Policy Comparison Under Unmeasured Confounding

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

Amanda Coston
Carnegie Mellon University
Pittsburgh, PA 15213
acoston@andrew.cmu.edu

Zhiwei Steven Wu*
Carnegie Mellon University
Pittsburgh, PA 15213
zstevenwu@cmu.edu

Kenneth Holstein*
Carnegie Mellon University
Pittsburgh, PA 15213
kjholste@cs.cmu.edu

Abstract

Predictive models are often introduced under the rationale that they improve performance over an existing decision-making policy. However, it is challenging to directly compare an algorithm against a status quo policy due to uncertainty introduced by confounding and selection bias. In this work, we develop a regret estimator which evaluates differences in classification metrics across decision-making policies under confounding. Theoretical and experimental results demonstrate that our regret estimator yields tighter regret bounds than existing auditing frameworks designed to evaluate predictive models under confounding. Further, we show that our regret estimator can be combined with a flexible set of causal identification strategies to yield informative and well-justified policy comparisons. Our experimental results also illustrate how confounding and selection bias contribute to uncertainty in subgroup-level policy comparisons. Our auditing framework provides a step towards the effective operationalization of regulatory frameworks calling for more direct assessments of predictive model efficacy.

1 Introduction

Regulatory frameworks such as the NIST AI Risk Management Framework (RMF) and Algorithmic Accountability Act of 2023 have called for model developers to more rigorously assess the efficacy of algorithmic systems prior to deployment in critical settings such as employment, healthcare, and education [Tabassi, 2023, Clarke, 2023]. A key question underlying efficacy evaluations involves whether an algorithm offers an improvement over an existing decision-making policy. However, reliably comparing an algorithm against the status quo is challenging due to uncertainty introduced by confounding and selection bias. Policy comparisons can be especially challenging among subpopulations that received low selection rates under an existing policy due to limited outcome data.

In this work, we support the operationalization of newly developed regulatory frameworks by characterizing and reducing uncertainty from confounding in policy comparisons. Given a proposed algorithmic policy and observational data collected under a status quo policy, our approach bounds the difference (i.e., regret) in predictive performance measures such as the false negative rate across policies. We provide tight asymptotic regret bounds which characterize the most informative comparison possible at a given level of confounding. We develop a one step difference-based regret estimator which yields tight regret bounds by isolating uncertainty relevant to the policy comparison. We show

* Equal Contribution.

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

that our approach yields more informative regret bounds than an existing two step auditing framework (i.e., [Rambachan et al., 2022]), which requires (1) bounding the predictive performance of each policy individually under confounding, then (2) taking a difference in upper and lower bounds across policies. Our estimator can be combined with a flexible set of causal identification strategies such as worst-case Manski-style bounds [Manski and Pepper, 1998] and an instrumental variable [Lakkaraju et al., 2017, Kleinberg et al., 2015, Chen et al., 2023]. This allows practitioners to conduct the most informative policy audit possible given a set of well justified causal assumptions.

We conduct a synthetic experiment assessing the bounds recovered by our approach. Results show that our one step difference-based regret estimator yields more informative bounds which can support qualitatively stronger conclusions when comparing decision-making policies under confounding (e.g., by ruling out a regret interval that contains zero). Our evaluation also illustrates how confounding contributes to uncertainty in subgroup-level policy comparisons by yielding wider bounds among subpopulations with lower selection rates under the status quo policy. In summary, we offer the following three contributions:

- We formulate the problem of policy comparison under confounding (§ 2). We provide several examples of policy-relevant auditing settings captured by our framework.
- We develop a regret estimator which reduces uncertainty in policy comparisons (§ 3.1). We show that our technique is compatible with multiple causal identification strategies (§ 3.2).
- We conduct an experimental evaluation validating our regret estimator with synthetic data (§ 4). Results show that our estimator reduces uncertainty in policy comparisons across multiple identification strategies, at both the population and subgroup level.

2 Problem formulation

Let $\pi_0 : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{A}$ be a status quo decision-making policy mapping measured covariates $X \in \mathcal{X} \in \mathbb{R}^d$ and unmeasured confounders $U \in \mathcal{U} \in \mathbb{R}^w$ to a binary action $\mathcal{A} \in \{0, 1\}$. This existing decision-making policy (e.g., a physician or judge making decisions without a risk assessment) observes recorded features X available for modeling (e.g., electronic medical records), as well as unobserved side information U (e.g., real-time test results [Mullainathan and Obermeyer, 2019]). Let $\pi : \mathcal{X} \rightarrow \mathcal{A}$ be a proposed algorithmic policy, which observes measured covariates but does not have access to side information. We take D^{π_0}, T^{π} to be random variables indicating actions selected under the status quo and updated policies, respectively.[1] A key *policy comparison* question involves whether π constitutes an improvement over the status quo π_0, as measured by predictive performance measures of interest.

In particular, let $Y(a) \in \{0, 1\}$ be the potential outcome which *would occur* under action $A = a$ [Rubin, 2005] and let $Y = Y(D^{\pi_0})$ be outcomes observed under the status quo policy. Given observational data $O = \{(X_i, T_i^{\pi}, D_i^{\pi_0}, Y_i) : i = 1, ..., n\}$, we would like to evaluate the *regret* of π against π_0, defined as a difference in predictive performance measures such as the false negative rate[2]

$$R(\pi, \pi_0) = p(T^{\pi} = 0 \mid Y(1) = 1) - p(D^{\pi_0} = 0 \mid Y(1) = 1). \tag{1}$$

For example, if $Y(1)$ were positive disease diagnosis given admission of a medical test ($a = 1$), $R(\pi, \pi_0) < 0$ would indicate that π yields fewer missed diagnoses than the status quo π_0. However, because counterfactual outcomes $Y(1 - D^{\pi_0})$ are unobserved in observational data, the policy regret is *partially identified* within an interval $\mathcal{H}(R(\pi, \pi_0)) = [\underline{R}(\pi, \pi_0), \overline{R}(\pi, \pi_0)]$. Our goal in this work is to recover the tightest (i.e., most informative) regret interval possible given observational data.

2.1 Framework examples

Our framework captures an extensive set of auditing scenarios of interest to model developers and regulators. We briefly list a few examples of policy comparisons supported by our framework:

[1]We sometimes omit policy superscripts over random variables to ease notation.

[2]For ease of exposition, we focus on false negative rate regret in the main text and include regret bounds for additional predictive performance measures in Appendix A.2.

- Let π_0 be an existing human-only decision-making policy (e.g., judicial decisions [Kleinberg et al., 2018], physician testing decisions [Mullainathan and Obermeyer, 2019], insurance application approvals [Lakkaraju et al., 2017]). Let π be an algorithmic decision support (ADS) tool which recommends actions based on thresholded risk predictions (e.g., [Obermeyer et al., 2019, Chouldechova et al., 2018, Angwin et al., 2016]). While humans will retain final autonomy over decisions under the updated policy, we would like to assess whether algorithmic recommendations alone constitute an improvement in predictive performance measures over human-only decisions (e.g., [Rambachan et al., 2021]).
- Let π_0 be a status quo policy in which humans make decisions with the support of ADS recommendations (e.g., social workers making child welfare hotline screening decisions with the support of a predictive model [Chouldechova et al., 2018]). Let π be the actions which *would have been taken* by an algorithmic policy alone *without* human intervention. We would like to assess performance differences across the deployed human+algorithm policy versus the algorithmic policy alone (e.g., [Cheng et al., 2022]).
- Let π_0 be a rule-based policy which can be manually overridden when justified by context-dependent (i.e., unobserved) situational factors (e.g., MELD score for liver transplantation [Kamath et al., 2001], Supplemental Nutrition Assistance Program eligibility [Cady, 2022]). We would like to assess whether an updated scoring system π would improve predictive performance measures over the existing scoring system [Ben-Michael et al., 2021].

3 Methodology

In § 3.1, we derive asymptotic bounds on the FNR policy regret. We leverage these bounds to propose a one step difference-based regret estimator, which we show yields tighter asymptotic regret bounds than existing auditing strategies. In § 3.2, we show that a flexible set of causal identification strategies can be applied to our one step estimator to yield well-justified and informative regret bounds. Proofs are included in Appendix A.1.

3.1 Asymptotic policy regret bounds

We begin by providing a theorem which characterizes the tightest bounds possible on the FNR regret, given asymptotic data. This form of result bounding a target quantity (i.e., FNR policy regret) within an interval given asymptotic data is also known as partial identification in causal inference literature.

Theorem 3.1. *Let* $v_{td} = p(T = t, D = d, Y(1) = 1)$*, with* $v_{t0} \in [\underline{v}_{t0}, \overline{v}_{t0}]$*. Then the FNR policy regret is partially identified within the interval* $\mathcal{H}(R(\pi, \pi_0)) = [\underline{R}(\pi, \pi_0), \overline{R}(\pi, \pi_0)]$*, with*

$$\begin{aligned} \underline{R}(\pi, \pi_0) &= \frac{v_{01} - \overline{v}_{10}}{\gamma + \overline{v}_{10} + v_{01} + v_{11}}, \quad \gamma = \begin{cases} \overline{v}_{00}, & v_{01} > \overline{v}_{10} \\ \underline{v}_{00}, & v_{01} \leq \overline{v}_{10} \end{cases}, \\ \overline{R}(\pi, \pi_0) &= \frac{v_{01} - \underline{v}_{10}}{\gamma + \underline{v}_{10} + v_{01} + v_{11}}, \quad \gamma = \begin{cases} \underline{v}_{00}, & v_{01} > \underline{v}_{10} \\ \overline{v}_{00}, & v_{01} \leq \underline{v}_{10} \end{cases} \end{aligned} \tag{2}$$

This result decomposes regret into comparison quadrants $v_{td} = p(T = t, D = d, Y(1) = 1)$. The v_{t1} terms are identifiable given observational data. However, the v_{t0} terms are partially identified within the interval $[\underline{v}_{t0}, \overline{v}_{t0}]$ because potential outcomes Y(1) are unobserved when $D = 0$. Therefore, these terms are the key driver of uncertainty in asymptotic regret bounds. We refer Theorem 3.1 as a *one step approach* because it directly bounds the performance difference across decision policies.

An alternative *two step approach* for deriving regret bounds involves bounding the predictive performance of each policy separately (e.g., via [Rambachan et al., 2022]), then estimating the overall regret by taking a difference across policy-specific performance intervals. However, the following result shows that this two step approach yields less informative bounds than our one step estimator, so long as $v_{11} > 0$ and $\underline{v}_{00} \neq \overline{v}_{00}$.

Theorem 3.2. *Let FNR*(π) *and FNR*(π_0) *be bounded as in Proposition (A.1) and (A.2) respectively, and let* $R(\pi, \pi_0)$ *be bounded as in Proposition (3.1). Assume that* $\overline{v}_{00} > \underline{v}_{00}$ *and* $v_{11} > 0$*. Then the* *following bounds hold on* $\mathcal{H}(R(\pi, \pi_0))$.

$$\begin{aligned} \underline{FNR}(\pi) - \overline{FNR}(\pi_0) < \underline{R}(\pi, \pi_0) \\ \overline{FNR}(\pi) - \underline{FNR}(\pi_0) > \overline{R}(\pi, \pi_0) \end{aligned} \tag{3}$$

Intuitively, this result holds because our one step regret bound isolates uncertainty relevant to the policy comparison. Whereas a two step estimator inherits uncertainty from the v_{00} term in the numerator of both policy-specific bounds, this uncertainty cancels in the one step difference-based estimator.

3.2 Partial identification framework

Our one step difference estimator can be combined with a flexible set of causal identification strategies to yield informative and well-justified regret estimates. When no assumptions on confounding are reasonable, a worst case Manski-style bound ([Manski and Pepper, 1998]) can be derived by plugging $\underline{v}_{t0} = 0, \overline{v}_{t0} = p(T = t, D = 0)$ into the one step regret interval (Theorem 3.1).

While worst case bounds are unlikely to be informative in most settings, we can also leverage additional causal assumptions to tighten bounds on v_{t0}. This will translate to more informative downstream regret estimates. For example, a popular partial identification framework for risk assessment evaluation under confounding involves leveraging random assignment of cases to decision-makers as an instrument [Lakkaraju et al., 2017, Kleinberg et al., 2018, Rambachan et al., 2022].

Assumption 1. *(Random instrument) Let Z be an instrument satisfying the following conditions*

1. *Relevance:* $Z \not\perp\!\!\!\perp D \mid X$
2. *IV independence:* $Z \perp\!\!\!\perp U \mid X$
3. *Exclusion restriction:* $Z \perp\!\!\!\perp Y \mid D, X, U$

This instrument has previously been used to bound predictive performance metrics [Rambachan et al., 2022], social welfare [Rambachan et al., 2021], and failure rates [Lakkaraju et al., 2017] of predictive models under confounding. In the following result, we leverage this instrument to establish sharp bounds on the unobserved comparison quadrants v_{t0}.

Theorem 3.3. *Assume that 1 holds. Let* $\mu_d(x; t, z) = \mathbb{E}[Y(1) \mid D = d, X = x, T = t, Z = z]$ *and* $e_d(x; t, z) = p(D = d \mid X = x, T = t, Z = z)$*. Then* $\forall t \in \{0, 1\}$*, unobserved comparison quadrants* v_{t0} *are bounded by*

$$\underline{v}_{t0} = \sum_{x \in X} \sum_{z \in Z} \max \left\{ \frac{\underline{\mu}(x; t) - \mu_1(x; t, z) \cdot e_1(x; t, z)}{e_0(x; t, z)}, 0 \right\} \cdot p(Z = z, T = t, D = 0, X = x),$$

$$\overline{v}_{t0} = \sum_{x \in X} \sum_{z \in Z} \min \left\{ \frac{\overline{\mu}(x; t) - \mu_1(x; t, z) \cdot e_1(x; t, z)}{e_0(x; t, z)}, 1 \right\} \cdot p(Z = z, T = t, D = 0, X = x)$$

$\underline{\mu}(x; a) = \max_{\tilde{z} \in \mathcal{Z}} \{\mu_1(x; t, \tilde{z}) e_1(x; t, \tilde{z})\}$, $\bar{\mu}(x; t) = \min_{\tilde{z} \in \mathcal{Z}} \{e_0(x; t, \tilde{z}) + \mu_1(x; t, \tilde{z}) e_1(x; t, \tilde{z})\}$.

In comparison to prior uses of this instrument, our setting requires learning a propensity function $e_d(x; t, z)$ and outcome regression $\mu_d(x; t, z)$ conditional on both policy recommendations D and T. IV-based estimates can then be leveraged by our one step regret estimator (e.q. 22) to yield an overall regret bound. In the future, it may be possible to apply a broader set of identification frameworks (e.g., sensitivity models such as the *marginal sensitivity model* [Tan, 2006] or Rosenbaum's τ-sensitivity model [Rosenbaum, 1987]) to bound v_{t0} when an instrument is unavailable.

4 Experiment

We conduct a synthetic experiment assessing the improvement in regret bounds recovered by our one step estimator. Our synthetic evaluation allows us to introduce confounding of known magnitude by varying which predictors are available to the status quo policy and the updated algorithmic policy. Our

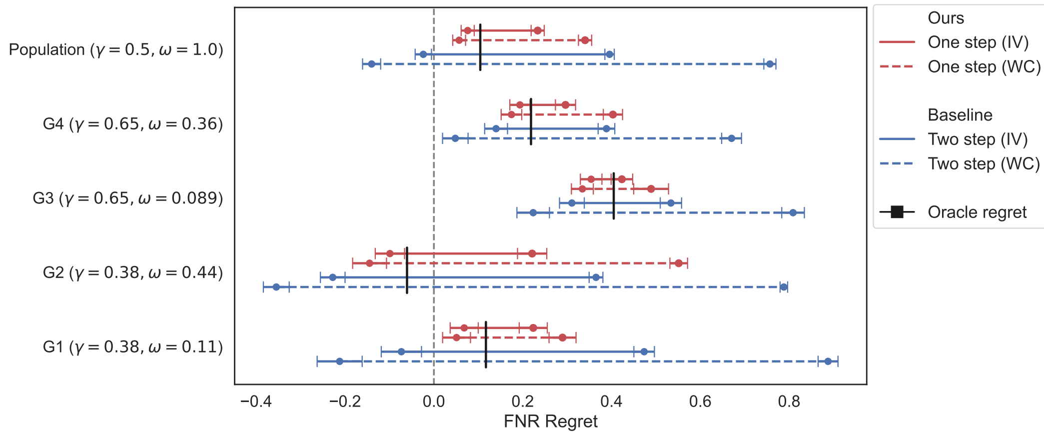

Figure 1: Our one step estimator narrows regret bounds across both instrumental variable (IV) and worst case (WC) partial identification strategies. Top row indicates full population bounds, while lower rows show regret over subgroups of varying selection rates (γ) and sizes (ω). Horizontal bars indicate asymptotic regret bounds under confounding. We show statistical uncertainty over $N = 10$ runs by plotting 95% confidence intervals centered at each upper and lower asymptotic bound.

synthetic setup also allows us to compare estimated bounds to ground-truth policy regret. This would not normally be feasible via real-world observational datasets because the magnitude of unmeasured confounding is typically unknown.

We compare the size of bounds resulting from our one step FNR regret estimator against a two step baseline. We partially identify unobserved outcome probabilities via a no assumptions worst case (**WC**) bound with $v_{t0} \in [0, p(T = t, D = 0)]$ and an instrumental variable (**IV**) bound as defined in Theorem 3.3. Given our focus on identification, we assume knowledge of oracle outcome and propensity functions when estimating IV bounds. Constructing finite sample estimators of these IV regret bounds is an important next step for practical application of our techniques.

4.1 Data generating process

We draw five covariates from a multivariate Gaussian $X \in \mathbb{R}^5 \sim \mathcal{N}(0, \Sigma)$ where Σ is the 5×5 identity matrix. We take $Z \sim \lfloor U(0, 10) \rfloor$ to be an instrumental variable with ten finite values (i.e., loan officers randomly assigned to cases). $X = (X_1, X_2, X_3)$ are measured covariates observed by the propensity function, conditional outcome functions, and algorithmic policy, while $U = (X_4, X_5)$ are confounders unobserved by algorithmic policy. We define the propensity function $e_1(X, U, Z)$, conditional outcome functions $\mu_d(X, U)$, and updated policy $\pi(X)$ as follows

$$\begin{aligned}
e_1(X, U, Z) &:= \sigma(-1.1 + 1.7 \cdot X_1 + .01 \cdot X_2 + 1.5 \cdot X_3 + .2 \cdot X_4 + 0.2 \cdot X_5 + .25 \cdot Z) \\
\mu_1(X, U) &:= \sigma(1.3 + 0.8 \cdot X_1 + .02 \cdot X_2 + 0.9 \cdot X_3 + 0.3 \cdot X_4 + 0.6 \cdot X_5) \\
\mu_0(X, U) &:= \sigma(1.5 + .05 \cdot X_1 + 1.8 \cdot X_2 + .05 \cdot X_3 + 0.2 \cdot X_4 + 0.5 \cdot X_5) \\
\pi(X) &:= \sigma(1.5 + .05 \cdot X_1 + 1.8 \cdot X_2 + .05 \cdot X_3 + .05 \cdot X_3)
\end{aligned} \tag{4}$$

where σ is the sigmoid function. The $X_1, ..., X_5$ coefficients in the function definitions above were selected to introduce (1) subgroup-level heterogeneity in ground-truth regret estimates and (2) an improvement in bounds from heterogeneity in the instrument Z.

The oracle probability of $Y(1)$ can be given by $\mu(X) = e_1(X, Z) \cdot \mu_1(X) + (1 - e_1(X, Z)) \cdot \mu_0(X)$. We sample data by drawing $D \sim \text{Bern}(e_1(X, Z))$, $T \sim \text{Bern}(\pi(X))$, $Y \sim \text{Bern}(\mu(X))$. We estimate one step (e.q. 22) and two-step regret by computing empirical estimates of comparison quadrants $\hat{v}_{t,d}$, then plugging these into the relevant estimator. To examine subgroup-level regret estimates, we define two protected attributes with two levels each. The first correlates with X_1, while the second correlates with X_2. This yields four intersectional subgroups G1-G4 with varying sizes (ω) and selection rates under the status quo policy (γ).

4.2 Results

Figure 1 compares the size of bounds returned by a one step and two step regret estimator across worst case (WC) and instrumental variable (IV) partial identification strategies. Results are averaged over 10 runs with $N = 5000$ samples each. Overall, this plot shows a net increase in the FNR under π, with more nuanced differences at the subgroup level (e.g., G2 experiences an FNR *decrease* under π, while G3 experiences a *large increase* in FNR under π). Our one step regret estimator narrows bounds across all subgroups and partial identification strategies. Surprisingly, we also observe that one step WC bounds are tighter than the two step IV bounds in several cases (e.g., population, G1, G3). This indicates that, compared to a two step estimator, a one step estimator can sometimes yield more reduction in uncertainty while making *no additional assumptions* on confounding. Critically, tighter regret bounds can support stronger conclusions when comparing decision policies. In this case, our one step estimator supports the interpretation that π yields in a net FNR increase across the full population. This conclusion would not be warranted on the basis of a two step estimator because its regret interval contains zero.

At the subgroup level, we see that groups with lower selection rates under the status quo policy (e.g., G1, G2) have wider asymptotic bounds than those with higher selection rates (e.g., G3, G4). This indicates that the key driver of uncertainty at the asymptotic level is confounding. We can also examine statistical uncertainty in subgroup regret estimates by inspecting the size of confidence intervals. We observe that smaller subgroups (e.g., G1, G3) have larger variance in regret estimates than larger subgroups (e.g., G2, G4, full population). We can isolate subgroup size (ω) as the source of this uncertainty because we fix selection rates across subgroups such that G1(γ)=G2(γ), G3(γ)=G4(γ). Taken together, these results indicate that distinct sources of uncertainty can complicate subgroup-level policy comparisons.

5 Discussion

In this work, we have proposed an approach for comparing decision-making policies under uncertainty from confounding. We have theoretically and empirically shown that our one step regret estimator reduces uncertainty in policy comparisons by narrowing bounds on regret across decision-making policies. In the future, it will be important to extend our identification framework to be compatible with a broader set of causal assumptions (e.g., the marginal sensitivity model [Tan, 2006], Rosenbaum model [Rosenbaum, 1987], and mean outcome sensitivity model [Rambachan et al., 2022]) and examine differences in one step and two step bounds across a wider set of predictive performance measures. Additionally, while this work has primarily focused on asymptotic identification of regret (i.e., characterizing uncertainty under given infinite data), it will also be important to develop practical finite sample regret estimators and analyze their convergence properties.

More broadly, we hope to extend these approaches into a flexible toolkit which supports more robust assessments of predictive model efficacy. While the NIST AI Risk Management Framework cites model reliability and validity as foundational characteristics of trustworthy AI systems, no existing software packages available via the RMF Playbook are designed to directly assess the impacts of confounding and related challenges on decision-making policy comparisons [Tabassi, 2023]. This presents a pressing barrier to the effective operationalization of regulatory guidelines governing the introduction of automated decision-making systems. Our work offers one step towards a better set of tools for practical efficacy evaluations.

6 Related work

Recent work has drawn attention to a number of technical challenges which can limit the efficacy of predictive models in real-world deployments [Wang et al., 2022, Coston et al., 2023, Raji et al., 2022, Hutchinson et al., 2022]. In this work, we propose evaluation tools to help practitioners conduct more reliable policy comparisons under uncertainty from counterfactual outcomes, selection bias, and unmeasured confounding. Lakkaraju et al. [2017] introduce the term *selective labels* to describe a setting in which outcomes are only observed under one of the possible screening decisions (e.g., re-arrest observed under pre-trial release). The authors introduce a contraction-based technique which estimates the failure rate of a model under confounding by leveraging heterogeneity in leniency rates across decision-makers. Kleinberg et al. [2018] leverages this approach to conduct an in-depth empirical comparison of judicial decision-making with a pre-trial risk assessment. More recent frameworks have formalized heterogeneity in screening rates as an instrumental variable, and used this approach to learn confounding robust models [Chen et al., 2023].

Most directly related to our work, [Rambachan et al., 2022] develop an auditing framework for learning and evaluating predictive models in the presence of unmeasured confounding. This work offers an approach for *partially identifying* common predictive performance measures under a mean outcome sensitivity model (MOSM). In principle, this approach could be used to bound regret across classification metrics by partially identifying performance of each policy separately, then taking a difference in upper and lower bounds. We include this two step approach as a baseline in our experiment. [Rambachan et al., 2021] leverages a similar set of tools to identify systematic prediction mistakes in a status quo decision-making policy under uncertainty outcome utilities.

An additional body of work proposes techniques for learning and ranking decision-making policies under confounding and overlap violations. [Kallus and Zhou, 2021, 2018] learn minimax-optimal policies under unobserved confounding which are guaranteed to provide no performance reduction in comparison to a status quo observational policy. [Ben-Michael et al., 2021] learns robust decision-making policies in cases where some subpopulations had no chance of selection under the status quo policy (i.e., *overlap violations*). [Zhang et al., 2021] develop an approach for ranking individualized treatment assignment rules under confounding. Additional work in off policy evaluation aims to estimate the value function of decision-making policies under confounding, but does not focus on policy comparisons [Kallus and Zhou, 2020, Jung et al., 2020]. Most recently, Gao and Yin [2023] develops a deferral-based framework to route examples to decision-makers at *runtime* given unmeasured confounding. Like [Kallus and Zhou, 2021], this work models confounding via the marginal sensitivity model.

Our work differs from prior off-policy evaluation frameworks in two key ways. First, our approach is the first to support comparisons against a *confounded data-generating baseline policy*. Whereas prior work in the non-sequential setting assumes that the data-generating mechanism and baseline policy (i.e., treat all, treat none) differ [Zhang et al., 2021, Kallus and Zhou, 2021, Gao and Yin, 2023], we study a more realistic and challenging setting in which the confounded baseline policy being bench-marked against generated the observational auditing dataset. This formulation is critical for supporting policy comparisons of interest in key regulatory settings (i.e., examples described in Section 2.1). Second, while prior work is tightly coupled to a specific causal partial identification strategy (i.e., the marginal sensitivity model [Kallus and Zhou, 2021] or Rosenbaum's τ-sensitivity model [Zhang et al., 2021]), our framework generalizes to a broad set of identification approaches. This offers flexibility required to support a diverse set of auditing contexts.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. 2016.

Eli Ben-Michael, D James Greiner, Kosuke Imai, and Zhichao Jiang. Safe policy learning through extrapolation: Application to pre-trial risk assessment. *arXiv preprint arXiv:2109.11679*, 2021.

Rob Cady. A quick guide to snap eligibility and benefits-recent changes to snap benefits and. 2022.

Jian Chen, Zhehao Li, and Xiaojie Mao. Learning under selective labels with data from heterogeneous decision-makers: An instrumental variable approach. *CoRR*, 2023.

Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. How child welfare workers reduce racial disparities in algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2022.

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.

Yvette D. Clarke. H.r. 5628 (ih) - algorithmic accountability act of 2023. `https://www.govinfo.gov/app/details/BILLS-118hr5628ih`, 2023. Accessed: 2023-10-01.

Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 690–704. IEEE, 2023.

Ruijiang Gao and Mingzhang Yin. Confounding-robust policy improvement with human-ai teams. *arXiv preprint arXiv:2310.08824*, 2023.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1859–1876, 2022.

Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. Bayesian sensitivity analysis for offline policy evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 64–70, 2020.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31, 2018.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in neural information processing systems*, 33:22293–22304, 2020.

Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

Patrick S Kamath, Russell H Wiesner, Michael Malinchoc, Walter Kremers, Terry M Therneau, Catherine L Kosberg, Gennaro D'Amico, Rolland E Dickson, and Ray W Kim. A model to predict survival in patients with end–stage liver disease. *Hepatology*, 33(2):464–470, 2001.

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

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.

Charles F Manski and John V Pepper. Monotone instrumental variables with an application to the returns to schooling, 1998.

Sendhil Mullainathan and Ziad Obermeyer. *A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions*. National Bureau of Economic Research, 2019.

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

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

Ashesh Rambachan, Amanda Coston, and Edward Kennedy. Counterfactual risk assessments under unmeasured confounding. *arXiv preprint arXiv:2212.09844*, 2022.

Ashesh Rambachan et al. Identifying prediction mistakes in observational data. *Harvard University*, 2021.

Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

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

Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). 2023.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *Available at SSRN*, 2022.

Bo Zhang, Jordan Weiss, Dylan S Small, and Qingyuan Zhao. Selecting and ranking individualized treatment rules with unmeasured confounding. *Journal of the American Statistical Association*, 116(533):295–308, 2021.

A Appendix

A.1 Omitted results and proofs

Proposition A.1. *Let* $v_{td} = p(T = t, D = d, Y(1) = 1)$*, with* $v_{t0} \in [\underline{v}_{t0}, \overline{v}_{t0}]$*. Then FNR*$(\pi) = \mathbb{E}[1 - T^{\pi} \mid Y(1) = 1]$ *is bounded by*

$$\underline{FNR}(\pi) = \frac{\underline{v}_{00} + v_{01}}{\underline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}}, \quad \overline{FNR}(\pi) = \frac{\overline{v}_{00} + v_{01}}{\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}}. \tag{5}$$

Proof.

$$\begin{aligned}
FNR(\pi) &= \mathbb{E}[1 - T^{\pi} \mid Y(1) = 1] && (6)\\
&= p(T = 0 \mid Y(1) = 1) && (7)\\
&= \frac{p(T = 0, Y(1) = 1)}{p(Y(1) = 1)} && (8)\\
&= \frac{v_{01} + v_{00}}{v_{00} + v_{10} + v_{01} + v_{11}} && (9)\\
& && (10)
\end{aligned}$$

where the overall FNR is bounded under $v_{t0} \in [\underline{v}_{t0}, \overline{v}_{t0}]$ via

$$\frac{\underline{v}_{00} + v_{01}}{\underline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} \leq \frac{v_{01} + v_{00}}{v_{00} + v_{10} + v_{01} + v_{11}} \leq \frac{\overline{v}_{00} + v_{01}}{\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}}$$

□

Proposition A.2. *Let* $v_{td} = p(T = t, D = d, Y(1) = 1)$*, with* $v_{t0} \in [\underline{v}_{t0}, \overline{v}_{t0}]$*. Then FNR*$(\pi_0) = \mathbb{E}[1 - D^{\pi_0} \mid Y(1) = 1]$ *is bounded by*

$$\underline{FNR}(\pi_0) = \frac{\underline{v}_{00} + \underline{v}_{10}}{\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}}, \quad \overline{FNR}(\pi_0) = \frac{\overline{v}_{00} + \overline{v}_{10}}{\overline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}}. \tag{11}$$

The bounds on $FNR(\pi_0)$ follow by a similar argument. We now prove Theorem 3.1.

Proof. Decomposing our expression for the FNR regret, we see that

$$\begin{aligned}
R(\pi, \pi_0) &= \mathbb{E}[1 - T^{\pi} \mid Y(1) = 1] - \mathbb{E}[1 - D^{\pi_0} \mid Y(1) = 1] && (12)\\
&= p(T = 0 \mid Y(1) = 1) - p(D = 0 \mid Y(1) = 1) && (13)\\
&= \frac{p(T = 0, Y(1) = 1)}{p(Y(1) = 1)} - \frac{p(D = 0, Y(1) = 1)}{p(Y(1) = 1)} && (14)\\
&= \frac{v_{01} + v_{00}}{v_{00} + v_{10} + v_{01} + v_{11}} - \frac{v_{10} + v_{00}}{v_{00} + v_{10} + v_{01} + v_{11}} && (15)\\
&= \frac{v_{01} - v_{10}}{v_{00} + v_{10} + v_{01} + v_{11}} && (16)
\end{aligned}$$

where v_{t1} is identifiable from observational data and $v_{t0} \in [\underline{v}_{t0}, \overline{v}_{t0}]$ is partially identified.

$$\frac{v_{01} - \overline{v}_{10}}{\gamma + \overline{v}_{10} + v_{01} + v_{11}} \leq \frac{v_{01} - v_{10}}{v_{00} + v_{10} + v_{01} + v_{11}} \leq \frac{v_{01} - \underline{v}_{10}}{\gamma + \underline{v}_{10} + v_{01} + v_{11}}$$

□

We now prove Theorem 3.2. A similar argument can be applied to prove asymptotic bounds on the regret of the one-step FPR, TNR, and TPR estimator.

Proof. We will begin by showing that $\overline{\text{FNR}}(\pi) - \underline{\text{FNR}}(\pi_0) > \overline{R}(\pi, \pi_0)$. Case 1: $v_{01} > \underline{\mathrm{v}}_{10}$.

$$\frac{\overline{v}_{00} + v_{01}}{\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} - \frac{\underline{v}_{00} + \underline{v}_{10}}{\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} > \frac{v_{01} - \underline{v}_{10}}{\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}}$$

$$\frac{\overline{v}_{00} + v_{01}}{\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} - \frac{\underline{v}_{00} + \underline{v}_{10}}{\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} - \frac{v_{01} - \underline{v}_{10}}{\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} > 0$$

$$\frac{(v_{11} + \underline{v}_{10}) \cdot (\overline{v}_{00} - \underline{v}_{00})}{(\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}) \cdot (\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11})} > 0$$

Case 2: $v_{01} \leq \underline{v}_{01}$.

$$\frac{\overline{v}_{00} + v_{01}}{\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} - \frac{\underline{v}_{00} + \underline{v}_{10}}{\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} > \frac{v_{01} - \underline{v}_{10}}{\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}}$$

$$\frac{\overline{v}_{00} + v_{01}}{\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} - \frac{\underline{v}_{00} + \underline{v}_{10}}{\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} - \frac{v_{01} - \underline{v}_{10}}{\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}} > 0$$

$$\frac{(v_{11} + v_{01}) \cdot (\overline{v}_{00} - \underline{v}_{00})}{(\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}) \cdot (\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11})} > 0$$

In both cases, the left hand term must by positive by the assumption that $\overline{v}_{00} > \underline{v}_{00}$ and $v_{11} > 0$. Next, we will show that $\underline{\text{FNR}}(\pi) - \overline{\text{FNR}}(\pi_0) < \underline{R}(\pi, \pi_0)$. Case 1: $v_{01} > \overline{v}_{10}$.

$$\frac{\underline{v}_{00} + v_{01}}{\underline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} - \frac{\overline{v}_{00} + \overline{v}_{10}}{\overline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} < \frac{v_{01} - \overline{v}_{10}}{\overline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}}$$

$$\frac{\underline{v}_{00} + v_{01}}{\underline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} - \frac{\overline{v}_{00} + \overline{v}_{10}}{\overline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} - \frac{v_{01} - \overline{v}_{10}}{\overline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} < 0$$

$$\frac{(v_{11} + \overline{v}_{10}) \cdot (\underline{v}_{00} - \overline{v}_{00})}{(\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}) \cdot (\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11})} < 0$$

Case 2: $v_{01} < \overline{v}_{01}$.

$$\frac{\underline{v}_{00} + v_{01}}{\underline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} - \frac{\overline{v}_{00} + \overline{v}_{10}}{\overline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} < \frac{v_{01} - \overline{v}_{10}}{\underline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}}$$

$$\frac{\underline{v}_{00} + v_{01}}{\underline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} - \frac{\overline{v}_{00} + \overline{v}_{10}}{\overline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} - \frac{v_{01} - \overline{v}_{10}}{\underline{v}_{00} + \overline{v}_{10} + v_{01} + v_{11}} < 0$$

$$\frac{(v_{11} + v_{01}) \cdot (\underline{v}_{00} - \overline{v}_{00})}{(\underline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11}) \cdot (\overline{v}_{00} + \underline{v}_{10} + v_{01} + v_{11})} < 0$$

In both cases, the left hand term must by negative by the assumption that $\overline{v}_{00} > \underline{v}_{00}$ and $v_{11} > 0$. ☐

We now prove Theorem 3.3.

Proof. Observe that $\mu(x; a) = \mu(x; t, z) = \mu_1(x; t, z) \cdot e_1(x; t, z) + \mu_0(x; t, z) \cdot e_0(x; t, z)$, where the first equality holds by IV independence (1) and the second by iterated expectations. Applying worst-case bounds on $\mu_0(x; t, z) \in [0, 1]$ (Manski and Pepper [1998]), we have that $\forall z \in \mathcal{Z}$

$$\mu_1(x; t, z)e_1(x; t, z) \leq \mu(x; t, z) \leq e_0(x; t, z) + \mu_1(x; t, z)e_1(x; t, z). \tag{17}$$

Because $Y(0), Y(1) \perp\!\!\!\perp Z \mid X$, we have that $\mathbb{E}[Y(1) \mid X = x] = \mathbb{E}[Y(1) \mid X = x, Z = z]$. This implies the intersection bound

$$\max_{\tilde{z} \in \mathcal{Z}} \{\mu_1(x; t, \tilde{z})e_1(x; t, \tilde{z})\} \leq \mu(x; t) \leq \min_{\tilde{z} \in \mathcal{Z}} \{e_0(x; t, \tilde{z}) + \mu_1(x; t, \tilde{z})e_1(x; t, \tilde{z})\} \tag{18}$$

Solving for $\mu_0(x;t,z)$ yields

$$
\begin{aligned}
\underline{\mu}_0(x;t,z) &= \max\left\{\frac{\underline{\mu}(x;t) - \mu_1(x;t,z)\cdot e_1(x;t,z)}{e_0(x;t,z)}, 0\right\},\\
\overline{\mu}_0(x;t,z) &= \min\left\{\frac{\overline{\mu}(x;t) - \mu_1(x;t,z)\cdot e_1(x;t,z)}{e_0(x;t,z)}, 1\right\}.
\end{aligned}
\tag{19}
$$

The result follows by marginalizing over Z.

□

A.2 Regret bounds on additional performance measures

The proof of one step regret bounds below follow by the same argument as the proof for Theorem 3.1.

Proposition A.3. *(TNR regret bounds) Let $w_{td} = p(T=t, D=d, Y(1)=0)$, with $w_{t0} \in [\underline{w}_{t0}, \overline{w}_{t0}]$. Then the TNR policy regret is partially identified within the interval $\mathcal{H}(R(\pi,\pi_0)) = [\underline{R}(\pi,\pi_0), \overline{R}(\pi,\pi_0)]$, with*

$$
\begin{aligned}
\underline{R}(\pi,\pi_0) &= \frac{w_{01} - \underline{w}_{10}}{\gamma + \underline{w}_{10} + w_{01} + w_{11}},\ \ \gamma = \begin{cases}\overline{w}_{00}, & w_{01} > \underline{w}_{10}\\ \underline{w}_{00}, & w_{01} \leq \underline{w}_{10}\end{cases},\\
\overline{R}(\pi,\pi_0) &= \frac{w_{01} - \overline{w}_{10}}{\gamma + \overline{w}_{10} + w_{01} + w_{11}},\ \ \gamma = \begin{cases}\overline{w}_{00}, & w_{01} > \overline{w}_{10}\\ \underline{w}_{00}, & w_{01} \leq \overline{w}_{10}\end{cases}
\end{aligned}
\tag{20}
$$

Proposition A.4. *(FPR regret bounds) Let $w_{td} = p(T=t, D=d, Y(1)=0)$, with $w_{t0} \in [\underline{w}_{t0}, \overline{w}_{t0}]$. Then the FPR policy regret is partially identified within the interval $\mathcal{H}(R(\pi,\pi_0)) = [\underline{R}(\pi,\pi_0), \overline{R}(\pi,\pi_0)]$, with*

$$
\begin{aligned}
\underline{R}(\pi,\pi_0) &= \frac{\underline{w}_{10} - w_{01}}{\gamma + \underline{w}_{10} + w_{01} + w_{11}},\ \ \gamma = \begin{cases}\overline{w}_{00}, & \underline{w}_{10} > w_{01}\\ \underline{w}_{00}, & \underline{w}_{10} \leq w_{01}\end{cases},\\
\overline{R}(\pi,\pi_0) &= \frac{\overline{w}_{10} - w_{01}}{\gamma + \overline{w}_{10} + w_{01} + w_{11}},\ \ \gamma = \begin{cases}\overline{w}_{00}, & \overline{w}_{10} > w_{01}\\ \underline{w}_{00}, & \overline{w}_{10} \leq w_{01}\end{cases}
\end{aligned}
\tag{21}
$$

Proposition A.5. *(TPR regret bounds) Let $v_{td} = p(T=t, D=d, Y(1)=1)$, with $v_{t0} \in [\underline{v}_{t0}, \overline{v}_{t0}]$. Then the FPR policy regret is partially identified within the interval $\mathcal{H}(R(\pi,\pi_0)) = [\underline{R}(\pi,\pi_0), \overline{R}(\pi,\pi_0)]$, with*

$$
\begin{aligned}
\underline{R}(\pi,\pi_0) &= \frac{\underline{v}_{10} - v_{01}}{\gamma + \underline{v}_{10} + v_{01} + v_{11}},\ \ \gamma = \begin{cases}\overline{v}_{00}, & \underline{v}_{10} > v_{01}\\ \underline{v}_{00}, & \underline{v}_{10} \leq v_{01}\end{cases},\\
\overline{R}(\pi,\pi_0) &= \frac{\overline{v}_{10} - v_{01}}{\gamma + \overline{v}_{10} + v_{01} + v_{11}},\ \ \gamma = \begin{cases}\overline{v}_{00}, & \overline{v}_{10} > v_{01}\\ \underline{v}_{00}, & \overline{v}_{10} \leq v_{01}\end{cases}
\end{aligned}
\tag{22}
$$