Counterfactual Decision Support Under Treatment-Conditional Outcome Measurement Error

Luke Guerdan
Carnegie Mellon University
Pittsburgh, PA 15213
lguerdan@cs.cmu.edu

Amanda Coston
Carnegie Mellon University
Pittsburgh, PA 15213

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, PA 15213

Zhiwei Steven Wu
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Growing work in algorithmic decision support proposes methods for combining predictive models with human judgment to improve decision quality. A challenge that arises in this setting is predicting the risk of a decision-relevant target outcome under multiple candidate actions. While counterfactual prediction techniques have been developed for these tasks, current approaches do not account for measurement error in observed labels. This is a key limitation because in many domains, observed labels (e.g., medical diagnoses, test scores) serve as a proxy for the target outcome of interest (e.g., biological medical outcomes, student learning). We develop a method for counterfactual prediction of target outcomes observed under treatment-conditional outcome measurement error (TC-OME). Our method minimizes risk with respect to target potential outcomes given access to observational data and estimates of measurement error parameters. We also develop a method for estimating error parameters in cases where these are unknown in advance. Through a synthetic evaluation, we show that our approach achieves performance parity with an oracle model when measurement error parameters are known and retains performance given moderate bias in error parameter estimates.

1 Introduction

Predictive models are increasingly being introduced to support expert decision-making in real-world tasks. In the medical domain, clinical models have been developed to inform patient treatment decisions by predicting the likelihood of adverse health outcomes (e.g, heart attack, stroke) [Mullainathan and Obermeyer, 2022]. In the educational domain, learning analytics tools have been introduced to allocate additional tutoring resources to at-risk students [Livieris et al., 2016]. In these settings, policy makers often wish to estimate the risk of a downstream target outcome under multiple alternative actions [Schulam and Saria, 2017].

However, because outcomes were observed under the past decision-making policy, we do not observe the counterfactual outcome that *would* under a different decision [Pearl, 2009]. This makes it challenging to learn a counterfactual model given observational data. Recent work proposes counterfactual modeling and evaluation approaches designed for decision-support settings Coston et al. [2020b,a]. This work builds upon causal inference methods for conditional average treatment effect (CATE) estimation [Abrevaya et al., 2015], which have received growing interest within the machine learning community [Johansson et al., 2018]. We consider counterfactual modeling in two real-world decision support tasks. In *selective intervention* tasks, a model estimates baseline risk

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

under no intervention (e.g., risk of child neglect given no welfare services [Chouldechova et al., 2018], or poor learning outcomes given no tutoring). In *selective opportunity* settings, a model estimates risk of a target outcome under a proposed opportunity (e.g., likelihood default given receiving a loan, or performance ratings given a new job). We study these counterfactual prediction settings where the target outcome is subject to measurement error.

Outcome measurement error occurs when policy-relevant target outcomes (e.g., heart attack, student learning) are imperfectly approximated by *proxies* (e.g., heart attack *diagnosis*, test scores) in data [Jacobs and Wallach, 2021]. While measurement error and label noise have been studied in previous literature [Menon et al., 2015, Natarajan et al., 2013, Wang et al., 2021, Fogliato et al., 2020, 2021], this challenge has not been addressed in counterfactual prediction settings. Therefore, in this work, we study the novel setting of counterfactual prediction in the presence of treatment-conditional outcome measurement error (TC-OME). We offer the following contributions:

- We formalize the problem of counterfactual prediction under treatment-conditional outcome measurement error.
- We develop a risk minimization approach (FRM-SL; Algorithm 1) for estimating target potential outcomes given observed covariates, past decisions, and proxy outcomes observed under treatment-conditional outcome measurement error. Our method combines covariate adjustment techniques designed for CATE inference [Johansson et al., 2020] with a surrogate loss developed by Natarajan et al. [2013] for label noise correction.
- We develop a method for estimating treatment-conditional measurement error parameters (CCPE; Algorithm 2). Our approach builds on class probability estimation (CPE) techniques designed for label noise settings [Menon et al., 2015, Scott et al., 2013].
- We evaluate FRM-SL and CCPE via synthetic experiments and show that FRM-SL achieves performance parity with an oracle model given access to ground-truth error parameters. We also show that FRM-SL performance is tolerant to moderate bias in CCPE parameter estimates, and show that bias in estimates decreases as a function of sample size.

2 Problem setup

We consider a counterfactual distribution defined over $p^*(X, D, Y_0^*, Y_1^*)$, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$ are covariates, $D \in \{0, 1\}$ are past decisions, and Y_0^*, $Y_1^* \in \{0, 1\}$ are target binary potential outcomes of interest to human decision-makers. Under potential outcomes [Rubin, 2005], Y_0^* is the hypothetical outcome we would see under the baseline condition when $d = 0$, while Y_1^* is the outcome we would observe under the proposed intervention (selective intervention) or opportunity (selective opportunity) when $d = 1$. Following the standard setup in causal inference, we only observe Y_0^* or Y_1^* for a given instance such that $Y^* = D \cdot Y_1^* + (1 - D) \cdot Y_0^*$ [Pearl, 2009].

Given the counterfactual joint p^*, we would like to estimate the target quantity

$$\eta_d^*(x) \coloneqq \mathbb{P}(Y_d^* = 1 \mid X = x), \ \forall x \in X. \tag{1}$$

where $d = 0$ in selective intervention settings, and $d = 1$ in selective opportunity settings.

However, rather than sampling directly from p^*, we draw samples i.i.d. from $p(X, D, Y)$, where $Y \in \{0, 1\}$ is a binary proxy outcome. Two challenges complicate estimation of $\eta_d^*(x)$ given samples from p. First, observed proxies Y arise from potential outcomes such that $Y = D \cdot Y_1 + (1 - D) \cdot Y_0$. Therefore, we do not know the outcome that *would have* occurred had the counterfactual decision been made in the past. Second, proxy potential outcomes Y_d are subject to outcome measurement error. In our *treatment-conditional outcome measurement error* (TC-OME) model, the proxy potential outcome is observed under a false positive rate α_d and false negative rate β_d given by

$$\alpha_d \coloneqq \mathbb{P}(Y_d = 1 \mid Y_d^* = 0), \beta_d \coloneqq \mathbb{P}(Y_d = 0 \mid Y_d^* = 1), \ \ \forall d \in D \tag{2}$$

where $\alpha_d + \beta_d < 1$. Figure 1 shows an example of TC-OME in a heart attack prediction context, including factual and counterfactual target potential outcomes that can be observed under different error rates for $d = \{0, 1\}$.

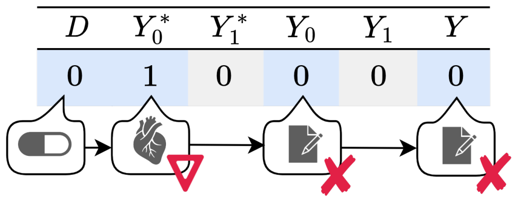

Figure 1: Example of TC-OME in a heart attack prediction setting. Under the past decision not to treat (factual in blue), heart attack occurred ($Y_0^* = 1$) but went undiagnosed ($Y_0 = 0$) and no heart attack was recorded ($Y = 0$). Had the patient received treatment (counterfactual in grey), the patient would not have had a heart attack ($Y_1^* = 0$) and would be correctly diagnosed ($Y_1 = 0$) and recorded ($Y = 0$) as such.

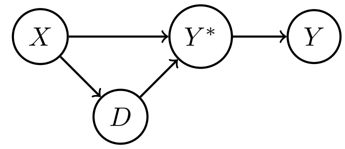

Figure 2: Causal diagram of treatment conditional outcome measurement error.

Under this model, the class probability function of the proxy potential outcome Y_d can be given by

$$\begin{aligned}\mathbb{P}(Y_d = 1 \mid X = x) &= \mathbb{P}(Y_d = 1 \mid Y_d^* = 1) \cdot \mathbb{P}(Y_d^* = 1 \mid X = x) \\ &+ \mathbb{P}(Y_d = 1 \mid Y_d^* = 0) \cdot \mathbb{P}(Y_d^* = 0 \mid X = x) \\ &= (1 - \beta_d)\ \eta_d^*(x) + \alpha_d\ (1 - \eta_d^*(x)),\ \forall x \in X, d \in D.\end{aligned} \tag{3}$$

which gives an expression for the observed class probability $\eta_d(x)$ in terms of the target class probability $\eta_d^*(x)$. Modeling measurement error requires making assumptions on the relationship between target and proxy outcomes [Jacobs and Wallach, 2021]. We make the following assumptions in our model:

Assumption 2.1 (Measurement Error Model)**.** Error parameters not depend on covariates X or unmeasured confounders Z: $\mathbb{P}(Y_d \mid Y_d^*) = \mathbb{P}(Y_d \mid Y_d^*, X = x) = \mathbb{P}(Y_d \mid Y_d^*, Z = z)$.

While this assumption follows from the class-conditional model studied in past literature [Menon et al., 2015, Scott et al., 2013], our methods can be readily extended to settings where $\mathbb{P}(Y_d \mid Y_d^*) \neq \mathbb{P}(Y_d \mid Y_d^*, X = x)$ by applying a group-dependent error model [Wang et al., 2021]. It also follows from e.q. 3 that $\forall d \in D$, $\eta_d(x)$ is a strictly monotone increasing transform of $\eta_d^*(x)$.

2.1 Identifiability conditions

Figure 2 shows a causal diagram specifying the assumptions we make on the data generating process in a TC-OME setting. Class probability functions η_d^* and η_d are *identifiable* if they can be computed uniquely from the observed distribution $p(X, D, Y)$. Our target causal estimand $\eta_d^*(x)$ is not identifiable directly because potential outcomes $\{Y_0^*, Y_1^*\}$ are unobserved. However, $\{Y_0, Y_1\}$ are identifiable under a standard set of causal identifiability assumptions:

Assumption 2.2 (Consistency)**.** An instance receiving decision $d \in \{0, 1\}$ has outcome $Y = Y_d$: $Y = D \cdot Y_1 + (1 - D) \cdot Y_0$.

Assumption 2.3 (Ignorability)**.** Potential outcomes and decisions are conditionally independent given X: $\{Y_0, Y_1\} \perp\!\!\!\perp D \mid X$.

Assumption 2.4 (Positivity)**.** For any set of covariates $x \in X$, both decisions have non-zero probability of observation in the data: $\forall x \in X, d \in \{0, 1\} : p(D = d|X = x) > 0$.

3 Methodology

First, we develop an estimator for $\eta_d^*(x)$ given i.i.d. samples drawn from the observational joint $p(X, D, Y)$ assuming *a priori* knowledge of error terms (Section 3.1). In practice, α_d and β_d are unknown in advance. Therefore, in Section 3.2, we develop an approach for error parameter estimation (Algorithm 2). This approach requires access to observational data from p and an additional *weak separability* assumption commonly applied in class-conditional error settings [Menon et al., 2015].

3.1 Risk minimization

Let $f \in \mathcal{H}$ be a probabilistic decision function belonging to $\mathcal{H} \subset \{h : \mathcal{X} \to [0,1]\}$ and let $\ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}_+$ be a bounded loss function. The ℓ-risk of f over Y_d^* can be given by

$$R_\ell^*(f) \coloneqq \mathbb{E}_{X,Y_d^*}[\ell(f(X), Y_d^*)] \tag{4}$$

where $R_\ell^*(f)$ is the marginal risk over the full population.

Because we wish to recover a class probability $\eta_d^*(x)$, we restrict ℓ to be *strictly proper composite* such that $\arg\min_f R_\ell^*(f)$ is a monotone transform ϕ of η_d^* (e.g., the logistic and exponential loss) [Agarwal, 2014, Menon et al., 2015]. This allows for recovering class probabilities from the optimal prediction via the link function ϕ.

Johansson et al. [2020] show that the marginal risk over the full population can be decomposed into *factual* and *counterfactual* components scaled by $\pi = p(D = 1)$ via

$$R_\ell^*\left(f\right) \coloneqq \pi \underbrace{R_{d,\ell}^*\left(f\right)}_{\text{Factual}} + \left(1-\pi\right) \underbrace{R_{1-d,\ell}^*\left(f\right)}_{\text{Counterfactual}}. \tag{5}$$

for $d \in \{0, 1\}$. The factual risk, denoted by subscript d, can be directly estimated over the sample that received treatment $D = d$. Under ignorability, this factual risk is identifiable as

$$R_{d,\ell}^*(f) \coloneqq \mathbb{E}_{X,Y_d^*}[\ell(f(X), Y_d^*)] = \mathbb{E}_{X,Y^*|D}[\ell(f(X), Y^*)|D = d] \tag{6}$$

The factual risk $R_{d,\ell}^*(f)$ is a biased estimator for population risk $R_\ell^*(f)$ in observational settings (i.e., when $X \not\perp D$) [Johansson et al., 2020]. However, previous work has demonstrated empirically that bias correction techniques such as propensity re-weighting offer limited performance improvements in counterfactual risk assessment settings given sufficient sample size and an expressive model class [Coston et al., 2020b]. Therefore, in this work, we develop a minimizer for the factual risk $R_{d,\ell}^*(f)$ and leave a comparison with re-weighting based approaches for future work.

The factual risk $R_{d,\ell}^*(f)$ cannot be estimated directly because target potential outcomes are unobserved. Instead, we construct a surrogate loss $\tilde{\ell}$ such that minimizing factual $\tilde{\ell}$-risk w.r.t. proxy outcomes Y is equivalent to minimizing factual ℓ-risk w.r.t. target outcomes Y^* in expectation [Natarajan et al., 2013]. Under ignorability, the *factual risk* over the proxy potential outcome Y_d is identifiable by conditioning on $D = d$ via

$$R_{d,\ell}(f) \coloneqq \mathbb{E}_{X,Y_d}[\ell(f(X), Y_d)] = \mathbb{E}_{X,Y|D}[\ell(f(X), Y)|D = d] \tag{7}$$

Given knowledge of error parameters, we wish to construct a surrogate loss $\tilde{\ell}$ s.t. $R_{d,\tilde{\ell}}(f)$ gives an unbiased estimator for $R_{d,\ell}^*(f)$

$$\mathbb{E}_{Y|Y^*=y^*,D=d}[\tilde{\ell}(f(x), Y)] = \ell(f(x), y^*) \tag{8}$$

$\forall x \in X$ s.t. $D = d$, where ℓ is computed on the target y^* and $\tilde{\ell}$ is computed over proxy outcomes Y. By Lemma 1 in Natarajan et al. [2013] , such an $\tilde{\ell}$ can be constructed via

$$\begin{aligned} \tilde{\ell}(f(x), +1) &\coloneqq \frac{(1-\alpha_d)\ \ell(f(x), +1) - \beta_d\ \ell(f(x), -1)}{1 - \beta_d - \alpha_d} \\ \tilde{\ell}(f(x), -1) &\coloneqq \frac{(1-\beta_d)\ \ell(f(x), -1) - \alpha_d\ \ell(f(x), +1)}{1 - \beta_d - \alpha_d} \end{aligned} \tag{9}$$

$\forall x \in X$ s.t. $D = d$ and arbitrary loss ℓ.

Algorithm 1: Factual risk minimization with surrogate loss (FRM-SL)

Input: Data $W \sim p$
Output: Learned estimator for $\hat{\eta}^*_d(x)$
Compute parameter estimates $\hat{\alpha}_d, \hat{\beta}_d \leftarrow \text{CCPE}(W)$
Construct $\tilde{\ell}$ parameterized by $\hat{\alpha}_d$, $\hat{\beta}_d$
Learn $\hat{\eta}^*_d(x) := \arg\min_{f \in \mathcal{H}} \hat{R}_{d,\tilde{\ell}}(f)$

Algorithm 2: Conditional class probability estimation (CCPE)

Input: Data $W = \{X_i, D_i, Y_i\}_{i \in N} \sim p$
Output: Parameter estimates $\hat{\alpha}_d$, $\hat{\beta}_d$
Partition W into subsets W^1, W^2
Learn $\hat{\eta}_d(x) := \arg\min_{f \in \mathcal{H}} \hat{R}_{d,\ell}(f)$ on W^1
Use $\hat{\eta}_d(x)$ to estimate error terms on W^2:
$\hat{\alpha}_d = \min_{x \in X}\{\hat{\eta}_d(x)\}, \hat{\beta}_d = 1 - \max_{x \in X}\{\hat{\eta}_d(x)\}$

Because $R_{d,\tilde{\ell}}(f)$ is over the observed joint $p(X, D, Y)$, we can compute the empirical $\tilde{\ell}$-risk on samples $(X_1, D_1, Y_1), ..., (X_n, D_n, Y_n)$ from p

$$\hat{R}_{d,\tilde{\ell}}(f) := \frac{1}{|D'|} \sum_{i \in D'} \tilde{\ell}(f(X_i), Y_i) \tag{10}$$

where $D^{'} = \{i \in 1, ..., N : D_i = d\}$. We can then learn a predictor from observed data by minimizing the empirical risk

$$\hat{f} \leftarrow \arg\min_{f \in \mathcal{H}} \hat{R}_{d,\tilde{\ell}}(f) \tag{11}$$

By Lemma 1 in Natarajan et al. [2013], e.q. 11 converges to the factual risk of the target potential outcome $R^*_{d,\ell}(f)$ in expectation. Under the condition that ℓ is strictly proper composite, $\hat{R}^*_d(f)$ can be used to recover desired class probabilities $\eta^*_d(x)$. Therefore, given a priori knowledge of α_d, β_d, we can learn an estimator for $\eta^*_d(x)$ given samples from $p(X, D, Y)$.

3.2 Error parameter estimation

Directly minimizing $\hat{R}_{d,\tilde{\ell}}(f)$ is challenging because error parameters are often unknown in advance. Therefore, we develop a procedure for estimating error terms, then use estimates $\hat{\alpha}_d$, $\hat{\beta}_d$ to construct the surrogate loss. Our parameter estimation approach places a weak separability assumption on p^*.

Assumption 3.1 (Weak Separability)**.** For fixed $d \in \{0, 1\}$, there exist target potential outcomes Y^*_d such that $\inf_{x \in X}\{\eta^*_d(x)\} = 0$ and $\sup_{x \in X}\{\eta^*_d(x)\} = 1$.

This assumption has been widely applied in observational label noise settings [Menon et al., 2015, Xia et al., 2019, Wang et al., 2021]. In our *counterfactual* setting, this assumption stipulates that there exists an individual at no risk ($Y^*_d = 0$) and another individual at certain risk ($Y^*_d = 1$) under $d = 0$ in selective intervention settings or $d = 1$ in selective opportunity settings. This assumption need only hold for $d = 0$ or $d = 1$ depending on the target quantity of interest in a given setting.

Given weak separability, error parameters can be estimated by substituting $\eta^*_d(x) = 0$, $\eta^*_d(x) = 1$ into the TC-OME model (e.q. 3) and solving

$$\begin{aligned} \eta^*_d(x) = 0 &\implies \alpha_d = \inf_{x \in X} \eta_d(x) \\ \eta^*_d(x) = 1 &\implies \beta_d = 1 - \sup_{x \in X} \eta_d(x) \end{aligned} \tag{12}$$

where we can take the infimum and supremum of $\eta_d(x)$ because it is a monotone transform of $\eta^*_d(x)$ by e.q. 3. Therefore, we estimate α_d, β_d by learning $\hat{\eta}_d(x)$, then computing the minimum and maximum over class probabilities predicted on a held-out sample (Algorithm 2). For statistical purposes, we perform each step on disjoint data folds [Menon et al., 2015]. Estimates $\hat{\alpha}$, $\hat{\beta}$ can then be used to construct $\tilde{\ell}$ and minimize e.q. 11 to learn $\hat{\eta}^*_d(x)$ (Algorithm 1).

	$(0,0)$	$(.1,.3)$	$(.2,.2)$	$(.3,.1)$
OBS	55.6 (± 6.0)	54.5 (± 1.7)	53.7 (± 4.6)	58.5 (± 3.5)
FRM	75.6 (± 3.4)	63.0 (± 8.7)	**72.1** (± 7.7)	67.9 (± 6.8)
FRM+SL	**77.1** (± 0.5)	**76.2** (± 0.8)	69.3 (± 9.6)	**73.0** (± 6.8)
PROXY*	**75.1** (± 3.6)	65.8 (± 11.7)	**75.4** (± 3.8)	64.6 (± 12.4)
TARGET*	72.6 (± 7.5)	**75.7** (± 2.9)	70.7 (± 9.4)	**72.9** (± 9.3)

Table 1: Mean accuracy (%) over 10 runs with standard error reported in parentheses $N = 10,000$. Results for configuration (α_d, β_d) shown in each column.

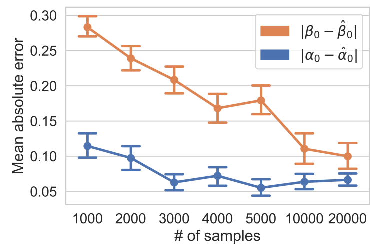

Figure 3: Parameter estimation error as a function of sample size.

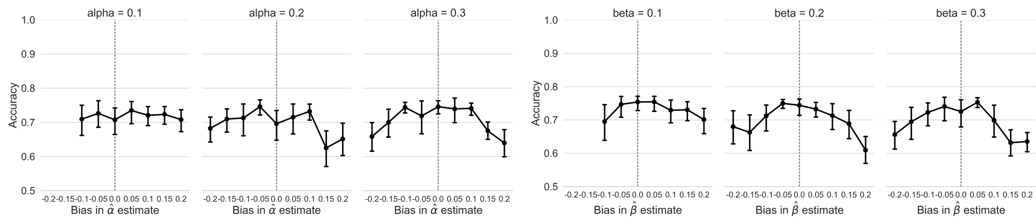

Figure 4: FRM-SL performance as a function of bias in $\hat{\alpha}_0$, $\hat{\beta}_0$. We vary α_0 (β_0) across columns keeping β_0 (α_0) fixed at 0.

4 Experiments

Setup. TC-OME evaluation on real-world data is challenging due to compounding uncertainty from (1) unobserved potential outcomes [Pearl, 2009, Coston et al., 2020b] and (2) measurement error [De-Arteaga et al., 2021, Fogliato et al., 2021]. Therefore, we conduct a controlled synthetic evaluation to validate our approach. Our evaluation emulates a selective intervention setting with target quantity $\eta_0^*(x)$. We use a unidimensional covariate $X \sim U(-1,1)$ and sinusoidal functions $\eta_0^*(x)$, $\eta_1^*(x)$ satisfying weak separability. We sample $Y_0^*, Y_1^* \sim \text{Bern}(\eta_d^*(x)), \forall d \in D$ and generate proxy outcomes by flipping Y_0^*, Y_1^* with probability given by α_0, β_0. We observe outcomes Y by sampling from a propensity function $\pi(x) = \mathbb{P}(D = 1 \mid X = x)$ that is linear in X. We use an MLP trained via binary cross-entropy loss in all experiments. Appendix A.1 contains additional details.

Experiments. Experiment one (Table 1) compares (i) an observational model targeting Y (OBS); (ii) factual risk minimization with unmodified loss (FRM); and (iii) factual risk minimization with a surrogate loss parameterized by α_0, β_d (FRM+SL; Algorithm 1). We compare against oracles predicting Y_0^* (TARGET*) and Y_0 (PROXY*). We report accuracy with respect to Y_0^* on a held-out sample in line with a selective intervention setting targeting $\eta_0^*(x)$. As shown in Table 1, OBS performs poorly across all configurations. FRM and oracle methods perform comparably in configurations with (1) no measurement error ($\alpha_0 = 0, \beta_0 = 0$) and (2) with symmetric error terms $(0.2, 0.2)$. In the former case, a surrogate loss is not needed to correct for measurement error. In the latter, FRM and PROXY* perform well because the optimal threshold minimizing misclassification risk can be computed directly from observed labels given symmetric error terms (see Menon et al. [2015], Appendix F.1). In more realistic asymmetric error configurations $(.1, .3)$, $(.3, .1)$, FRM+SL outperforms FRM and performs at parity with TARGET*. While this experiment assumes oracle access to α_0, β_d, we conduct an additional experiment (Figure 4) showing that FRM+SL performance is robust to bias in $\hat{\alpha}_0$, $\hat{\beta}_0$ in the neighborhood of $\pm.05$ to $\pm.1$ depending magnitude of α_0, β_0.

Experiment two (Figure 3) evaluates CCPE as a function of sample size. We construct $\pi(x)$ and $\eta_0^*(x)$ such that $\pi(x) \propto \eta_0^*(x)$. This mirrors the real-world setting in which individuals at high baseline risk are more likely to receive a risk-reducing intervention. Because this results in lower sample density over the high risk region of $\eta_0^*(x)$, the learned approximation $\hat{\eta}_0(x)$ is likely to be worse near the supremum of $\eta_0^*(x)$ than at its infimum. As a result, we expect more bias in $\hat{\beta}_d$ than $\hat{\alpha}_d$. Figure 3 shows that estimates improve as N increases, and confirms slower convergence for $\hat{\beta}_d$ than $\hat{\alpha}_d$.

5 Discussion

We introduce a novel treatment-conditional outcome measurement error model that formalizes key challenges in counterfactual prediction settings. We provide risk minimization (Algorithm 1) and parameter estimation (Algorithm 2) techniques designed for this setting. Synthetic results demonstrate that FRM+SL provides a strong improvement over FRM alone in asymmetric error settings. Results also show that FRM+SL performance remains robust given bias in $\hat{\alpha}_d$, $\hat{\beta}_d$, and that CCPE recovers reasonable estimates for α_d, β_d.

Our counterfactual identification and measurement model assumptions can be violated in some real-world settings. For instance, ignorability (Assumption 2.3) can be violated if humans make use of predictive contextual factors that are not recorded as covariates [Kleinberg et al., 2018, Lakkaraju et al., 2017]. This issue can be partially circumvented if unobservables are only un-available at runtime [Coston et al., 2020b]. In decision support settings, positivity violations are not of major concern. This is because the instances that require support from predictive models often have uncertain risk profiles and could receive either decision. In contrast, the "cut-and-dry" cases potentially violating positivity (Assumption 2.4) are normally routed through other administrative decision-making procedures. It is also important to carefully consider, based on domain expertise, whether weak separability (Assumption 2) is likely to hold in a given modeling context. This assumption has shown to be unreasonable in some settings (e.g., criminal justice [Fogliato et al., 2020, 2021]), and should not be relied upon without careful cross-checking with existing domain knowledge.

6 Related work

To our knowledge, the *treatment-conditional* error setting we study is novel in this work. Techniques have been developed for addressing class-conditional [Menon et al., 2015, Scott et al., 2013], group-dependent [Wang et al., 2021], and instance-dependent [Xia et al., 2020] label noise models. Menon et al. [2015], Scott et al. [2013] and Northcutt et al. [2021] develop noise rate estimation approaches commonly used in label noise settings. Recently, De-Arteaga et al. [2021] propose a method leveraging inter-expert consistency to adjust for measurement error in proxy outcomes. Yet this approach is designed for observational, rather than counterfactual, prediction settings. Coston et al. [2020b] develop counterfactual modeling and evaluation approaches for decision-support settings without accounting for measurement error.

References

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

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

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR, 2018.

Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. Counterfactual predictions under runtime confounding. *Advances in Neural Information Processing Systems*, 33:4150–4162, 2020a.

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, 2020b.

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

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

Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. 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*, pages 100–111, 2021.

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.

Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

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

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.

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.

Ioannis Livieris, Tassos Mikropoulos, and Panagiotis Pintelas. A decision support system for predicting students' performance. *Themes in Science and Technology Education*, 9(1):43–57, 2016.

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

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

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

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

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.

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

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

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

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We highlight opportunities for future work in 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] https://anonymous.4open.science/r/counterfactual-decision-support-under-TCE-F53E/
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Details of experimental setup

Synthetic setup: Conducting evaluations for settings in which measurement error and treatment effects impact observed outcomes is challenging because often there is no way of recovering the target counterfactual outcome or the measurement error parameters. Therefore, we design an initial empirical evaluation via synthetic data to validate our proposed approach. Previous approaches have also used a uni-variate feature and protected attribute Coston et al. [2020b] or 2D synthetic data Natarajan et al. [2013], De-Arteaga et al. [2021]. We plan to extend to these settings and others with semi-synthetic and real-world data in later evaluations.

In all experiments, we use sinusoidal target class probability functions

$$\eta_0^*(x) = \begin{cases} .4 + .4\cos(9x + 5.5) & x \in [-1, -.61] \\ .5 + .3\sin(8x + .9) + .15\sin(10x + .2) + .05\sin(30x + .2) & x \in (-.61, .921] \\ x^3 & x \in (.921, 1] \end{cases} \quad (13)$$

and $\eta_1^*(x) = .5 + .5\sin(2.9x + .1)\ \forall x \in [-1, 1]$. We design $\eta_0^*(x)$ to be more challenging to estimate because we use it as our target quantity in this selective intervention setting. In all experiments, we use a linear propensity function $\pi(x) = .35x + .5, \forall x \in [-1, 1]$.

Model and hyperparameters: All experiments use a MLP implemented via PyTorch with layer sizes $(1, 40, 20, 4, 1)$ and a binary cross-entropy loss satisfying our strictly proper composite criteria for class probability estimation. We run experiments with $\alpha = .001$.

Evaluation. In all experiments, we split data 70/30 into training and validation folds. We evaluate accuracy with respect to Y_0^* on a held-out validation fold. We use accuracy as a performance measure rather than a metric such as AU-ROC because AU-ROC is immune to corruption from our error model (see Menon et al. [2015] for additional details).

Computing Environment Experiments were run on a MacBook Pro with 2.6 GHz 6-Core processor with 32 GB of RAM and Google Colab environment with standard runtime configuration.

A.2 Details of experiment configurations

- **Experiment 1**. Reported in Table 1. We run each setting with $N = 10000$ and average performance over 10 runs with 40 epochs of training per run.
- **Experiment 2**. Reported in Figure 3. We run each setting with $N = \{1000, 2000, 3000, 4000, 5000, 10000, 20000\}$ with 150 runs per setting and 30 epochs of training per round. Each round, we sample $\alpha_0, \beta_0 \sim U(0, .3)$.
- **Experiment 3**. Reported in Figure 4. We run each setting with $N = 10000$ with 15 runs per setting and 30 epochs of training per round. We vary α_d (β_d) from .1, .2, .3 and hold β_0 (α_0) fixed at 0. We then construct the surrogate loss $\tilde{\ell}$ with biased parameter estimates $\hat{\alpha}_0$, $\hat{\beta}_0$ parameter estimates.

A.3 Baselines

- $\hat{P}_\ell[Y = 1|X = x]$ (OBS)
- $\hat{P}_\ell[Y = 1|D = d, X = x]$ (FRM)
- $\hat{P}_{\tilde{\ell}}[Y = 1|D = d, X = x]$ (FRM-SL)
- $\hat{P}_\ell[Y_d = 1|X = x]$ (PROXY*)
- $\hat{P}_\ell[Y_d^* = 1|X = x]$ (TARGET*)

In causal inference settings, FRM is also referred to as a backdoor covariate adjustment or plug-in estimator Coston et al. [2020b,a], Pearl [2009].