\mathcal{W}_2$, learn $\hat{\pi}(x)$ by regressing $T \sim X$
On $\mathcal{W}_3$, use $\hat{\pi}(x), \hat{\alpha}_t, \hat{\beta}_t$ to solve
  $\hat{\eta}^*_t(x) \leftarrow \arg\min_{f_t \in \mathcal{H}} \hat{R}^{\hat{w}}_{t,\hat{\ell}}(f_t)$
```

Algorithm 2: Conditional class probability estimation (CCPE)

```
Input: Data $\mathcal{V} \sim p$
Output: Parameter estimates $\hat{\alpha}_t, \hat{\beta}_t$
Partition $\mathcal{V}$ into $\mathcal{V}_1, \mathcal{V}_2$
On $\mathcal{V}_1$, learn $\hat{\eta}_t(x)$ by regressing $Y \sim X \mid T = t$
On $\mathcal{V}_2$, estimate error parameters:
  $\hat{\alpha}_t = \min_{x \in X}\{\hat{\eta}_t(x)\}, \quad \hat{\beta}_t = 1 - \max_{x \in X}\{\hat{\eta}_t(x)\}$
```

4.2 Error parameter identification and estimation

Because our risk minimization approach requires knowledge of OME parameters, we develop a method for estimating α_t, β_t from observational data. Error parameter estimation is challenging in decision support applications because target outcomes often result from nuanced social and organizational processes. Understanding the measurement error properties of proxies targeted in criminal justice, medicine, and hiring domains remains an ongoing focus of domain-specific research [3, 8, 24, 46, 50, 80]. *Therefore, we develop an approach compatible with multiple sources of domain knowledge about proxies, which can be flexibly combined depending on which assumptions are deemed reasonable in a specific context.*

Error parameters are *identifiable* if they can be uniquely computed from observational data. Because our error model (e.q. 1) expresses the proxy class probability as a linear equation with two unknowns, α_t, β_t are identifiable if the target class probability $c^*_{t,i} = \eta^*_t(x_i)$ and proxy class probability $c_{t,i} = \eta_t(x_i)$ are known at two distinct points $(c^*_{t,i}, c_{t,i})$ and $(c^*_{t,j}, c_{t,j})$ such that $c^*_{t,i} \neq c^*_{t,j}$. Following prior literature [26], we refer to knowledge of $(c^*_{t,i}, c_{t,i})$ as an *anchor assumption* because it requires knowledge of the unobserved quantity η^*_t. We now introduce several anchor assumptions that are practical in ADS, before showing that these can be flexibly combined to identify α_t, β_t in Theorem 4.2.

Min anchor. A min anchor assumption holds if there is an instance at no risk of the target potential outcome under intervention t: $c^*_{t,i} = \inf_{x_i \in \mathcal{X}} \{\eta^*_t(x_i)\} = 0$. Because η_t is a strictly monotone increasing transform of η^*_t, the corresponding value of η_t can be recovered via $c_{t,i} = \inf_{x_i \in \mathcal{X}}\{\eta_t(x_i)\}$ [44]. Min anchors are reasonable when there are cases that are confirmed to be at no risk based on domain knowledge of the data generating process. For example, a min anchor may be reasonable in diagnostic testing if a patient is confirmed to be negative for a medical condition based on a high-precision gold standard medical test [19].

	Know α_t	**Min**	**Base rate**	**Max**	**Know β_t**
Know α_t	×	×	✓	✓	✓
Min	×	×	✓	✓	✓
Base rate	✓	✓	×	✓	✓
Max	✓	✓	✓	×	×
Know β_t	✓	✓	✓	×	×

Table 1: Multiple combinations of min, max, and base rate anchor assumptions (shown via ✓) enable identification of α_t, β_t.

Max anchor. A max anchor assumption holds if there is an instance at certain risk of the target outcome under intervention t: $c^*_{t,i} = \sup_{x_i \in \mathcal{X}}\{\eta^*_t(x_i)\} = 1$. The corresponding value of η_t can be recovered via $c_{t,i} = \sup_{x_i \in \mathcal{X}}\{\eta_t(x_i)\}$ because η_t is a strictly monotone increasing transform of η^*_t. Max anchors are reasonable when there are confirmed instances of a positive target potential outcome based on domain knowledge of the data generating process. For example, a max anchor may be justified in a medical setting if a subset of patients have confirmed disease diagnoses based on biopsy results [5], or if a disease prognosis (and resulting health outcomes) are known from pathology.

Base rate anchor. A base rate anchor assumption holds if the expected value of η^*_t is known under intervention t: $c^*_{t,i} = \mathbb{E}[\eta^*_t(X)]$. The corresponding value of η_t can be recovered by taking the expectation over the proxy class probability $c_{t,i} = \mathbb{E}[\eta_t(X)]$. Base rate anchors are practical because the prevalence of unobservable target outcomes (e.g., medical conditions [75], crime [37, 42], student performance [17, 63]) is routinely estimated via domain-specific analyses of measurement error. For instance, studies have been conducted to estimate the base rate of undiagnosed heart attacks (i.e., accounting for measurement error in diagnosis proxy outcomes) [51]. Additionally, the conditional average treatment effect $\mathbb{E}[\eta^*_1(X)] - \mathbb{E}[\eta^*_0(X)]$ is commonly estimated in randomized controlled trials (RCTs) while assessing treatment effect heterogeneity [27]. While the conditional average treatment effect is normally estimated via proxies Y_0 and Y_1, measurement error analysis is a routine component of RCT design and evaluation [25].

Anchor assumptions can be flexibly combined to identify error parameters based on which set of assumptions are reasonable in a given ADS domain. In particular, Theorem 4.2 shows that combinations of anchor assumptions listed in Table 1 are sufficient for identifying error parameters under our causal assumptions.

Theorem 4.2. *Assume treatment-conditional error (1), consistency (2), ignorability (3) and positivity (4). Then α_t, β_t are identifiable from observational data $p(X, T, Y)$ given any identifying pair of anchor assumptions provided in Table 1.*

We prove Theorem 4.2 in Appendix A.2. In practice, we estimate the error rates on finite samples $(X_i, T_i, Y_i) \sim p$, which gives an approximation $\hat{\eta}_t$. Therefore, we propose a conditional class probability estimation (CCPE) method for parameter estimation which estimates $\hat{\alpha}_t, \hat{\beta}_t$ by fitting $\hat{\eta}_t$ on observational data then applying the relevant pair of anchor assumptions to estimate error rates. Algorithm 2 provides pseudocode for this approach with min and max anchors, which can easily be extended to other pairs of identifying assumptions shown in Table 1. The combination of min and max anchors is known as *weak separability* [44] or *mutual irreducibility* [64, 65] in the observational label noise literature. Prior results in the observational setting show that unconditional class probability estimation (i.e., fitting $\hat{\eta}(x) = p(Y = 1|X = x)$) yields a consistent estimator for observational error rates under weak seperability [57, 65]. Statistical consistency results extend to the treatment-conditional setting under positivity (4) because $p(T = t|X = x) > 0,\ \forall t \in \{0, 1\},\ x \in \mathcal{X}$. However, asymptotic convergence rates may be slower under strong selection bias if $p(T = t|X = x)$ is near 0.

5 EXPERIMENTS

Experimental evaluation under treatment-conditional OME is challenging due to compounding sources of uncertainty. We do not observe counterfactual outcomes in historical data, making it challenging to estimate the quality of new models via observational data. Further, because the target outcome is not observed directly, we rely on measurement assumptions when studying proxy outcomes in naturalistic data. We address this challenge by conducting a controlled evaluation with synthetic data where ground truth potential outcomes are fully observed. To better reflect the ecological settings of real-world deployments, we also conduct a semi-synthetic evaluation with real data collected through randomized controlled trials (RCTs) in healthcare and employment domains. Our evaluation (1) validates our proposed risk minimization approach, (2) underscores the need to carefully consider measurement assumptions during error rate estimation, and (3) shows that correcting for OME or treatment effects in isolation is insufficient.[2]

5.1 Models

We compare several modeling approaches in our evaluation to examine how existing modeling practices are impacted by treatment-conditional outcome measurement error:

- **Unconditional proxy (UP)**. Predict the observed outcome unconditional on treatment: $X \rightarrow Y$. This model *does not adjust for OME or treatment effects.*, and reflects model performance in a scenario in which practitioners overlook all challenges examined in this work. [3]
- **Unconditional target (UT)**. Predict the target outcome unconditional on treatment: $X \rightarrow Y^*$. Here, we determine Y^* by applying consistency: $Y^* = (1-T) \cdot Y_0^* + T \cdot Y_1^*$. This method reflects a setting in which no OME is present but modeling does not account for treatment effects [44, 47, 50, 77].
- **Conditional proxy (CP).** Predict the proxy outcome conditional on treatment: $X, T \rightarrow Y$. This is a counterfactual model that estimates a conditional expectation *without correcting for OME* [13, 38, 66].[4]
- **Re-weighted surrogate loss (RW-SL).** Our proposed risk minimization approach, as defined in equation (8). This method corrects for both OME and treatment effects in parallel. Additionally, this method corrects for distribution shift due to selection bias in the prior decision-making policy via reweighting.
- **Target Potential Outcome (TPO).** Directly predict the target potential outcome: $X \rightarrow Y_t^*$. This model is an *oracle* that provides an upper-bound on model performance under no OME or treatment effects.

We also perform an ablation of our proposed RW-SL method by including a model that applies a surrogate loss correction $\tilde{\ell}$ over the treatment population without re-weighting (**SL**).

5.2 Experiments on synthetic data

We begin by experimentally manipulating treatment effects and measurement error via a synthetic evaluation. Because this provides full control over the data generating process, we can evaluate methods against target potential outcomes. This evaluation would not possible with real-world data because counterfactual target outcomes are unobserved. Our experiment design is consistent with prior synthetic evaluations of counterfactual risk assessments [13] and causal inference methods [48, 67]. In our evaluation, we sample outcomes via the following data generating process:

(1) $Y_t^* := \sim \text{Bern}(\eta_t^*(X)),\ \forall t \in \{0, 1\}$

(2) $Y_t := \begin{cases} 1 - \epsilon_+ & \text{if } Y_t^* = 1, \text{ where } \epsilon_+ \sim \text{Bern}(\beta_t) \\ \epsilon^- & \text{if } Y_t^* = 0, \text{ where } \epsilon_- \sim \text{Bern}(\alpha_t) \end{cases},\ \forall t \in \{0, 1\}$

(3) $T := \sim \text{Bern}(\pi(X))$

(4) $Y^* := (1 - T) \cdot Y_0^* + T \cdot Y_1^*;\ Y := (1 - T) \cdot Y_0 + T \cdot Y_1$

As shown in Figure 3, we draw $X \sim U(-1, 1)$ and sample target potential outcomes from sinusoidal class probability functions (see Appendix A.4 for details). Note that our choice of $\eta_0^*(x)$, $\eta_1^*(x)$ satisfies min and max anchor assumptions. Because $\eta_0^*(x)$ and $\eta_1^*(x)$ differ, models that do not condition on treatment (i.e., UP, UT) will learn an average of the two class probability functions. Under our choice of $\pi(x)$, fewer samples are drawn from $\eta_1^*(x)$ in the region where $\pi(x)$ is small (near $x = -1$), and fewer samples are drawn from $\eta_0^*(x)$ in the region where $1 - \pi(x)$ is small (near $x = 1$). This introduces selection bias when sampling from $\pi(x)$.

5.2.1 Setup details. We train each model in § 5.1 to predict risk under no intervention ($t = 0$) and vary (α_0, β_0). We keep (α_1, β_1) fixed at $(0, 0)$ across settings. When estimating OME parameters, we run CCPE with sample splitting and cross-fitting (Algorithm 4) with min and max anchor assumptions for identification. These assumptions hold precisely under this controlled evaluation (Figure 3). We run all methods with sample splitting and cross-fitting (Algorithm A.3) and report performance on Y_0^*.

5.2.2 Results. Figure 4 shows the performance of each model as a function of sample size. TPO provides an upper bound on performance because it learns directly from target potential outcomes. RW-SL with oracle parameters (α, β) outperforms all other methods trained on observational data across across the full range of sample sizes. Thus, while Thm. 4.1 shows that RW-SL recovers an unbiased risk estimator *in expectation*, this method also demonstrates favorable finite-sample performance characteristics in practice. This finding is inline with prior experimental evaluations of unbiased risk estimators reported in the standard supervised learning setting

[2] Code for all experiments can be found at: https://github.com/lguerdan/CP_OME.

[3] This baseline is also called an *observational risk assessment* in experiments reported by Coston et al. [13].

[4] This model is known by different names in the causal inference literature, including the backdoor adjustment (G-computation) formula [53, 59], T-learner [38], and plug-in estimator [34].

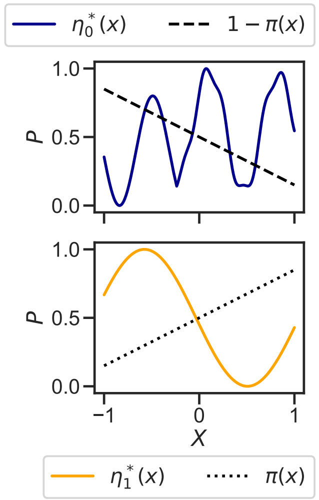

Figure 3: Synthetic setup.

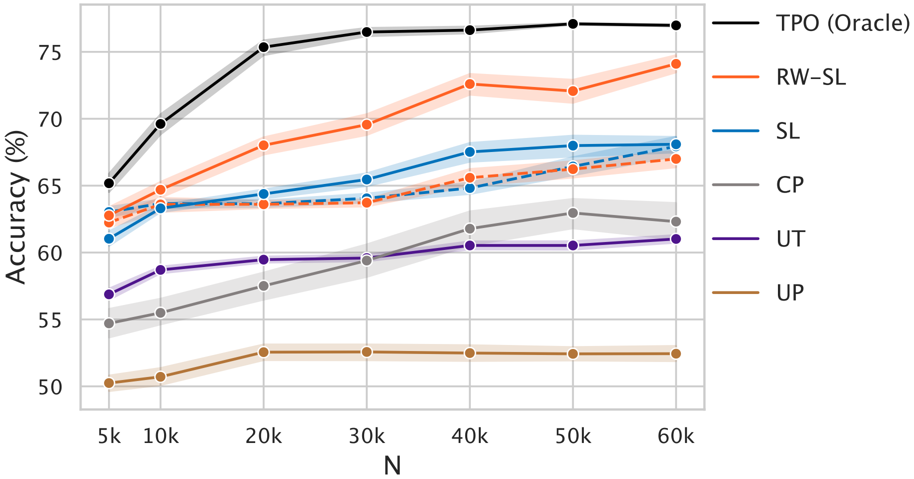

Figure 4: Accuracy on Y_0^* as a function of sample size. RW-SL and SL with oracle parameters plotted with solid lines. RW-SL and SL with learned parameters plotted with dashed lines. Results averaged over asymmetric error settings reported in Table 2.

(α_0, β_0)	(0.0, 0.4)	(0.1, 0.3)	(0.2, 0.2)	(0.3, 0.1)	(0.4, 0.0)
UP	54.18 (0.09)	53.00 (0.39)	54.89 (1.09)	55.81 (0.74)	46.76 (0.33)
UT	61.57 (0.63)	60.95 (0.50)	60.49 (0.41)	61.00 (0.49)	60.54 (0.70)
CP	51.36 (1.83)	68.24 (2.61)	**75.05** (0.92)	67.77 (1.33)	61.88 (0.28)
SL $(\hat{\alpha}, \hat{\beta})$	72.38 (1.65)	65.45 (0.66)	67.43 (1.64)	68.01 (0.99)	65.92 (1.34)
RW-SL $(\hat{\alpha}, \hat{\beta})$	69.08 (1.55)	65.96 (1.18)	66.57 (1.32)	68.39 (1.33)	64.56 (0.52)
SL (α, β)	67.09 (1.24)	67.58 (1.19)	67.75 (1.08)	69.11 (1.17)	68.59 (1.41)
RW-SL (α, β)	**73.68** (1.49)	**73.39** (1.60)	72.52 (1.66)	**74.34** (1.15)	**75.01** (1.24)
TPO	**77.08** (0.11)	**77.09** (0.20)	**76.98** (0.08)	**76.84** (0.18)	**76.90** (0.16)

Table 2: Model accuracy (s.e.) across error parameter settings (α_0, β_0) at $N = 60k$ samples over 10 runs. Top-2 performance across each (α_0, β_0) setting shown in bold.

[47, 77], and is further supported by reliable performance characteristics we observe in small sample regimes (see Appendix A.4).

In contrast, both models that do not condition on treatment (UP and UT), and the conditional regression trained on proxy outcomes (CP), reach a performance plateau by 50k samples and do not benefit from additional data. This indicates that (1) learning a counterfactual model and (2) correcting for measurement error is necessary to learn η_t^* in this evaluation. We likely observe a sharper plateau in UP and UT above 20k samples because these approaches fit a weighted average of η_0^* and η_1^* (where η_1^* differs from η_0^* considerably). We observe that RW-SL and SL performance deteriorates with learned parameters $(\hat{\alpha}, \hat{\beta})$ across all sample size settings due to misspecification in learned parameter estimates and weights.

Table 2 shows a breakdown across error rates (α_0, β_0) at $60k$ samples. RW-SL outperforms SL when oracle parameters are known. However, RW-SL and SL perform comparably when weights and parameters are learned. This may be because RW-SL relies on estimates $\hat{w}$ in addition to $\hat{\alpha}_0, \hat{\beta}_0$, which could introduce instability given misspecification in $\hat{w}$. CP performs notably well under high error parameter symmetry (i.e., $\alpha_0 = \beta_0 = .2$). This is consistent with prior results from the class-conditional label noise literature [44, 47], which show that the optimal classifier threshold for misclassification risk does not change under symmetric label noise. CP performs worse under high error asymmetry. We do not observe a similar performance improvement in UP and UT in the symmetric error setting because these baselines learn a weighted combination of η_0 and η_1, which differs from the target function η_0^* at all classification thresholds.

5.3 Semi-synthetic experiments on healthcare and employment data

In addition to our synthetic evaluation, we conduct experiments using real-world data collected as part of randomized controlled trials (RCTs) in healthcare and employment domains. While this evaluation affords less control over the data generating process, it provides a more realistic sample of data likely to be encountered in real-world model deployments. Evaluation via data from randomized or partially randomized experimental studies is useful for validating counterfactual prediction approaches because random assignment ensures that causal assumptions are satisfied [12, 31, 66].

5.3.1 Randomized Controlled Trial (RCT) Datasets. In 2008, the U.S. state of Oregon expanded access to its Medicare program via a lottery system [21]. This lottery provided an opportunity to study the effects of Medicare enrollment on healthcare utilization and medical outcomes via an experimental design, commonly referred to as the Oregon Health Insurance Experiment (OHIE). Lottery enrollees completed a pre-randomization survey recording demographic factors and baseline health status and were given a one-year follow-up assessment of health status and medical care utilization. We refer the reader to Finkelstein et al. [21] for details. We use the OHIE dataset to construct an evaluation task that parallels the label choice bias analysis of Obermeyer et al. [50]. We use this dataset rather than synthetic data released by Obermeyer et al. [50] because (1) treatment was randomly assigned, ruling out positivity and ignorability violations possible in observational data, and (2) OHIE data contains covariates necessary for predictive modeling. We predict diagnosis with an active chronic medical condition over the one-year follow-up period given $D = 58$ covariates, including health history, prior emergency room visits, and public services use. We predict chronic health conditions because findings from Obermeyer et al. [50] indicate that this outcome variable is a reasonable choice of proxy for patient medical need. We adopt the randomized lottery draw as the treatment. [5]

We conduct a second real-world evaluation using the JOBS dataset, which investigates the effect of job retraining on employment status [66]. This dataset includes an experimental sample collected by LaLonde [40] via the National Supported Work (NSW) program (297 treated, 425 control) consisting primarily of low-income individuals seeking job retraining. Smith and Todd [71] combine this sample with a "PSID" comparison group (2,490 control) collected from the general population, which resulted in a final sample with 297 treated and 2,915 control. This dataset includes $D = 17$ covariates including age, education, prior earnings, and interaction terms. 482 (15%) of subjects were unemployed at the end of the study. Following Johansson et al. [31], we construct an evaluation task predicting unemployment under enrollment ($t = 1$) and no enrollment ($t = 0$) in a job retraining program conditional on covariates.

5.3.2 Synthetic OME and selection bias. We experimentally manipulate OME to examine how outcome regressions perform under treatment-conditional error of known magnitude. We adopt diagnosis with a new chronic condition and future unemployment as a *target outcome* for OHIE and JOBS, respectively. We observe proxy outcomes by flipping target outcomes with probability (α_0, β_0). We keep (α_1, β_1) fixed at $(0, 0)$. This procedure of generating proxy outcomes by flipping available labels is a common approach for vetting the feasibility of new methodologies designed to address OME [44, 47, 77]. This approach offers precise control over the magnitude of OME but suffers from less ecological validity than studying multiple naturalistic proxies [50]. We opt for this semi-synthetic evaluation because (1) the precise measurement relationship between naturally occurring proxies may not be fully known, (2) the measurement relationship between naturally occurring proxies cannot be manipulated experimentally, and (3) there are few RCT datasets (i.e., required to guarantee causal assumptions) that contain multiple proxies of the same target outcome.

Models used for decision support are typically trained using data gathered under a historical decision-making policy. When prior decisions were made non-randomly, this introduces selection bias ($T \not\perp X$) and causes distribution shift between the population that received treatment t in training data, and the full population encountered at deployment time. Therefore, we emulate selection bias in the *training dataset*, and evaluate models over a held-out test set of randomized data. We insert selection bias in OHIE data by removing individuals from the treatment (lottery winning) arm with household income above the federal poverty line (10% of the treatment sample). This resembles an observational setting in which low-income individuals are more likely to receive an opportunity to enroll in a health insurance program (e.g., Medicaid, which determines eligibility based on household income in relation to the federal poverty line). We restrict our analysis to single-person households, yielding $N = 12,994$ total samples ($6,189$ treatment, $6,805$ control).

We model selection bias in JOBS data by including samples from the observational and experimental cohorts in the training data. Because the PSID comparison group consists of individuals with higher income and education than the NSW group, there is considerable distribution shift across the NSW and PSID cohorts [31, 40, 71]. Therefore, a model predicting unemployment over the control population (consisting of NSW and PSID samples) may suffer from bias when evaluated against test data that only includes samples from the NSW experimental arm. Thus we split data from the NSW experimental cohort 50-50 across training and test dataset, and only include PSID data in the training dataset.

5.3.3 Experimental setup. We include a Conditional Target (CT) model in place of a TPO model because counterfactual outcomes are not available in experimental data. CT provides a reasonable upper-bound on performance because identifiability conditions are satisfied in an experimental setting [53]. However, it is not possible to report accuracy over potential outcomes because counterfactual outcomes are unobserved. Therefore, we report error in ATE estimates $\tau - \hat{\tau}$, for

$$\tau := \mathbb{E}[Y^* \mid T = 1] - \mathbb{E}[Y^* \mid T = 0], \quad \hat{\tau} := \mathbb{E}[\hat{\eta}_1(X)] - \mathbb{E}[\hat{\eta}_0(X)]$$

where τ corresponds to the ground-truth treatment effect reported by prior work [16, 31] and $\hat{\eta}_t$ is a learned model discussed in § 5.1. One subtlety of this comparison is that our outcome regressions target the *conditional* average treatment effect, while τ reflects the ATE across the full population. Following prior evaluations [31], we compare all methods against the ATE because the ground-truth CATE is not available for JOBS or OHIE data. [6] We report results

[5] The OHIE experiment had imperfect compliance ($\approx$ 30 percent of selected individuals successfully enrolled [21]). Therefore, we predict diagnosis with a new chronic health condition given the *opportunity to enroll* in Medicare. This evaluation is consistent with many high-stakes decision-support settings granting opportunities to individuals, which they have a choice to pursue if desired.

[6] While our insertion of synthetic selection bias (§5.3.2) introduces distribution shift such that $p(X|T = 1)$ differs from $p(X|T = 0)$, it does not alter ground-truth

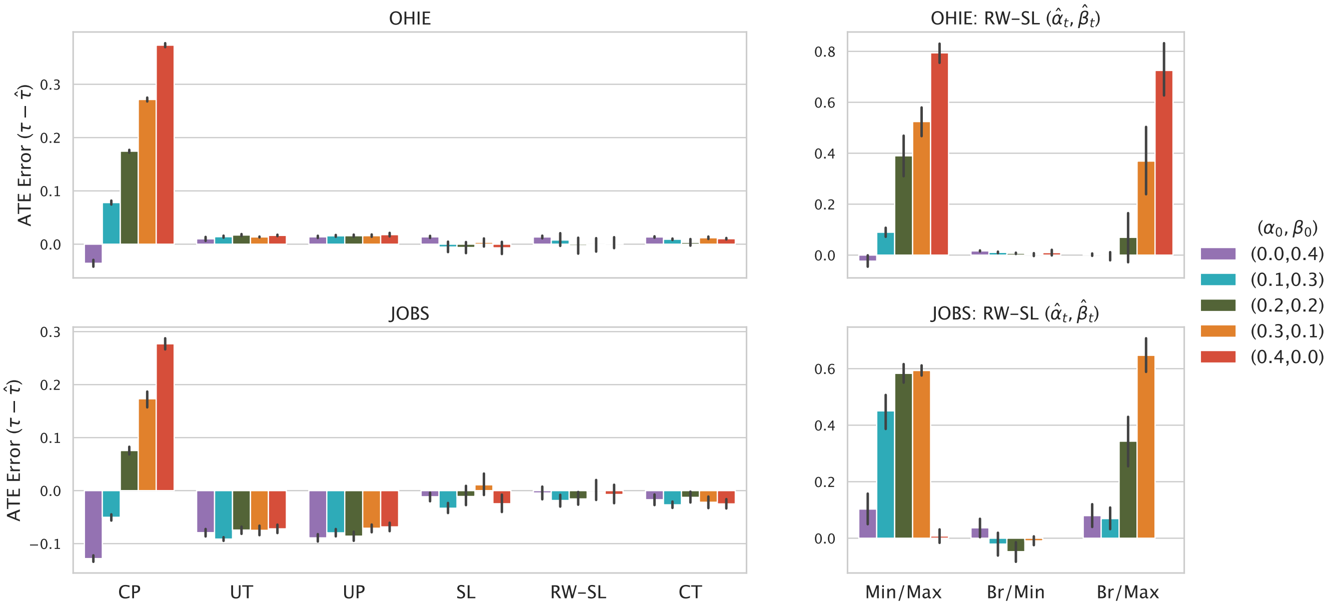

Figure 5: Bias in ATE estimates on OHIE and JOBS data. Error bars indicate standard error over ten runs. CT is a model with oracle access to target outcomes and RW-SL is our proposed approach.

over a test fold of randomized data that does not contain flipped outcomes or selection bias. Appendix A.4 contains additional setup details.

5.3.4 Results. Figure 5 shows bias in ATE estimates $\tau - \hat{\tau}$ over 10 runs on JOBS and OHIE data. The left panel compares CP, UT, UP, and the oracle CT model against RW-SL/SL with oracle parameters (α_0, β_0), (α_1, β_1). We show performance of RW-SL with learned parameters $(\hat{\alpha}_0, \hat{\beta}_0)$, $(\hat{\alpha}_1, \hat{\beta}_1)$ on the right panel. The left panel shows that CP is highly sensitive to measurement error. This is because measurement error introduces bias in estimates of the conditional expectations, which propagates to treatment effect estimates. Because UT and UP do not condition on treatment, they estimate an *average* of the outcome functions η_0^* and η_1^*, and generate predictions near 0. Therefore, while UT and UP perform well on OHIE data due to a small ground-truth ATE ($\tau = 0.015$), they perform poorly on JOBS ($\tau = -0.077$). SL and RW-SL with oracle parameters α_t, β_t perform comparably to the CT model with oracle access to target outcomes across all measurement error settings.

While we observe that re-weighting improves performance in our synthetic evaluation (given oracle parameters), we do not observe a similar advantage of RW-SL over SL in this experiment. Our results parallel other empirical evaluations of re-weighting for counterfactual modeling tasks on real-world data (e.g., see § 3.4.2 in [13]). One potential explanation for this finding is that our predictive model class (multi-layer MLPs) is large enough to learn the target regressions η_0^* and η_1^* for OHIE and JOBS data, even after our insertion of synthetic selection bias. As a result, re-weighting may not be required to learn a reasonable estimate of η_0^* and η_1^* given available data. This interpretation is supported by strong performance of the oracle CT model.

As shown on the right panel of Figure 5, RW-SL performance is highly sensitive to the choice of anchor assumption used to estimate parameters $(\hat{\alpha}_0, \hat{\beta}_0)$, $(\hat{\alpha}_1, \hat{\beta}_1)$ as indicated by increased bias in $\hat{\tau}$ and greater variability over runs. In particular, RW-SL performs poorly when Min/Max and Br/Max pairs of anchor assumptions are used to estimate error rates because the max anchor assumption is violated on OHIE and JOBS data. We shed further light on this finding by fitting the CT model to estimate $\hat{\eta}_0^*$, $\hat{\eta}_1^*$ on OHIE data, then computing inferences over a validation fold X_{val}. This analysis reveals that

$$\min_{x \in X_{val}} \hat{\eta}_0^* \approx 2.23 \cdot e^{-6}, \quad \max_{x \in X_{val}} \hat{\eta}_0^* \approx 0.85$$
$$\min_{x \in X_{val}} \hat{\eta}_1^* \approx 1.02 \cdot e^{-5}, \quad \max_{x \in X_{val}} \cdot \hat{\eta}_1^* \approx 0.81$$

which suggests that the min anchor assumption that $\min_{x \in X_{val}} \hat{\eta}_t^* = 0$ is reasonable for $t \in \{0, 1\}$, while the max anchor assumption that $\max_{x \in X_{val}} \hat{\eta}_t^* = 1$ is violated for both $t \in \{0, 1\}$. Therefore, because the min anchor assumption is satisfied for these choices of target outcome, and the ground-truth base rate is known in this experimental setting, RW-SL demonstrates strong performance given the Br/Min combination of anchor assumptions. In contrast, because the max anchor is violated, estimating β_t by taking the supremium of $\hat{\eta}_t(x)$ introduces bias in $\hat{\beta}_t$, which results in poor performance of RW-SL with Min/Max and Br/Max anchors. Applying this same procedure to the unemployment outcome targeted in JOBS data also reveals a violation of the max anchor assumption. These results underscore the importance of selecting anchor assumptions in close consultation with domain experts because it is not possible to verify

τ because the conditional outcome distribution $p(Y^*|T)$ remains unchanged. This setup recreates the unconfounded observational setting in which causal identification assumptions are satisfied [61].

anchor assumptions by learning $\hat{\eta}_t^*$ when the target outcome of interest is unobserved.

6 DISCUSSION

In this work, we show the importance of carefully addressing intersectional threats to model reliability during the development and evaluation of predictive models for decision support. Our theoretical and empirical results validate the efficacy of our unbiased risk minimization approach. When OME parameters are known, our method performs comparably to a model with oracle access to target potential outcomes. However, our results underscore the importance of vetting anchoring assumptions used for error parameter estimation before using error rate estimates for risk minimization. Critically, our experimental results also demonstrate that correcting for a single threat to model reliability in isolation is insufficient to address model validity concerns [55], and risks promoting false confidence in corrected models. Below, we expand upon key considerations surfaced by our work.

6.1 Decision points and complexities in measurement error modeling

Our work speaks to key complexities faced by domain experts, model developers, and other stakeholders while examining proxies in ADS. One decision surfaced by our work entails which *measurement error model* best describes the relationship between the unobserved outcome of policy interest and its recorded proxy. We open this work by highlighting a measurement model decision made by Obermeyer et al. [50] during their audit of a clinical risk assessment: that error rates are fixed across treatments. Our work suggests that failing to account for treatment-conditional error in OME models may exacerbate reliability concerns. However, at the same time, the error model we adopt in this work intentionally abstracts over other factors known to impact proxies in decision support tasks, including error rates that vary across covariates. Although this simplifying assumption can be unreasonable in some settings [3, 24], including the one studied by Obermeyer et al. [50], it is helpful in foregrounding previously-overlooked challenges involving treatment effects and selection bias. In practice, model developers correcting for measurement error may wish to combine our methods with existing unbiased risk minimization approaches designed for group-dependent error where appropriate [77]. Further, analyses of measurement error should not overlook more fundamental conceptual differences between target outcomes and proxies readily available for modeling (e.g., when long-term child welfare related outcomes targeted by a risk assessment differ from *immediate* threats to child safety weighted by social workers [33, 33]). This underscores the need to carefully weigh the validity of proxies in consultation with multiple stakeholders (e.g., domain experts, data scientists, and decision-makers) while deciding whether OME correction is warranted.

A second decision point highlighted in this work entails the *specific measurement error parameters* that govern the relationship between target and proxy outcomes. In particular, our work underscores the need for a tighter coupling between domain expertise and data-driven approaches for error parameter estimation. Current techniques designed to address OME in the machine learning literature – which typically examine settings with "label noise" – rely heavily upon data-driven approaches without close consideration of whether the underlying measurement assumptions hold [44, 49, 77, 79]. While application of these assumptions may be practical for methodological evaluations and theoretical analysis [57, 64, 65], these assumptions should be carefully vetted when applying OME correction to real-world proxies. This is supported by our findings in Figure 5, which show that RW-SL performs poorly when the anchor assumptions used for error parameter estimation are violated. Our flexible set of anchor assumptions provides a step towards a tighter coupling between domain expertise and data-driven approaches in measurement parameter estimation.

6.2 Challenges of linking causal and statistical estimands

Our counterfactual modeling approach requires several causal identifiability assumptions [53], which may not be satisfied in all decision support contexts. Of our assumptions, the most stringent is likely ignorability, which requires that no unobserved confounders influenced past decisions and outcomes. While recent modeling developments may ease ignorability-related concerns in some cases [13, 56], model developers should carefully evaluate whether confounders are likely to impact a model in a given deployment context. At the same time, our results show that formulating algorithmic decision support as a *"pure prediction problem"* that optimizes predictive performance without estimating causal effects [36] imposes equally serious limitations. If the policy-relevant target outcome of interest is risk *conditional on intervention* (as is often the case in decision support applications), an observational model will generate invalid predictions for cases that historically responded most to treatment [13]. Our results, which empirically demonstrate poor performance of observational PU and TU models that overlook treatment-effects, corroborate prior findings indicating that counterfactual modeling is required to ensure the reliability of RAIs in decision support settings [13]. Taken together, our work suggests that domain experts and model developers should exercise considerable caution while mapping the causal estimand of policy interest to the statistical estimand targeted by a predictive model [43].

ACKNOWLEDGMENTS

We thank the anonymous reviewers and attendees of the NeurIPS 2022 Workshop on Causality for Real-world Impact for their helpful feedback. This work was supported by an award from the UL Research Institutes through the Center for Advancing Safety of Machine Intelligence (CASMI) at Northwestern University, the Carnegie Mellon University Block Center for Technology and Society (Award No. 53680.1.5007718), and the National Science Foundation Graduate Research Fellowship Program (Award No. DGE-1745016). ZSW is supported in part by an NSF FAI Award (No. 1939606), a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, an Okawa Foundation Research Grant, and a Mozilla Research Grant.

REFERENCES

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. 2015. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* 33, 4 (2015), 485–505.

[2] Shivani Agarwal. 2014. Surrogate regret bounds for bipartite ranking via strongly proper losses. *The Journal of Machine Learning Research* 15, 1 (2014), 1653–1674.

[3] Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. 2021. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 838–849.

[4] Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.

[5] Colin B Begg and Robert A Greenes. 1983. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* (1983), 207–215.

[6] Christopher M Bishop. 1998. Latent Variable Models. *Learning in graphical models* 371 (1998).

[7] Bradley Butcher, Chris Robinson, Miri Zilka, Riccardo Fogliato, Carolyn Ashurst, and Adrian Weller. 2022. Racial Disparities in the Enforcement of Marijuana Violations in the US. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.* 130–143.

[8] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. Productivity and selection of human capital with machine learning. *American Economic Review* 106, 5 (2016), 124–127.

[9] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. 2021. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11442–11450.

[10] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. Double/debiased/neyman machine learning of treatment effects. *American Economic Review* 107, 5 (2017), 261–265.

[11] Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. 2020. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *International Conference on Machine Learning.* PMLR, 1929–1938.

[12] Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. 2020. Counterfactual predictions under runtime confounding. *Advances in Neural Information Processing Systems* 33 (2020), 4150–4162.

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

[14] Amanda Lee Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2022. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. *First IEEE Conference on Secure and Trustworthy Machine Learning* (2022).

[15] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2021. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648* (2021).

[16] Augustine Denteh and Helge Liebert. 2022. Who Increases Emergency Department Use? New Insights from the Oregon Health Insurance Experiment. *arXiv preprint arXiv:2201.07072* (2022).

[17] Cécile Di Folco, Ava Guez, Hugo Peyre, and Franck Ramus. 2022. Epidemiology of reading disability: A comparison of DSM-5 and ICD-11 criteria. *Scientific Studies of Reading* 26, 4 (2022), 337–355.

[18] Iván Díaz and Mark J van der Laan. 2013. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics* 9, 2 (2013), 149–160.

[19] Claes Enøe, Marios P Georgiadis, and Wesley O Johnson. 2000. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive veterinary medicine* 45, 1-2 (2000), 61–81.

[20] ME Falagas, KZ Vardakas, and PI Vergidis. 2007. Under-diagnosis of common chronic diseases: prevalence and impact on human health. *International journal of clinical practice* 61, 9 (2007), 1569–1579.

[21] Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. 2012. The Oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics* 127, 3 (2012), 1057–1106.

[22] Noam Finkelstein, Roy Adams, Suchi Saria, and Ilya Shpitser. 2021. Partial identifiability in discrete data with measurement error. In *Uncertainty in Artificial Intelligence.* PMLR, 1798–1808.

[23] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. 2020. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics.* PMLR, 2325–2336.

[24] Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. 2021. On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 100–111.

[25] Victoria Gamerman, Tianxi Cai, and Amelie Elsäßer. 2019. Pragmatic randomized clinical trials: best practices and statistical guidance. *Health Services and Outcomes Research Methodology* 19 (2019), 23–35.

[26] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. 2020. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406* (2020).

[27] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.

[28] Sui L Hui and Steven D Walter. 1980. Estimating the error rates of diagnostic tests. *Biometrics* (1980), 167–171.

[29] Paul Hur, HaeJin Lee, Suma Bhat, and Nigel Bosch. 2022. Using Machine Learning Explainability Methods to Personalize Interventions for Students. *International Educational Data Mining Society* (2022).

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

[31] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. 2020. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426* (2020).

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

[33] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *CHI Conference on Human Factors in Computing Systems.* 1–18.

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

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

[36] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–495.

[37] Candace Kruttschnitt, William D Kalsbeek, Carol C House, et al. 2014. Estimating the incidence of rape and sexual assault. (2014).

[38] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.

[39] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. 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.* 275–284.

[40] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* (1986), 604–620.

[41] Tongliang Liu and Dacheng Tao. 2015. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence* 38, 3 (2015), 447–461.

[42] Sharon Lohr. 2019. *Measuring crime: Behind the statistics.* Chapman and Hall/CRC.

[43] Ian Lundberg, Rebecca Johnson, and Brandon M Stewart. 2021. What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review* 86, 3 (2021), 532–565.

[44] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. 2015. Learning from corrupted binary labels via class-probability estimation. In *International conference on machine learning.* PMLR, 125–134.

[45] Sendhil Mullainathan and Ziad Obermeyer. 2017. Does machine learning automate moral hazard and error? *American Economic Review* 107, 5 (2017), 476–480.

[46] Sendhil Mullainathan and Ziad Obermeyer. 2021. On the inequity of predicting a while hoping for B. In *AEA Papers and Proceedings*, Vol. 111. 37–42.

[47] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. *Advances in neural information processing systems* 26 (2013).

[48] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. 2021. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861* (2021).

[49] Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* 70 (2021), 1373–1411.

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

[51] Francesco Orso, Gianna Fabbri, and Aldo Pietro Maggioni. 2017. Epidemiology of heart failure. *Heart Failure* (2017), 15–33.

[52] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1944–1952.

[53] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.

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

[55] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[56] Ashesh Rambachan, Amanda Coston, and Edward Kennedy. 2022. Counterfactual Risk Assessments under Unmeasured Confounding. *arXiv preprint arXiv:2212.09844* (2022).

[57] Henry Reeve et al. 2019. Classification with unknown class-conditional label noise on non-compact feature spaces. In *Conference on Learning Theory*. PMLR, 2624–2651.

[58] Fred S Roberts. 1985. Measurement theory. (1985).

[59] James Robins. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling* 7, 9-12 (1986), 1393–1512.

[60] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[61] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688.

[62] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[63] Henri C Schouwenburg. 2004. Procrastination in Academic Settings: General Introduction. (2004).

[64] Clayton Scott. 2015. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*. PMLR, 838–846.

[65] Clayton Scott, Gilles Blanchard, and Gregory Handy. 2013. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on learning theory*. PMLR, 489–511.

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

[67] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems* 32 (2019).

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

[69] Patrick E Shrout and Sean P Lane. 2012. Psychometrics. (2012).

[70] Di Shu and Grace Y Yi. 2019. Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Statistical methods in medical research* 28, 7 (2019), 2049–2068.

[71] Jeffrey A Smith and Petra E Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics* 125, 1-2 (2005), 305–353.

[72] Bill Turque. 2012. Creative... motivating'and fired. *The Washington Post* 6 (2012).

[73] Brendan Van Rooyen et al. 2015. Machine learning via transitions. (2015).

[74] Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. 2015. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems* 28 (2015).

[75] Steven D Walter and Les M Irwig. 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of clinical epidemiology* 41, 9 (1988), 923–937.

[76] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2022. Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *Available at SSRN* (2022).

[77] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 526–536.

[78] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. 2020. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems* 33 (2020), 7597–7610.

[79] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. 2019. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems* 32 (2019).

[80] Laura Zwaan and Hardeep Singh. 2015. The challenges in defining and measuring diagnostic error. *Diagnosis* 2, 2 (2015), 97–103.

A APPENDIX

This appendix contains the following subsections:

- A.1 provides a discussion of our re-analysis of audit data released by Obermeyer et al. [50].
- A.2 contains omitted proofs for theorems introduced in § 4.
- A.3 contains omitted algorithm pseudocode.
- A.4 contains additional details and results for experiments reported in Section 5.

Sample	FPR	FNR	N
Full population	0.37	0.38	48,784
Unenrolled	0.37	0.39	48,332
Enrolled	0.64	0.13	452

Table 3: Treatment-conditional OME parameters computed using synthetic data released by Obermeyer et al. [50].

Sample	FPR	FNR	N
Full population	0.36	0.39	48,784
Unenrolled	0.36	0.39	48,360
Enrolled	0.65	0.14	424

Table 4: Treatment-conditional OME parameters computed after re-applying synthpop on released synthetic data.

A.1 Re-analysis of data published by Obermeyer et al. [50]

Obermeyer et al. [50] release publicly available synthetic dataset corresponding to their audit of a clinical risk assessment.[7] Synthetic data was generated via the R package synthpop, which preserves moments and covariances of the original dataset. The synthetic data release is sufficient to replicate the main analyses reported over the raw (unmodified dataset) reported in [50]. This makes it likely that our analysis closely preserves the true statistics reported on raw data, as our only analysis step involves thresholding raw scores and computing conditional probabilities.

We probe the implications of naively estimating population OME parameters by reanalyzing public synthetic data published as part of the audit study. Our analysis estimates proxy error parameters by binarizing continuous cost (Y) and chronic active condition (Y^*) outcomes at the 55th risk percentile: the threshold used in practice to drive enrollment recommendations. While this choice of target outcome is itself imperfect [20], we use chronic active conditions as a reference outcome to match the original comparison conducted by Obermeyer et al. [50].

Our analysis (Table 3) finds that the false positive and false negative rates of the cost of care proxy varies substantially across program enrollment status. In particular, the false negative rate is 65.8% lower among patients enrolled in the program as compared to the full population, while the FPR is 72.9% higher. This difference is consistent with closer medical supervision: under enrollment, patients may incur greater costs due to expanded care, even after controlling for the number of underlying active chronic conditions. In contrast, OME parameters among the unenrolled resemble the population average because the vast majority of patients ($\approx 99\%$) are turned away from the program. We verify that this finding is not an artifact of synthetic data generation by re-applying synthpop on data provided by [50] and re-computing error parameters via the same procedure described above (Table 4). While we observe minor variations in error parameters after re-applying synthpop, the large difference in error rates across the full population and enrollment conditions persists.

Triangulating the downstream impacts of this error parameter discrepancy is challenging. To preserve privacy, the research team did not release covariates needed to re-train an algorithm. Prior program enrollment decisions were also non-randomized, meaning that differences in error parameters could be attributed to unmeasured confounders. Nevertheless, the difference in error parameters across enrolled and unenrolled carries serious implications for the diagnosis and mitigation of outcome measurement error.[8]

A.2 Omitted results and proofs

We begin by providing a roof of Theorem 4.1. This proof follows from unbiased risk minimization results from the label noise [11, 47, 52, 73] and counterfactual prediction [31] literature.

Proof. We will show that $R^{w}_{t,\tilde{\ell}}(f_t) = R_{\tilde{\ell}}(f_t) = R^{*}_{\ell}(f_t), \forall t = \{0,1\}$. We begin by showing the first equality. We have that

$$\begin{aligned} R^{w}_{t,\tilde{\ell}}(f_t) &:= \mathbb{E}_{p}\left[w(X)\tilde{\ell}(f_t(X),Y) \mid T=t\right] \\ &= \mathbb{E}_{p^*}\left[w(X)\tilde{\ell}(f_t(X),Y_t) \mid T=t\right] \\ &= \mathbb{E}_{p^*}\left[w(X)\tilde{\ell}(f_t(X),Y_t)\right] \end{aligned}$$

where the first equality holds by consistency (2) and the second by ignorability (3). As a result, we can express both equalities over potential outcomes $Y_t, Y^*_t \sim p^*$. Next, let $p_t(X) := p(X|T=t)$ and let $\tilde{\ell}_{f_t}(x) := \mathbb{E}_{Y_t}[\tilde{\ell}(f_t(x),Y_t) \mid X = x]$ be the *expected pointwise surrogate loss* of f_t evaluated at x. Then by Lemma 2 of [31], we have

$$\begin{aligned} R^{w}_{t,\tilde{\ell}}(f_t) &= \int_{x\in X} w(x)\tilde{\ell}_{f_t}(x)p_t(x)dx \\ &= \int_{x\in X} \frac{p_t(x)}{p(x)} w(x)\tilde{\ell}_{f_t}(x)p(x)dx \\ &= R_{\tilde{\ell}}(f_t) \end{aligned}$$

for $w(x) = p(x)/p_t(x)$. The second equality assumes positivity (4) and ignorability (3). Applying Bayes' to $p(x)/p_t(x)$ and rearranging

$$w(x) = \frac{p(x)}{p(X=x|T=t)} = \frac{p(T=t)}{p(T=t|X=x)} = \frac{p(T=t)}{(2t-1)\cdot\pi(x)+1-t}$$

which is the weighting function in (4). Next, we show that $R_{\tilde{\ell}}(f_t) = R^{*}_{\ell}(f_t)$, which follows from Lemma 1 of [47]. Given

[7] https://gitlab.com/labsysmed/dissecting-bias

[8] Obermeyer et al. [50] report robustness checks examining whether differential program effects by race could impact their study of label bias. The authors found no such differential effects by race. As a result, their main analyses are not likely to be impacted by the findings of our re-analysis. Nevertheless, the model reliability challenges we study in this work could impact *all individuals in the study population*, if unaddressed.

$$\boldsymbol{\eta_t}(x) = \begin{pmatrix} 1-\eta_t(x) \\ \eta_t(x) \end{pmatrix}, \quad \boldsymbol{T} = \begin{pmatrix} 1-\alpha_t & \alpha_t \\ \beta_t & 1-\beta_t \end{pmatrix}$$

we can express (1) as $\boldsymbol{\eta_t}(x) = \boldsymbol{T}\boldsymbol{\eta^*_t}(x)$ by assumption 1. This error model is invertable via $\boldsymbol{\eta^*_t}(x) = \boldsymbol{T}^{-1}\boldsymbol{\eta_t}(x)$ for

$$\boldsymbol{T}^{-1} = \frac{1}{1-\alpha_t-\beta_t}\begin{pmatrix} 1-\beta_t & -\alpha_t \\ -\beta_t & 1-\alpha_t \end{pmatrix}.$$

Let $\boldsymbol{\ell}(f_t(x)) = (\ell(f_t(x),0), \ell(f_t(x),1))^\top$ be a vectorized loss corresponding to labels $\boldsymbol{e} \in \{0,1\}^2$. Then we have that

$$\begin{aligned} R^*_\ell(f_t) &= \mathbb{E}_X\mathbb{E}_{Y^*_t \sim \boldsymbol{\eta}^*_t(X)}[\ell(f_t(X), Y^*_t)] = \mathbb{E}_X[\boldsymbol{\eta}^*_t(X)^\top \boldsymbol{\ell}(f_t(X))] \\ &= \mathbb{E}_X[\boldsymbol{\eta}_t(X)^\top(\mathbf{T}^{-1})\boldsymbol{\ell}(f_t(X))] = \mathbb{E}_X[\mathbf{e}^\top(\mathbf{T}^{-1})\boldsymbol{\ell}(f_t(X))] \\ &= \mathbb{E}_{X,Y_t}[\tilde{\ell}(f_t(X), Y_t)] = R_{\tilde{\ell}}(f_t) \end{aligned}$$

Therefore, $R_{\tilde{\ell}}(f_t) = R^*_\ell(f_t)$ for a surrogate loss constructed via $\tilde{\boldsymbol{\ell}} = (\mathbf{T}^{-1})\boldsymbol{\ell}(f_t(X))$. Multiplying $\boldsymbol{\ell}(f_t(X))$ by $\mathbf{T}^{-1}$ and rearranging yields (5). □

Next, we prove Theorem 4.2 showing that error parameters are identifiable under combinations of assumptions stated in Table 1.

Proof. To begin, observe that the error model (1) expresses the conditional proxy class probability η_t as a linear function of η^*_t with two unknowns. Therefore, given knowledge of the target class probability $c^*_{t,i} = \eta^*_t(x_i)$ and proxy class probability $c_{t,i} = \eta_t(x_i)$ at two distinct points $(c^*_{t,i}, c_{t,i})$ and $(c^*_{t,j}, c_{t,j})$, we can set up a linear equation

$$\begin{aligned} c_{t,i} &= (1-\beta_t)\cdot c^*_{t,i} + \alpha_t \cdot (1-c^*_{t,i}) \\ c_{t,j} &= (1-\beta_t)\cdot c^*_{t,j} + \alpha_t \cdot (1-c^*_{t,j}) \end{aligned} \tag{9}$$

and solve for error parameters

$$\alpha_t = \frac{c^*_{t,i}\cdot c_{t,j} - c_{t,i}\cdot c^*_{t,j}}{c^*_{t,i} - c^*_{t,j}} \tag{10}$$

$$\beta_t = \frac{c_{t,i}\cdot c^*_{t,j} - c_{t,i} + c^*_{t,i} - c^*_{t,j} + c_{t,j} - c^*_{t,i}\cdot c_{t,j}}{c^*_{t,i} - c^*_{t,j}} \tag{11}$$

provided that $c^*_{t,i} \neq c^*_{t,j}$. Identification of the specific cases in Table 1 follows from application of (10). When α_t and β_t are both known, identification is not required. When one of β_t (α_t) is known, the corresponding α_t (β_t) can be given by

$$\alpha_t = \frac{c_{t,i} - (1-\beta_t)\cdot c^*_{t,i}}{(1-c^*_{t,i})}, \quad \beta_t = \frac{c^*_{t,i} - c_{t,i} + \alpha_t\cdot(1-c^*_{t,i})}{c^*_{t,i}} \tag{12}$$

Therefore, only one anchor assumption $(c^*_{t,i}, c_{t,i})$ is required given knowledge of α_t or β_t. However, by (12), note that $c^*_{t,i} \neq 1$ is required for identification of α_t and $c^*_{t,j} \neq 0$ is required for identification of β_t. This rules out combinations denoted by (✕) in Table 1. Error parameters can be derived directly from (10) if α_t and β_t are both unknown so long as $c^*_{t,i} \neq c^*_{t,j}$. The specific values of $(c_{t,i}, c^*_{t,i})$ corresponding to min, max, and base rate anchors can be computed via

$$c^*_{t,i} = \inf_{x_i \in X}\{\eta^*_t(x_i)\}, \quad c_{t,i} = \inf_{x_i \in X}\{\eta_t(x_i)\} \quad \text{(Min anchor)}$$

$$c^*_{t,i} = \sup_{x_i \in X}\{\eta^*_t(x_i)\}, \quad c_{t,i} = \sup_{x_i \in X}\{\eta_t(x_i)\} \quad \text{(Max anchor)}$$

$$c^*_{t,i} = \mathbb{E}[\eta^*_t(X)], \quad c_{t,i} = \mathbb{E}[\eta_t(X)] \quad \text{(Base rate anchor)}$$

Above, the min anchor holds because η_t is a strictly monotone increasing transform of η^*_t by 1 such that $c_{t,i} = \arg\inf_{x_i \in X}\{\eta_t(x)\} = \arg\inf_{x_i \in X}\{\eta^*_t(x)\}$. The max anchor holds by the same argument. The base rate anchor holds because $\mathbb{E}_X[\eta_t(x)] = \mathbb{E}_X[\eta^*_t(x)\cdot(1-\beta_t-\alpha_t)+\alpha_t]$.

Finally, observe that $\eta_t(x)$ is defined over potential outcomes $Y_t \sim p^*$ rather than observational proxies $Y \sim p$. Identification of η_t from observational data follows from

$$\eta_t(x) := p(Y_t = 1|X = x) = p(Y = 1|X = x, T = t) \tag{13}$$

where the equality holds by ignorability (3) and consistency (2). By positivity (4), we have that the support of $\eta_t(x)$ is defined $\forall x \in X$, which guarantees that the min and max anchor will be defined. □

A.3 Algorithms

The RW-SL and CCPE algorithms presented in § 4 partition training data into disjoint folds to learn $\hat{\alpha}_t$, $\hat{\beta}_t$, $\hat{\pi}$, and minimize the re-weighted surrogate risk. We also provide a version of these algorithms with cross-fitting to improve data efficiency. Cross-fitting is useful when using limited data to fit multiple nuisance functions and improves data efficiency while limiting over-fitting [34].

Algorithm 3: Re-weighted risk minimization with surrogate loss (cross fitting)

```
Input: Data $\mathcal{W} = \{(X_i, T_i, Y_i)\}_{i=1}^n \sim p$
Output: Learned estimator $\hat{\eta}^*_t(x)$
Partition $\mathcal{W}$ into $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$
for $(m, n, p) \in \{(1,2,3), (3,1,2), (2,3,1)\}$ do
    On $\mathcal{W}_m$, estimate parameters $\hat{\alpha}^m_t, \hat{\beta}^m_t \leftarrow$ CCPE($\mathcal{W}_m$)
    On $\mathcal{W}_n$, learn $\hat{\pi}_n(x)$ by regressing $T \sim X$
    On $\mathcal{W}_p$, use $\hat{\pi}_n(x), \hat{\alpha}^m_t, \hat{\beta}^m_t$ to solve
      $\hat{\eta}^*_{t,p}(x) \leftarrow \arg\min_{f_t \in \mathcal{H}} \hat{R}^{\hat{w}}_{t,\tilde{\ell}}(f_t)$
end
Return combined predictions $\hat{\eta}^*_t(x) = \frac{1}{3}\sum_{p=1}^3 \hat{\eta}^*_{t,p}(x)$
```

Algorithm 4: Conditional class probability estimation (cross fitting)

```
Input: Data $\mathcal{V} \sim p$
Output: Parameter estimates $\hat{\alpha}_t, \hat{\beta}_t$
Partition $\mathcal{V}$ into $\mathcal{V}_1, \mathcal{V}_2$
for $(m, n) \in \{(1,2), (2,1)\}$ do
    On $\mathcal{V}_m$, learn $\hat{\eta}^m_t(x)$ by regressing $Y \sim X \mid T = t$
    On $\mathcal{V}_n$, estimate error parameters:
      $\hat{\alpha}^n_t = \min_{x \in X}\{\hat{\eta}^m_t(x)\}, \quad \hat{\beta}^n_t = 1 - \max_{x \in X}\{\hat{\eta}^m_t(x)\}$
end
Return averaged parameters $\hat{\alpha}_t = \frac{1}{2}\sum_{n=1}^2 \hat{\alpha}^n_t$,
  $\hat{\beta}_t = \frac{1}{2}\sum_{n=1}^2 \hat{\beta}^n_t$
```

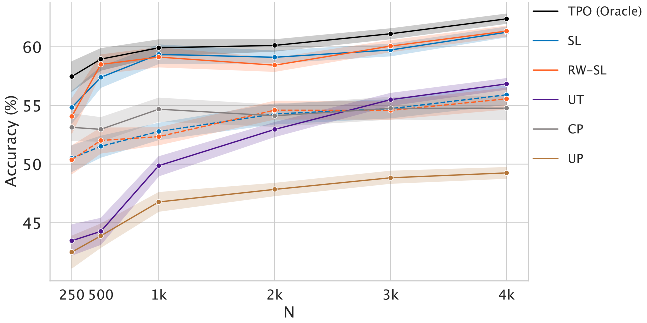

Figure 6: Comparison of models across small sample size regimes. SL and RW-SL with oracle parameters maintain performance parity with TPO across settings with 1k+ samples, but demonstrate worse performance than TPO in the most data scarce setting with 250 samples.

A.4 Additional experimental details and results

A.4.1 Setup details. In our synthetic evaluation, we sample from target class probability functions $\eta_0^*(x) = .4 + .4\cos(9x + 5.5), \forall x \in [-1, -.237]$; $.5 + .3\sin(8x + .9) + .15\sin(10x + .2) + .05\sin(30x + .2), \forall x \in (-.237, 1]$ and $\eta_1^*(x) = .5 - .5\sin(2.9x + .1)$ and sample treatments from the linear function $\pi(x) = .35x + .5$ (Figure 3).

We train all models with a binary-cross entropy loss. We use the same 4-layer MLP implemented via PyTorch with hidden layer sizes $(40, 30, 10)$ for all models discussed in § 5.1. Where relevant, we also fit $\pi(x)$ and $\eta_t(x)$ (used in CCPE) via the same architecture. We train all models for 10 each epochs each at learning rate $\eta = .5e^{-3}$. Hyperparameters were selected via a hyperparameter sweep optimizing accuracy on Y_0^* with respect to the TPO model.

In our semi-synthetic experiments, we run all models in the synthetic experiment without sample splitting and cross-fitting. While cross-fitting improves data efficiency and typically performs better in low sample settings, the treatment group in JOBS data had very few positive (unemployment) outcomes. As a result, we observed poor convergence of our MLP models across folds when performing sample splitting on this dataset. Therefore we run JOBS without sample splitting and cross-fitting, and maintain the same setting with OHIE data for consistency. We use a 4-layer MLP with layer sizes (30, 20, 10) for JOBS data and a 4-layer MLP with layer sizes (40, 30, 10) for OHIE data. We use $\eta = 1e - 3$ for JOBS data and $\eta = 5e - 3$ for OHIE data. We train JOBS and OHIE models for 15 and 20 epochs respectively. We with the synthetic experiment, we select hyperparamters by optimizing model performance with respect to the oracle TC model and use the same settings across all models. Note that $\tau = 0.015$ and $\tau = -0.077$ for the outcomes targeted in OHIE and JOBS, respectively.

A.4.2 Additional results. Theorem 4.1 shows that the re-weighted surrogate loss recovers the loss with respect to target potential outcomes in expectation. Because we do not provide a finite sample convergence rate for our method, we extend our synthetic evaluation to a low sample size regime to empirically test the performance of RW-SL on finite samples of limited size. Figure 6 shows a convergence plot for this experiment. We perform this analysis with the same set of hyperparameters used in the main experimental results reported in § 5. This plot indicates that the performance of all methods deteriorates as sample availability decreases, with performance upper bounded by the oracle TPO model. SL and RW-SL with oracle parameters achieve performance at near parity with TPO in sample settings above 500 samples, and begin to show rapid performance deterioration at 250 samples. This indicates that both SL and RW-SL tend to perform reliably in small sample settings *when parameters and weights are known.* However, SL and RW-SL with learned parameters performs poorly across all sample settings. This is likely due to cascading errors arising from bias in error parameter estimates. UT and UP both learn a function predicting the average outcome response $\hat{\eta}(x) \approx .61, \forall x \in X$ in the setting with 250 and 500 samples. As a result, these methods demonstrate accuracy lower than 50% in the small sample settings.