Bounding the Causal Impact of ML-assisted Decision-Making via Counterfactual Correctness

Jonathan Zhang[1] **Erik Skalnes**[1] **Jacob M. Chen**[1] **Michael Oberst**[1]

[1]Computer Science Dept., Johns Hopkins University, Baltimore, Maryland, USA

Abstract

Predictive machine learning (ML) models are increasingly used to aid human decision-makers across various high-risk domains such as healthcare and criminal justice. There is a growing recognition of the need to evaluate the causal impact of deploying these systems on downstream outcomes, such as patient survival or crime recidivism. Randomized control trials (RCTs) can provide high-quality evidence on the impact of a deployed model, but they run into a challenge: it is often infeasible to run repeated trials when models are updated or retrained to improve predictive performance. In this work, we present a partial-identification approach to using prior RCT data to construct bounds on the causal effect of a new model. The core innovation in our approach is to leverage assumptions relating fine-grained predictive accuracy to downstream outcomes. We do so via two monotonicity assumptions: first, on individual-level 'counterfactual correctness' (all else being equal, a correct prediction leads to non-inferior outcomes); and second, on the relation between subgroup predictive performance and outcomes, interpretable as an assumption regarding trust in model outputs. We demonstrate our method with a simulation study, illustrating how incorporating this information can lead to more informative bounds compared to prior work.

1 INTRODUCTION

Predictive machine learning (ML) models are being deployed more and more into high-risk decision-making domains, such as recidivism prediction for criminal justice or diagnostic aids for physicians [Travaini et al., 2022, Esteva et al., 2017]. In these domains, there is a growing recognition of the need to treat artificial intelligence (AI) systems as interventions, going beyond predictive accuracy to reason about their impact on decision-makers and downstream outcomes [Ben-Michael et al., 2025, Liu et al., 2025, Joshi et al., 2025, van Amsterdam et al., 2026].

In some settings, it is feasible to evaluate the impact of predictive systems via randomized control trials (RCTs), which are increasingly being used to assess ML/AI interventions in medicine [Shaukat et al., 2022, Yao et al., 2021, Upton et al., 2024, Gohil et al., 2024, Chen et al., 2025, Shaban et al., 2025] and criminal justice [Imai et al., 2023]. Such trials are often designed to evaluate the impact of deploying an ML/AI system as a discrete intervention, including impact on decision-makers and downstream outcomes, keeping in mind that the impact of predictive models is mediated by human behavior [Raji and Liu, 2025].

However, there are limits to the approach of viewing distinct ML models as distinct interventions. For example, it is typically infeasible to run a new RCT for every proposed update to a model. This infeasibility creates a challenge: If a randomized trial has been conducted for one model, and a new model (with higher predictive accuracy) is developed afterwards, what can we say about the expected impact of the latter model?

Prior work has made some steps towards addressing this gap. For instance, Nercessian et al. [2025] propose an adaptive trial design under the assumption that downstream outcomes do not differ between models whose predictions agree. Chen and Oberst [2025] propose a causal framework for partial identification that weakens this assumption, by allowing for differences in outcomes at the individual level, even if predictions are identical. They motivate this possibility from the perspective of user trust: For instance, a model that flags every patient as high-risk is likely to influence clinical actions less than a model with higher precision, even on the cases where they agree. To model this aspect of performance, they assume that residual variation in outcomes is driven by an observable per-model measure of global predictive

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:8026–8070.

performance (e.g., precision).

However, both of these frameworks fail to incorporate the accuracy of individual-level model predictions—in many scenarios, we might expect that correct predictions are no worse than suboptimal predictions. We refer to this notion as **counterfactual correctness**: Whether or not the predictions of a new model would have been correct on individual subjects in the original RCT. With that notion in mind, we make the following contributions in this work: (i) We extend the causal framework of Chen and Oberst [2025] to incorporate known 'correct' prediction labels: if a new model produces the correct prediction, we assume that outcomes are no worse than those of models that made incorrect predictions. (ii) We further extend this framework to consider conditional subgroup predictive performance as influencing outcomes, rather than a global measure of performance, reflecting the idea that users are likely to develop mental heuristics of where models perform well or poorly. (iii) We provide population-level upper and lower bounds under these assumptions, and further provide a consistent and asymptotically normal estimator of those bounds. Crucially, unlike Chen and Oberst [2025], we do so without assuming that model performance measures are known a priori, and allow for the fact that e.g., sub-group performance needs to be estimated from data. (iv) In a semi-synthetic simulation study, we illustrate that incorporating this information can lead to more informative bounds compared to prior work.

Our framework is most useful in settings where the "correct" predictions (which we often refer to as the "true labels") are observed for all individuals in the RCT, and explicitly assumes that predictions themselves do not influence the true label. These conditions may co-occur when the prediction task is fundamentally diagnostic, with the possibility of retrospective (even if expensive) assessment of the correct prediction for each individual. Our framework can be extended to settings where the true label is not always observed, under specific mechanisms of missingness. We discuss one of these examples in greater detail in Appendix G. We also discuss distinctions between our work and other related areas in Appendix F.

2 SETUP & CAUSAL MODEL

Notation: We use upper case letters X to denote random variables, lower case x to denote realizations, and calligraphic font $\mathcal{X}$ to denote the set of possible values. $\mathbf{1}\{E\}$ is the indicator function of an event E, equal to 1 if the event occurs and 0 otherwise. Unless otherwise noted, $\|\cdot\|$ refers to the $L^2(P)$ norm and $|\cdot|$ refers to the absolute value.

Problem Setup: We model the data-generating process as the directed acyclic graph (DAG) in Fig. 1a. A directed edge $A \to B$ means that A may cause B. No edge means A does not cause B. Each node is a random variable. $D \in \{1, \ldots, d\}$ are the arms of the RCT (e.g. with and without ML assistance). $\mathcal{T}$ is the set of models being trialed, with each model corresponding to a specific arm. The model we would like to evaluate (denoted π_e) is untrialed, so $\pi_e \notin \mathcal{T}$. $X \in \mathcal{X}$ are features used by the prediction model, $Z \in \mathcal{Z}$ is the "true label", and $\hat{Z} \in \mathcal{Z}$ is the predicted label. The performance of the model Π for individuals with covariates X is a real number $M \in \mathcal{M} \subseteq \mathbb{R}$, where $\mathcal{M}$ denotes the range of achievable metric values. $A \in \mathcal{A}$ is the action taken by the decision-maker, which we do not model explicitly and generally allow to be unobserved. Finally, $Y \in \mathbb{R}$ is the downstream outcome, influenced by A and X. We employ potential outcomes notation [Richardson and Robins, 2013] where $Y(\Pi = \pi)$ represents the counterfactual value of Y that would be observed had we set $\Pi = \pi$.

Example 2.1 (Diagnosis Recommendation System). To build intuition for our notation in the context of a concrete example, consider a machine learning model (π) designed to help radiologists make diagnoses. An RCT to determine whether model deployment improves patient outcomes might randomly assign physicians to either the control arm (no model, $D = 1$) or the treatment arm (model assistance, $D = 2$). The model takes in a chest x-ray, as well as other general patient information (X) as input, and returns a predicted diagnosis ($\hat{Z}$), such as pneumonia, pneumothorax, or healthy. This prediction may or may not agree with the true diagnosis (Z), which may be recorded later in time (e.g., at discharge). Over time, physicians will learn about the performance of the deployed models (M) on different types of patients. This learning can be captured quantitatively by calculating a per-subgroup metric on each of the models such as the sensitivity/specificity. Together, the knowledge of these variables will affect the physician's actions (A), in turn affecting some outcome of interest such as survival rate (Y). These actions are left unmodeled, and in general could represent a complex set of actions (e.g., specific language used in a radiology report). Our goal is to evaluate expected patient outcomes Y under the deployment of a new model π_e not included in the trial.

We now formally introduce our causal framework, shown graphically in Fig. 1.

Assumption 2.1 (Data Generating Process)**.** We assume that $(D, \Pi, X, Z, \hat{Z}, M, Y)$ are generated according to the following structural causal model (SCM) Pearl [2009], where all noise variables ϵ are mutually independent:

$$\begin{aligned} D &= f_D(\epsilon_D) & \Pi &= f_\Pi(D) \\ X, Z &= f_{XZ}(\epsilon_{XZ}) & \hat{Z} &= \Pi(X) \\ Y &= f_Y(X, Z, M, \hat{Z}, \epsilon_Y) & M &= f_M(X, \Pi) \end{aligned}$$

Note that our SCM allows for the possibility of various causal relationships between X and Z (e.g., where X causes

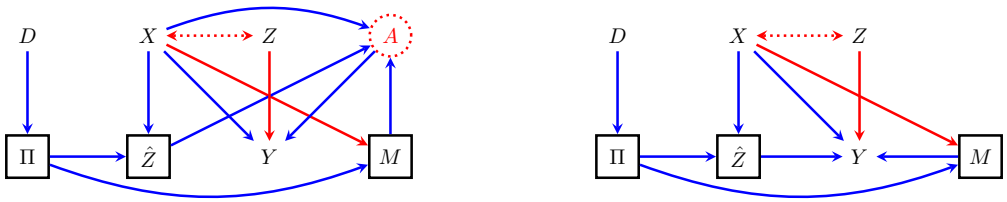


Figure 1: Fig. 1a is the causal DAG representing the structure of an RCT, where X denotes features, and D the arm of the RCT, which determines the model Π used for prediction. The model Π outputs predictions $\hat{Z}$ as a deterministic function of X, where Z is the true label. The dotted bi-directed edge between X, Z represents the possibility that X and Z share an unobserved common cause. M is a subgroup performance metric of Π, A is the decision-maker's actions, and Y is the downstream outcome. A solid square indicates the node is deterministic given its parents, and a dotted circle indicates the node is hidden. In Fig. 1b, we show the same graph after applying the latent projection operator on A, which only considers the observed variables in the RCT. Edges also present in prior work by Chen and Oberst [2025] are shown in blue (re-interpreting Chen and Oberst [2025] as having an implicit action A), while edges novel to our work are shown in red.

Z, vice versa, or where both are related by a common cause). We make a few additional observations: First, our causal model (reflected in Fig. 1b) does not include A, but can be seen as a projection of the directed acyclic graph (DAG) shown in Fig. 1a, where A mediates the effect of $\hat{Z}$ on Y. This mediation reflects our natural intuition that the decision-maker, not the model, is the final conduit of change, and serves as conceptual motivation for several of our assumptions. Second, we allow for X, Z to be arbitrarily related (e.g., by a common cause), but we assume that $\hat{Z}$ does not have any causal effect on Z. Finally, several elements are *deterministic* given their parents, $\hat{Z}$ and M in particular.

Our Goal: Consider a model $\pi_e \notin \mathcal{T}$. Our goal is to estimate the expectation of Y for π_e had it been in the RCT.

$$\mathbb{E}[Y(\pi_e)] = \mathbb{E}[Y(\hat{Z} = \pi_e(X), M = f_M(X, \pi_e))]. \quad (1)$$

This notation is used in conjunction with our SCM notation defined in Assump. 2.1. In particular, the right-hand side of Eq. (1) can be written as $\mathbb{E}\left[f_Y(X, Z, f_M(X, \pi_e), \pi_e(X), \epsilon_Y)\right]$,[1] where we set $\hat{Z} = \pi_e(X), M = f_M(X, \pi_e)$ and evaluate the expectation. We further illustrate this estimand with an example.

Example 2.1 (continuing from p. 2). Suppose we have data from a previous RCT on radiology models. The data records the values of Z and $\hat{Z}$ for all models on all patients. For each patient, we observe the value of Y under the deployment of a single model. In this context, given a new model, our goal would be to estimate its effect on patient survival (Y) without running a new RCT, leveraging only knowledge of the new model's predictions ($\hat{Z}$) on the data from that previous RCT.

[1]Throughout, we write f_Y with its arguments in the order given in Assump. 2.1, namely $(X, Z, M, \hat{Z}, \epsilon_Y)$.

3 STRUCTURAL ASSUMPTIONS AND FALSIFICATION TESTS

To draw conclusions about the downstream outcomes of an untrialed model, we need to make additional *structural* assumptions on the data generating process, beyond those implied by the causal graph. First, we formalize our notion of counterfactual correctness.

Assumption 3.1 (Counterfactual Correctness)**.** We assume that correct predictions cause no harm. That is, for all $m, m' \in \mathcal{M}$ and $z, z' \in \mathcal{Z}$ with $z' \neq z$,

$$Y(\hat{Z} = z, M = m, Z = z) \geq Y(\hat{Z} = z', M = m', Z = z)$$

This assumption states that, when predictions are correct, we don't expect them to harm outcomes.[2] While intuitive, this assumption may fail if e.g., physicians act contrary to predictions. That said, this assumption has testable implications, which can be assessed using RCT data.

Proposition 3.1 (Falsification of Assump. 3.1)**.** *Let $\pi_c, \pi_w \in \mathcal{T}$ be two trialed models (one of which will be "correct" and the other "wrong"), and let $\mathcal{X}_{c \neq w} := \{x \in \mathcal{X} \mid \pi_c(x) \neq \pi_w(x)\}$ be the set of covariates on which their predictions disagree. Consider the subpopulation where $X \in \mathcal{X}_{c \neq w}$ and $Z = \pi_c(X)$ (i.e., where π_c is correct and π_w incorrect). Then, the observation that outcomes are better under π_w,*[3]

$$\mathbb{E}[Y \mid X \in \mathcal{X}_{c \neq w}, \Pi = \pi_c, Z = \pi_c(X)]$$

[2]This inequality can be written equivalently under Assump. 2.1 as requiring that for all values of z, x, m, ϵ_Y and $z' \neq z$, $f_Y(x, z, m, z, \epsilon_Y) \geq f_Y(x, z, m', z', \epsilon_Y)$, where potential outcomes are taken to be pointwise functions of the noise variables.

[3]All falsification tests require that the relevant conditioning sets have non-zero probability of occurring. For this case, if D is uniformly distributed and independent of X, Z, it suffices that $P\big(X \in \mathcal{X}_{c \neq w}, Z = \pi_c(X)\big) > 0$.

$$< \mathbb{E}[Y \mid X \in \mathcal{X}_{c \neq w},\, \Pi = \pi_w,\, Z = \pi_c(X)]$$

falsifies Assump. 3.1.

This falsification test allows practitioners to use a simple hypothesis test (comparing two means) to assess if Assump. 3.1 is contradicted by RCT data. This falsification test, and others we propose, cannot conclusively demonstrate that an assumption is true, only that it is false, and passing the falsification tests does not validate the assumptions: The assumption may hold on RCT data but not more broadly across all future, untrialed models, and our assumptions are defined pointwise on potential outcomes, while these falsifications test for inequalities in expectation. Nonetheless, these tests have the appeal of being simple to implement. Proofs for all results, including the proof for the falsification tests, are provided in the Appendix.

Next, we introduce three additional assumptions adapted from Chen and Oberst [2025], but modified to better align with real-world conditions and our new data generating process. While the previous assumption dealt with outcomes under different predictions, the next assumptions focus on outcomes under the same prediction. We first introduce a notion of performance monotonicity, which was also explored in Chen and Oberst [2025].

Assumption 3.2 (Conditional Performance monotonicity)**.** Potential outcomes are monotonic in model performance relative to correctness: If $m_i < m_j$ for $m_i, m_j \in \mathcal{M}$, then for all $z \in \mathcal{Z}$ and all $\hat{z} \in \mathcal{Z}$ with $\hat{z} \neq z$,

$$Y(\hat{Z} = z, M = m_i, Z = z) \leq Y(\hat{Z} = z, M = m_j, Z = z)$$
$$Y(\hat{Z} = \hat{z}, M = m_i, Z = z) \geq Y(\hat{Z} = \hat{z}, M = m_j, Z = z)$$

This assumption states that, conditioned on correctness, potential outcomes are monotonic with respect to model performance. When the prediction is correct ($\hat{z} = z$), potential outcomes are monotonically increasing with respect to performance, and vice versa when the prediction is incorrect ($\hat{z} \neq z$). Conceptually, this assumption reflects the idea that decision-makers are more likely to trust higher performing models, and follow through on their predictions. When the prediction is correct, this increased trust leads to better outcomes, but when predictions are incorrect, this reliance leads to worse outcomes. Chen and Oberst [2025] considered a simplified form of this assumption, where improved performance always leads to better outcomes, which is likely invalid outside a narrow range of settings. They motivated the assumption in a binary alerting setting, where false alerts are assumed to not be harmful at the individual level, and $\hat{z} = 0$ does nothing to affect decision-making. We similarly present a falsification test for Assump. 3.2. This falsification test only covers the case when $\hat{Z} = z$; we provide an analogous test for the case when $\hat{Z} \neq z$ in the Appendix.

Proposition 3.2 (Falsification of Assump. 3.2)**.** *Let $\pi_+, \pi_- \in \mathcal{T}$ be two trialed models. Consider a subpopulation with $\mathcal{X}_{better} \subseteq \mathcal{X}$ such that for all $x \in \mathcal{X}_{better}$, $f_M(x, \pi_+) > f_M(x, \pi_-)$, and a subpopulation with covariates $\mathcal{X}_{agree} \subseteq \mathcal{X}$ such that for all $x \in \mathcal{X}_{agree}$, $\pi_+(x) = \pi_-(x)$. Then, for any $\mathcal{X}_s = \mathcal{X}_{better} \cap \mathcal{X}_{agree}$*

$$\mathbb{E}[Y|X \in \mathcal{X}_s, \Pi = \pi_+, Z = \pi_+(X)]$$
$$< \mathbb{E}[Y|X \in \mathcal{X}_s, \Pi = \pi_-, Z = \pi_+(X)]$$

falsifies Assump. 3.2.

Furthermore, we introduce the concept of a "neutral prediction", and assume that performance of the model does not affect the outcome in this case.

Assumption 3.3 (Neutral prediction)**.** There exists $\hat{z}_0 \in \mathcal{Z}$ such that for all $m_i, m_j \in \mathcal{M}$ and $z \in \mathcal{Z}$

$$Y(\hat{Z} = \hat{z}_0, M = m_i, Z = z) = Y(\hat{Z} = \hat{z}_0, M = m_j, Z = z)$$

We note that this assumption is "optional", in the sense that the method still produces bounds without it, at the cost of those bounds being potentially looser. This assumption is largely the same as Assumption 3.2 of Chen and Oberst [2025], and plays the role of linking the "control arm" of a trial to evaluation of a newer model, and can be seen as a special case of Assump. 3.2 in which the nonstrict inequality becomes an equality. In our running example, such a "prediction" could be choosing to defer to a radiologist, and the control arm could be viewed as an "always defer" policy. The intuition for this assumption is that when the model takes the neutral prediction, the physician behaves as if there were no model at all, and hence model performance does not impact actions. We also provide a falsification test for this assumption.

Proposition 3.3 (Falsification of Assump. 3.3)**.** *Let $\pi_1, \pi_2 \in \mathcal{T}$ be two trialed models. Letting $\mathcal{X}_{\hat{z}_0} := \{x \in \mathcal{X} | \pi_1(x) = \pi_2(x) = \hat{z}_0\}$, the observation that for some $z \in \mathcal{Z}$:*

$$\mathbb{E}\left[Y|X \in \mathcal{X}_{\hat{z}_0}, \Pi = \pi_1, Z = z\right]$$
$$\neq \mathbb{E}\left[Y|X \in \mathcal{X}_{\hat{z}_0}, \Pi = \pi_2, Z = z\right]$$

falsifies Assump. 3.3.

Finally, similar to Chen and Oberst [2025], we require that Y itself be bounded by known maximum and minimum values, which can be trivially falsified by observing values of Y outside of those bounds.

Assumption 3.4 (Bounded outcomes)**.** There exist $Y_{min}, Y_{max} \in \mathbb{R}$ such that $Y_{min} \leq Y \leq Y_{max}$

4 PARTIAL IDENTIFICATION BOUNDS

In this section, we introduce our main result, which provides upper and lower bounds on the expected outcome under deployment of a new model (Eq. (1)). Before we can introduce this result, we first introduce some notation for partitions of the model and covariate space in our RCT. These partitions will be used in the derivation of our bounds.

Definition 4.1 (Partitions)**.** We use π_i as a model in the model space and π_e as the untrialed model. For each value of x, we define the set of trialed models that make the same prediction as our untrialed model, $\pi_e(x)$[4].

$$\mathcal{S}(x) \coloneqq \{\pi_i \in \mathcal{T} \mid \pi_i(x) = \pi_e(x)\}$$

We further partition $\mathcal{S}(x)$ into models whose predictive performances imply outcomes that are no better / no worse than those of π_e

$$\begin{aligned}\mathcal{S}_{\downarrow}(x,z) \coloneqq &\{\pi_i \in \mathcal{S} \mid f_M(x,\pi_i) \le f_M(x,\pi_e), \pi_e(x) = z\} \\ &\cup \{\pi_i \in \mathcal{S} \mid f_M(x,\pi_i) \ge f_M(x,\pi_e), \pi_e(x) \ne z\} \\ \mathcal{S}_{\uparrow}(x,z) \coloneqq &\{\pi_i \in \mathcal{S} \mid f_M(x,\pi_i) \ge f_M(x,\pi_e), \pi_e(x) = z\} \\ &\cup \{\pi_i \in \mathcal{S} \mid f_M(x,\pi_i) \le f_M(x,\pi_e), \pi_e(x) \ne z\}\end{aligned}$$

We further define subsets of $\mathcal{S}_{\downarrow}, \mathcal{S}_{\uparrow}$ containing only the models whose performance is closest to that of π_e.

$$\begin{aligned}\mathcal{S}_{\Downarrow}(x,z) &\coloneqq \underset{\pi_i \in \mathcal{S}_{\downarrow}(x,z)}{\arg\min} |f_M(x,\pi_i) - f_M(x,\pi_e)| \\ \mathcal{S}_{\Uparrow}(x,z) &\coloneqq \underset{\pi_i \in \mathcal{S}_{\uparrow}(x,z)}{\arg\min} |f_M(x,\pi_i) - f_M(x,\pi_e)|\end{aligned}$$

Finally, for some value of x and an optimal prediction z, we define the set of all trialed models that give the correct prediction and its complement:

$$\mathcal{C}(x,z) \coloneqq \{\pi_i \in \mathcal{T} \mid \pi_i(x) = z\}$$

$$\overline{\mathcal{C}}(x,z) \coloneqq \{\pi_i \in \mathcal{T} \mid \pi_i(x) \ne z\}$$

We note that any of the defined sets can possibly be empty.

Example 2.1 (continuing from p. 3). Suppose in our radiology RCT, we have three models: π_1, π_2, π_3. The precision of these models are 0.1, 0.8, and 0.9, respectively, at a specific covariate value x. For a specific patient with covariates x, whose true diagnosis, z, is pneumonia, we want to evaluate an untrialed model, π_e, with precision 0.7. Suppose π_2, π_3 and π_e all suggest the diagnosis of pneumonia, while π_1 suggests bronchitis for this particular patient. Under this scenario, $\mathcal{S}(x) = \{\pi_2, \pi_3\}$, and π_e was optimal, and thus since both π_2 and π_3 have greater performance than π_e, $\mathcal{S}_{\uparrow}(x,z) = \{\pi_2, \pi_3\}$ and $\mathcal{S}_{\downarrow}(x,z) = \varnothing$. If π_e were not optimal the sets would be reversed. Further, $\mathcal{S}_{\Uparrow}(x,z) = \{\pi_2\}$, since its performance metric is closest to that of π_e. Finally, $\mathcal{C}(x,z) = \{\pi_2, \pi_3\}, \overline{\mathcal{C}}(x,z) = \{\pi_1\}$.

[4] As in Chen and Oberst [2025], these sets should be written with π_e as an argument, but we omit this for brevity. Note that $\mathcal{S}$ stands for "same" and $\mathcal{C}$ stands for "correct".

Next, under the notation established in Def. 4.1, we introduce a *branch selector* that assigns each (x,z) to the single assumption used to bound the outcome.

Definition 4.2 (Branch selector)**.** For each (x,z) we define lower- and upper-bound branch selectors $b_L, b_U : \mathcal{X} \times \mathcal{Z} \to \{\mathsf{neu}, \mathsf{mon}, \mathsf{cor}, \mathsf{def}\}$ (neutral, monotonic, correct, default), evaluated by first match:

$$b_L(x,z) = \begin{cases} \mathsf{neu} & \pi_e(x) = \hat{z}_0,\ \mathcal{S}(x) \ne \varnothing \\ \mathsf{mon} & \pi_e(x) \ne \hat{z}_0,\ \mathcal{S}_{\downarrow}(x,z) \ne \varnothing \\ \mathsf{cor} & \pi_e(x) = z,\ \mathcal{S}_{\downarrow}(x,z) = \varnothing,\ \overline{\mathcal{C}}(x,z) \ne \varnothing \\ \mathsf{def} & \text{otherwise,} \end{cases}$$

$$b_U(x,z) = \begin{cases} \mathsf{neu} & \pi_e(x) = \hat{z}_0,\ \mathcal{S}(x) \ne \varnothing \\ \mathsf{mon} & \pi_e(x) \ne \hat{z}_0,\ \mathcal{S}_{\uparrow}(x,z) \ne \varnothing \\ \mathsf{cor} & \pi_e(x) \ne z,\ \mathcal{S}_{\uparrow}(x,z) = \varnothing,\ \mathcal{C}(x,z) \ne \varnothing \\ \mathsf{def} & \text{otherwise.} \end{cases}$$

Because each selector is a function evaluated by first match, it returns exactly one value, so the induced indicators $\mathbf{1}\{b_L = k\}, \mathbf{1}\{b_U = k\}$ partition the space of $\mathcal{X} \times \mathcal{Z}$.

With this selector, we can now produce our main theorem, which provides bounds on the expected outcome of deploying an untrialed model in our novel RCT setting.

Theorem 4.1 (Bounds)**.** *Under Assumptions 2.1 and 3.1 to 3.4, we have the following bounds on the impact of our untrialed model π_e:*

$$L(\pi_e) \le \mathbb{E}[Y(\hat{Z} = \pi_e(X), M = f_M(X, \pi_e))] \le U(\pi_e)$$

Where:

$$\begin{aligned}L(\pi_e) = \mathbb{E}[&\mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}(X), Z] \qquad (2) \\ &+ \mathbf{1}\{b_L = \mathsf{mon}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}_{\Downarrow}(X,Z), Z] \\ &+ \mathbf{1}\{b_L = \mathsf{cor}\} \max_{i \in \overline{\mathcal{C}}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z] \\ &+ \mathbf{1}\{b_L = \mathsf{def}\}Y_{min}]. \qquad (3)\end{aligned}$$

$$\begin{aligned}U(\pi_e) = \mathbb{E}[&\mathbf{1}\{b_U = \mathsf{neu}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}(X), Z] \\ &+ \mathbf{1}\{b_U = \mathsf{mon}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}_{\Uparrow}(X,Z), Z] \\ &+ \mathbf{1}\{b_U = \mathsf{cor}\} \min_{i \in \mathcal{C}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z] \\ &+ \mathbf{1}\{b_U = \mathsf{def}\}Y_{max}]. \qquad (4)\end{aligned}$$

Note that each value of X, Z is covered by exactly one "branch", and the bound is constructed via pointwise upper

and lower bounds on the resulting outcome under π_e. We illustrate the intuition for the lower bound here, with symmetric logic for the upper bound: (i) When $b_L = \texttt{neu}$, our untrialed model takes the neutral prediction, and there are models in the RCT that agree. By Assump. 3.3, the downstream outcome will be equal regardless of performance across these neutral-prediction-taking models. (ii) When $b_L = \texttt{mon}$, our untrialed model does not take the neutral prediction, and there are models in the RCT that agree with it and have worse or equal performance. By Assump. 3.2, the downstream outcome of our untrialed model will be no worse than the best performing model in this set. (iii) When $b_L = \texttt{cor}$, our untrialed model takes the optimal prediction, and there are no models in the RCT that agree with it and have worse or equal performance. We also know that $b_L \neq \texttt{neu}$ and $\overline{\mathcal{C}}(x,z)$, or the set of incorrect models, is non-empty. By Assump. 3.1, the downstream outcome of our untrialed model will be no worse than any of these incorrect models, since they must have taken a suboptimal prediction. (iv) Finally, when $b_L = \texttt{def}$, none of the prior cases apply, leaving us with the maximally conservative choice of Y_{min}. We provide a flowchart for the lower bound in Fig. A.8, and note that the logic is similar (with slightly different relevant sets) for the upper bound.

Tightness of Bounds Chen and Oberst [2025] construct similar bounds under a different set of assumptions, and show that they are tight, in the sense that there exists a pair of SCMs that are allowable under their assumptions, and for which the upper and lower bounds match the causal effect. We do not claim that our bounds are tight, and provide intuition here as to why constructing tight bounds is not feasible under our counterfactual correctness assumption. Focusing on $L(\pi_e)$, Assump. 3.1 is stated pointwise on potential outcomes, while our bounds instead maximize over the conditional expectations. To illustrate, consider an RCT in which the untrialed model takes the optimal prediction. The data contains $\pi_1(x), \pi_2(x)$ that take incorrect predictions, while $\pi_e(x)$ takes the optimal prediction. Both trialed models have the same conditional expectation $E[Y|X,\Pi,Z] = 0.75$, so our lower-bound for $\pi_e(x)$ would be 0.75. A (likely tighter) lower bound would instead be $E[\max_{i=1,2} Y(\pi_i) \mid X, Z]$, but implementing such a bound is not feasible given that we only observe outcomes for a single trialed model for any individual.

5 ESTIMATION

The bounds in Theorem 4.1 have several values that must be estimated. For example, Eq. (2) requires $\mathbb{E}[Y \mid X, \Pi \in \mathcal{S}(X), Z]$ for all x such that $\mathbf{1}\{b_L = \texttt{neu}\} = 1$. If $\mathcal{S}(X)$ contains more than one model, it is impossible to have data on Y under both. We address this with consistent and asymptotically normal estimators, and asymptotically valid confidence intervals for the lower and upper bounds. Constructing these estimators requires addressing three challenges that do not arise in Chen and Oberst [2025]: (i) estimating $\mathbb{E}[Y \mid X, Z, \pi]$ via nuisance models, (ii) handling non-smooth max/min operators, and (iii) estimating model performance from finite data rather than assuming oracle knowledge. We address these challenges with an augmented inverse propensity weighted estimator for the bounds, and we prove its consistency and asymptotic normality.

Definition 5.1 (Nuisance Functions)**.** We define the following nuisance functions and estimators:

$$\mu_j(x,z) := \mathbb{E}[Y|x, \Pi = \pi_j, z] \tag{5}$$

$$\begin{aligned} \hat{d}_L(x,z) &:= \arg\max_{j \in \overline{\mathcal{C}}(x,z)} \widehat{\mu}_j(x,z) \\ \hat{d}_U(x,z) &:= \arg\min_{j \in \mathcal{C}(x,z)} \widehat{\mu}_j(x,z) \end{aligned} \tag{6}$$

$$\widehat{\lambda}_j(Y,\Pi,x,z) := \frac{\mathbf{1}\{\Pi = j\}}{P(\Pi = j \mid X = x)}\big(Y - \widehat{\mu}_j(x,z)\big) \tag{7}$$

$$\begin{aligned} \widehat{\mathcal{S}_{\Downarrow}}(x,z) &:= \arg\min_{\pi \in \widehat{\mathcal{S}}_{\downarrow}(x,z)} |\hat{f}_M(x,\pi) - \hat{f}_M(x,\pi_e)| \\ \widehat{\mathcal{S}}_{\Uparrow}(x,z) &:= \arg\min_{\pi \in \widehat{\mathcal{S}}_{\uparrow}(x,z)} |\hat{f}_M(x,\pi) - \hat{f}_M(x,\pi_e)| \end{aligned} \tag{8}$$

$$\begin{aligned} \varphi_L\big(x,z;\hat{d}_L\big) &:= \widehat{\mu}_{\hat{d}_L(x,z)}(x,z) + \widehat{\lambda}_{\hat{d}_L(x,z)}(Y,\Pi,x,z) \\ \varphi_U\big(x,z;\hat{d}_U\big) &:= \widehat{\mu}_{\hat{d}_U(x,z)}(x,z) + \widehat{\lambda}_{\hat{d}_U(x,z)}(Y,\Pi,x,z), \end{aligned} \tag{9}$$

Where $\widehat{\mathcal{S}}_{\downarrow}(x,z), \widehat{\mathcal{S}}_{\uparrow}(x,z)$ are defined as in Def. 4.1, with $\hat{f}$ in place of f, and $\widehat{\mu}(x,z)$ is the plug-in estimator of μ.

For our estimators to be consistent and asymptotically normal, we require the following assumptions:

Assumption 5.1 (Known Propensities)**.** The value of $P(\Pi \in S \mid X)$ is known for all $S \subseteq \mathcal{T}$, and is strictly positive for all S, X.

Given our RCT setting, this assumption will be satisfied by design. A similar assumption is implicit in Chen and Oberst [2025], but we make it explicit here. We require additional conditions to handle non-smooth max/min operators, which we describe next.

Assumption 5.2 (Margin Conditions)**.** Under Def. 5.1, there exists $\alpha > 0$ such that for any $t \geq 0$, we have

$$P\left[\min_{j \neq d(X,Z)} \big|\mu_{d(X,Z)}(X,Z) - \mu_j(X,Z)\big| \leq t\right] \leq t^{\alpha}$$

for both $d_L(x,z), d_U(x,z)$, as defined in Eq. (6), and μ defined in Eq. (5). Furthermore, there exists $\beta > 0$ such that for any $t \geq 0$, we have

$$P\left[\min_{i \neq j} |f_M(X,\pi_j) - f_M(X,\pi_i)| \leq t\right] \leq t^{\beta}$$

where f_M is defined in Assump. 2.1, and $\pi_i, \pi_j \in \mathcal{T} \cup \{\pi_e\}$.

These assumptions state that there is sufficient separation between the downstream outcomes of the best two models, the worst two models, and between the performance of any two models, untrialed or trialed.

Assumption 5.3 (Convergence of Nuisance Functions)**.** The nuisance functions $\widehat{\mu}_j(X,Z)$, $\widehat{f}_M(x;\pi_i)$, defined in Def. 5.1 converge at rates $o_p(n^{-\frac{1}{2(1+\alpha)}})$ and $o_p(n^{-\frac{1}{2\beta}})$, respectively, for all $\pi_i \in \mathcal{T}, x \in \mathcal{X}$, where α and β are the same constants as in the Margin Conditions (Assump. 5.2). That is:

$$\|\widehat{\mu}_j(X,Z) - \mu_j(X,Z)\| = o_p(n^{-\frac{1}{2(1+\alpha)}})$$

$$\|\widehat{f}_M(x;\pi_i) - f_M(x,\pi_i)\| = o_p(n^{-\frac{1}{2\beta}})$$

where $\pi_i \in \mathcal{T} \cup \{\pi_e\}$

We note that $o_p(n^{-1/4})$ convergence rates have been demonstrated for a variety of modern machine learning methods under e.g., sparsity constraints [Chernozhukov et al., 2018], making Assump. 5.3 plausible for moderate values of α, β (e.g., $\alpha = 1, \beta = 2$). Finally, we can define our estimator of the upper and lower bounds as the average of the pseudo-outcomes defined below.

Definition 5.2 (Pseudo-Outcomes)**.** We first define

$$\hat{h}(X,Y,Z,S) := \widehat{\mu}_S(X,Z) + \frac{\mathbf{1}\{\Pi \in S\}}{P(\Pi \in S \mid X)}\big(Y - \widehat{\mu}_S(X,Z)\big)$$

where $S \subseteq \mathcal{T}$ is an arbitrary set of models. Then, we define pseudo-outcomes for the lower and upper bounds as

$$\begin{aligned}
&\widehat{\psi}_L(Y,X,\Pi,Z;\widehat{f}_M) = \\
&\mathbf{1}\{\hat{b}_L = \mathsf{neu}\}\,\hat{h}(X,Y,Z,\mathcal{S}) + \mathbf{1}\{\hat{b}_L = \mathsf{mon}\}\,\hat{h}(X,Y,Z,\widehat{\mathcal{S}}_{\Downarrow}) \\
&+ \mathbf{1}\{\hat{b}_L = \mathsf{cor}\}\,\varphi_L(X,Z;\hat{d}_L) + \mathbf{1}\{\hat{b}_L = \mathsf{def}\}\,Y_{\min} \\
&\widehat{\psi}_U(Y,X,\Pi,Z;\widehat{f}_M) = \\
&\mathbf{1}\{\hat{b}_U = \mathsf{neu}\}\,\hat{h}(X,Y,Z,\mathcal{S}) + \mathbf{1}\{\hat{b}_U = \mathsf{mon}\}\,\hat{h}(X,Y,Z,\widehat{\mathcal{S}}_{\Uparrow}) \\
&+ \mathbf{1}\{\hat{b}_U = \mathsf{cor}\}\,\varphi_U(X,Z;\hat{d}_U) + \mathbf{1}\{\hat{b}_U = \mathsf{def}\}\,Y_{\max}
\end{aligned}$$

where the argument after the semicolon denotes the performance function used to construct the partitions and branch selectors: $\hat{b}_L, \hat{b}_U$ and the estimated sets are computed from the estimated performance $\hat{f}_M$ in place of f_M.

Theorem 5.1 (Consistency and asymptotic normality of $\widehat{L}(\pi_e), \widehat{U}(\pi_e)$)**.** *If Assumptions 5.1 to 5.3 hold for some $\alpha, \beta > 0$, then*

$$\widehat{L}(\pi_e) = n^{-1}\sum_{i=1}^{n} \widehat{\psi}_L(Y_i,X_i,\Pi_i,Z_i;\widehat{f}_M) \xrightarrow{P} L(\pi_e)$$

$$\widehat{U}(\pi_e) = n^{-1}\sum_{i=1}^{n} \widehat{\psi}_U(Y_i,X_i,\Pi_i,Z_i;\widehat{f}_M) \xrightarrow{P} U(\pi_e)$$

Where $\xrightarrow{P}$ denotes convergence in probability. Furthermore,

$$\sqrt{n}(\widehat{L}(\pi_e) - L(\pi_e)) \xrightarrow{d} N(0,\sigma_L^2)$$

$$\sqrt{n}(\widehat{U}(\pi_e) - U(\pi_e)) \xrightarrow{d} N(0,\sigma_U^2)$$

Where $\xrightarrow{d}$ denotes convergence in distribution, and $\sigma_L^2 = Var(\psi_L), \sigma_U^2 = Var(\psi_U)$, with ψ_L, ψ_U denoting the pseudo-outcomes of Def. 5.2 evaluated with the true nuisance functions μ_j, f_M in place of their estimates.

Since our estimator is asymptotically normal, this provides us the following corollary:

Corollary 5.1 (Confidence Intervals for $L(\pi_e), U(\pi_e)$)**.** *Under the conditions of Theorem 5.1, we can construct the following confidence intervals for $L(\pi_e), U(\pi_e)$:*

$$\widehat{L}(\pi_e) \pm z_{1-\gamma/2}\sqrt{\widehat{Var}(\widehat{\psi}_L)/n}$$

$$\widehat{U}(\pi_e) \pm z_{1-\gamma/2}\sqrt{\widehat{Var}(\widehat{\psi}_U)/n}$$

Where $z_{1-\gamma/2}$ is the $1-\gamma/2$ quantile of the standard normal distribution, and $\widehat{Var}(\widehat{\psi}_L), \widehat{Var}(\widehat{\psi}_U)$ are the empirical variance of $\widehat{\psi}_L, \widehat{\psi}_U$, respectively.

We note that our results require Assump. 5.2 to hold for some value of α, β. The first (α) condition controls the bias from selecting the wrong argmax/argmin in the $\mathbf{1}\{b_L = \mathsf{cor}\}, \mathbf{1}\{b_U = \mathsf{cor}\}$ branches; the second (β) condition controls a separate source of error specific to our setting, namely that $\hat{\psi}_L, \hat{\psi}_U$ are evaluated at the estimated partitions $\widehat{\mathcal{S}}_{\Downarrow}, \widehat{\mathcal{S}}_{\Uparrow}$ rather than at the population partitions in Theorem 4.1, and $\hat{f}_M$'s error can permute the relevant ordering. The probability of a permutation is bounded polynomially by the margin condition, contributing $o_p(n^{-1/2})$ to the overall error.

6 EXPERIMENTS

We now shift discussion to a semi-synthetic simulation study to demonstrate the efficiency of our proposed bounds. All experiment code is available on Github at oberst-lab/counterfactual-correctness-experiments.

We will use data from Imai et al. [2023], an experiment evaluating the impact of showing judges algorithmically generated risk scores (1-6) prior to bail sentencing. This data also records the action taken by judges, which includes assigning high bail, low bail, or release without bail to suspects (note that "imprisonment without bail" does not appear as an option in the data). Downstream outcomes were recorded for each suspect, such as failure to appear (FTA) at a trial hearing. The full data includes suspect covariates, judge group, judge decision, risk score, and FTA for every individual. FTA distributions in the data are shown in Table A.6.

Simulation Setup We use this data to construct a semi-synthetic simulation study, re-using the covariates of individuals in the dataset, as well as the original risk scores. We introduce simulation in order to reason about the 'ground truth' counterfactual outcomes under different forms of AI assistance. We focus primarily on hypothetical alerting models that notify the judge of high-risk individuals: This framing allows us to consider the "control arm" of the trial as providing information on judge behavior in the absence of an alert. We make the additional simplification, for convenience, of considering "requires bail" $(A = 1)$ and "does not require bail" as the only actions, collapsing the "high" and "low" bail options. Using the original data, we learn

$$\hat{f}_a(X) := \hat{P}(A = 1 \mid X, \text{Arm} = \text{control})$$
$$\hat{f}_y(X, A) := \hat{P}(\text{FTA} = 1 \mid X, A, \text{Arm} = \text{control})$$

where both $\hat{f}_a$ and $\hat{f}_y$ represent our estimates of the conditional probability of (a) the judge requiring some form of bail in the absence of ML assistance, and (b) the defendant failing to appear, conditioned on the type of bail required. We also retain the risk score (from 1-6) that is provided in the original dataset, which we denote $\mathfrak{R}$, and which is used to inform some of the models that we evaluate later on.

We then simulate judge and defendant behavior as follows: If no alert is raised ($\hat{Z} = 0$), then judges make decisions according to $\hat{f}_a(X)$, and if an alert is raised ($\hat{Z} = 1$), then their probability increases proportionally to the performance (in this case, the precision) of the model as follows

$$P(A = 1 \mid X, \Pi, \hat{Z})$$
$$= \sigma(\text{logit}(\hat{f}_a(X)) + \mathbf{1}\{\hat{Z} = 1\} \cdot f_M(X, \Pi))$$

where σ is the sigmoid function, and $f_M(X, \Pi)$ is the performance of the model Π at X. We then simulate failure-to-appear using $P(\text{FTA} = 1 \mid X, A) = \hat{f}_y(X, A)$, and construct our outcome Y as a net utility that rewards court appearance while imposing a penalty for requiring bail

$$Y_{\text{raw}} = 1 - c_{\text{FTA}} \cdot \text{FTA} - c_{\text{det}} \cdot A, \quad Y = \frac{Y_{\text{raw}} + c_{\text{det}}}{1 + c_{\text{det}}},$$

with $c_{\text{FTA}} = 1$ and $c_{\text{det}} = 0.1$. The affine rescaling is for interpretability, as Y_{raw} is already bounded by $[-0.1, 1]$. The bail penalty c_{det} reflects our intuition that over-detention is harmful and allows incorrect predictions, particularly false positives, to cause harm.

Finally, we take Z (the 'ground truth' label) to be whether or not the defendant would fail to appear if released without bail. While in reality, Z would only be observed for individuals released without bail, for the purpose of the illustrative simulation here, we take it to be observed for all individuals.[5] More broadly, while we use real data to inform our simulation of judge and defendant behavior, we do not claim that our simulation is a realistic surrogate for real-world behavior: Rather, we design our simulation primarily for ease of exposition, and to illustrate the application of our method. We now briefly describe the experiments run, with additional details and figures in Appendix E.

[5] In Appendix G we discuss assumptions under which our method can handle missingness, and in Exp. 5 we consider a misspecified experiment with missing values of Z.

Exp #	Width (Ours)	Width (Chen)	Lower Bd. (Ours)	Upper Bd. (Ours)	True value
1	0.331	0.704	0.430	0.761	0.726
2	0.121	0.362	0.616	0.737	0.728
3	0.140	0.281	0.624	0.764	0.696
4	0.454	0.802	0.349	0.803	0.695
5	0.204	0.281	0.586	0.790	0.697

Table 1: Semi-synthetic experiment results (see Section 6) showing the mean width of our bounds $\hat{U}(\pi_e) - \hat{L}(\pi_e)$ compared to those derived via the method of Chen and Oberst [2025], alongside the average upper and lower bound values and the true value. We see that our bounds are consistently tighter and always contain the true value of Y.

Evaluation of changing alert thresholds and re-training of a pre-existing model (Exp 1, 2)

To start, we consider a simulated setting of an RCT that has two arms: a control arm in which no alerts are raised, and a trial arm in which alerts are raised when the risk score $\mathfrak{R}$ present in the Imai et al. [2023] dataset is greater than 3. Judge decisions and outcomes are then simulated as described above. In Experiment 1, we evaluate a new alerting system that uses the same risk scores $\mathfrak{R}$, but raises an alert at a lower (less conservative) threshold. In particular, alerts are raised for every individual with a risk score $\mathfrak{R}$ greater than 1. In Experiment 2, we create a new risk score algorithm and evaluate an untrialed model that raises an alert when this new risk score is greater than 3. This algorithm is discussed in greater detail in Appendix E.1. This setting represents the motivating scenario where a freshly trained model needs to be evaluated.

Overall results are shown in Table 1, where we observe that the bounds produced by Chen and Oberst [2025] are extremely wide (0.70 and 0.36 for Exp. 1 and 2, respectively) given that the outcome (and hence the target to estimate) is bounded between zero and one. This gap reflects the fact that in both cases, the untrialed model tends to produce alerts in cases where no alerts occurred in the original trial (trialed model raises alerts on 28.1% of individuals, whereas the untrialed model raises an alert on 80.2% and 45.6% of individuals in Exp. 1 and 2, respectively). Our method produces bounds that are consistently narrower than those of prior work (0.33 and 0.12, respectively), because we are able to produce tighter lower bounds in settings where these previously unobserved alerts are correct, and tighter upper bounds when they are incorrect.

In the Appendix, we conduct a further ablation study, using data from Experiment 1, finding that the primary driver of improvement over Chen and Oberst [2025] is the use of the counterfactual correctness bounds, rather than e.g., stratification of M on X. In Table A.4, we also show the results of 100 Monte Carlo trials of Exp. 1, demonstrating that all methods consistently capture the true expected outcome, but our method produces a smaller bound interval and narrower confidence intervals.

Evaluation in the absence of an RCT (Exp 3, 4)

When no RCT has been conducted, but data exists from prior to the deployment of any AI system, such data is effectively a "single-arm" RCT that only contains a control arm that always takes the "neutral" action. We illustrate the application of our method in this setting. Here, the prediction of "no alert" is taken to be a neutral action (leading to similar outcomes as those in prior data), while the prediction of "alert", when correct/incorrect, implies that outcomes will be no worse/no better than those observed in prior data. In Experiment 3, we evaluate an untrialed model that raises an alert for individuals with a risk score greater than 3, and in Experiment 4, we evaluate an untrialed model that does the same for those with a risk score greater than 1.

Overall results are similarly shown in Table 1, where we again find our bounds are consistently tighter than Chen and Oberst [2025]. We also notice that bound width grows with the alert rate of the untrialed model. Experiment 3's conservative threshold ($\mathfrak{R} > 3$) raises few alerts and leads to a narrow bound (0.140), while Exp. 4's aggressive threshold ($\mathfrak{R} > 1$) raises many more alerts and leads to a much wider bound (0.454). Additionally, in Fig. 2, we illustrate our bounds across different age groups, compared to the method from Chen and Oberst [2025], after modifying the latter method to use per-age group performances. We see that across all groups, our bounds capture the true value and are tighter, agreeing with Table 1. This result suggests that our notion of counterfactual correctness is the primary driver of empirical improvement, rather than stratifying M by X.

Misspecified example with missing data (Exp 5)

As a sanity check, we consider the setting of Experiment 3, but where we hide the counterfactual failure-to-appear values for offenders that were not released without bail. This modification violates our assumptions, and this experiment is designed to illustrate the impact of missing values of Z, as discussed further in Appendix G.

In Table 1, we observe that our bounds are wider (0.2 in Exp. 5 vs 0.14 in Exp. 3), but still contain the true value. We show further details in Table A.2 in the Appendix, where we demonstrate that the bounds still capture the true values, and the confidence intervals for the lower and upper bounds still capture the ground truth values. This result suggests that the bias introduced by the misspecification in this particular simulation is not significant enough to yield invalid results from applying our method.

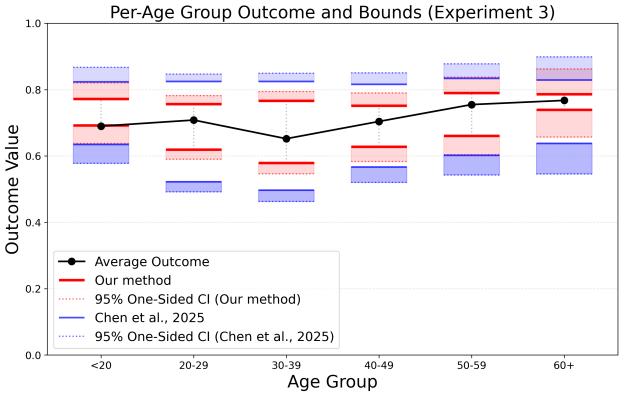


Figure 2: (Experiment 3) Comparison of bounds on downstream outcomes for an untrialed model with a single trialed model in the RCT. The figure displays the lower bound and upper bound computed using our proposed method and a modified version of the method from Chen and Oberst [2025], also plotted in Fig. A.1b with performance as functions of individual age.

7 CONCLUSION AND LIMITATIONS

When a newly trained model has better predictive performance than models previously evaluated in an RCT, there will naturally be limited data on how its predictions impact outcomes, insofar as it makes correct predictions where prior models failed. Our work improves over prior work in partial identification of the causal impact of AI systems, by recognizing that, when ground truth labels exist in a prior RCT, the model's predictions can be evaluated against that ground truth, and correctness used as a directional indicator of relative impact at the individual level.

Several limitations and areas for future investigation merit discussion. Since our falsification tests are in-expectation, individual-level assumption violations may go undetected. Handling cases where an identical prediction from a better model has a worse outcome, or where neutral action depends on performance, is an important direction for future work. Many deployed models also do not provide discrete decision recommendations. Instead, many may output continuous predictions. Our method can still be applied by imposing discrete buckets, or using $\mathbf{1}\{b_L = \mathsf{cor}\}, \mathbf{1}\{b_U = \mathsf{cor}\}, \mathbf{1}\{b_L = \mathsf{def}\}, \mathbf{1}\{b_U = \mathsf{def}\}$ branches only, but more efficient methods of evaluation may be of interest. As discussed, our method has limitations when values of Z are missing, which presents avenues for future research. Finally, investigating an approach that incorporates observational data, which is often significantly more available than RCT data, would be valuable.

References

Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume 1, pages 230–237. AAAI Press/MIT Press, 1994.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 113–120, Catalina Island, CA, 2012. AUAI Press.

Alexis Bellot. Towards bounding causal effects under Markov equivalence. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 308–332. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/bellot24a.html.

Alexis Bellot and Silvia Chiappa. Towards estimating bounds on the effect of policies under unobserved confounding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 104556–104594. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bd0194a23b3b0787e54a0173d65b0b37-Paper-Conference.pdf.

Eli Ben-Michael, D James Greiner, Melody Huang, Kosuke Imai, Zhichao Jiang, and Sooahn Shin. Does AI help humans make better decisions? a statistical evaluation framework for experimental and observational studies. *Proceedings of the National Academy of Sciences*, 122 (38):e2505106122, 2025.

Yewon Byun, Dylan Sam, Michael Oberst, Zachary Lipton, and Bryan Wilder. Auditing fairness under unobserved confounding. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4339–4347. PMLR, 5 2024. URL https://proceedings.mlr.press/v238/byun24a.html.

Jacob M. Chen and Michael Oberst. Just trial once: Ongoing causal validation of machine learning models. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, 2 2025. URL https://proceedings.mlr.press/v286/chen25b.html.

Yao Chen, D. Ohlssen, A. Readie, G. Ligozio, Ruvie Martin, and T. Coroller. The framework that survives bad models: Human-AI collaboration for clinical trials. *ArXiv*, abs/2510.06567, 2025. URL https://api.semanticscholar.org/CorpusId:281891637.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 02 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

David Maxwell Chickering and Judea Pearl. A clinician's tool for analyzing non-compliance. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1269–1276. AAAI, 1996.

Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020a.

Juan D. Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 10093–10100. AAAI Press, 2020b.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793, 2024.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Shruti K. Gohil, Edward Septimus, Ken Kleinman, Neha Varma, Taliser R. Avery, Lauren Heim, Risa Rahm, William S. Cooper, Mandelin Cooper, Laura E. McLean, Naoise G. Nickolay, Robert A. Weinstein, L. Hayley Burgess, Micaela H. Coady, Edward Rosen, Selsebil Sljivo, Kenneth E. Sands, Julia Moody, Justin Vigeant, Syma Rashid, Rebecca F. Gilbert, Kim N. Smith, Brandon Carver, Russell E. Poland, Jason Hickok, S. G. Sturdevant, Michael S. Calderwood, Anastasiia Weiland, David W. Kubiak, Sujan Reddy, Melinda M. Neuhauser, Arjun Srinivasan, John A. Jernigan, Mary K. Hayden, Abinav Gowda, Katyuska Eibensteiner, Robert Wolf, Jonathan B.

Perlin, Richard Platt, and Susan S. Huang. Stewardship prompts to improve antibiotic selection for urinary tract infection. *JAMA*, 331(23):2018, 6 2024. doi: 10.1001/jama.2024.6259. URL http://dx.doi.org/10.1001/jama.2024.6259.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349. IEEE, 2021.

Kosuke Imai, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2):167–189, 2023.

Shalmali Joshi, Iñigo Urteaga, Wouter A C van Amsterdam, George Hripcsak, Pierre Elias, Benjamin Recht, Noémie Elhadad, James Fackler, Mark P Sendak, Jenna Wiens, Kaivalya Deshpande, Yoav Wald, Madalina Fiterau, Zachary Lipton, Daniel Malinsky, Madhur Nayan, Hongseok Namkoong, Soojin Park, Julia E Vogt, and Rajesh Ranganath. AI as an intervention: improving clinical outcomes relies on a causal approach to AI development and validation. *Journal of the American Medical Informatics Association*, 1 2025. doi: 10.1093/jamia/ocae301. URL https://doi.org/10.1093/jamia/ocae301.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Apoorva Lal, Wenjing Zheng, and Simon Ejdemyr. A framework for generalization and transportation of causal estimates under covariate shift. *arXiv preprint arXiv:2301.04776*, 2023.

Sanghack Lee, Juan D. Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pages 389–398. PMLR, 2019.

Lydia T. Liu, Inioluwa Deborah Raji, Angela Zhou, Luke Guerdan, Jessica Hullman, Daniel Malinsky, Bryan Wilder, Simone Zhang, Hammaad Adam, Amanda Coston, Ben Laufer, Ezinne Nwankwo, Michael Zanger-Tishler, Eli Ben-Michael, Solon Barocas, Avi Feller, Marissa Gerchick, Talia Gillis, Shion Guha, Daniel Ho, Lily Hu, Kosuke Imai, Sayash Kapoor, Joshua Loftus, Razieh Nabi, Arvind Narayanan, Ben Recht, Juan Carlos Perdomo, Matthew Salganik, Mark Sendak, Alexander Tolbert, Berk Ustun, Suresh Venkatasubramanian, Angelina Wang, and Ashia Wilson. Bridging prediction and intervention problems in social systems. *arXiv preprint (2507.05216v3)*, 7 2025. URL http://arxiv.org/abs/2507.05216v3.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2):319–323, 1990. Papers and Proceedings.

Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7521–7531. IEEE, 2022.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*, pages 10520–10545. PMLR, 2023.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33: 18819–18831, 2020.

Michael Nercessian, Wenxin Zhang, Alexander Schubert, Daphne Yang, Maggie Chung, Ahmed Alaa, and Adam Yala. Data reuse enables cost-efficient randomized trials of medical AI models. *arXiv preprint arXiv:2511.08986*, 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Generalizing experimental findings. *Journal of causal inference*, 3(2):259–266, 2015.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 247–254, 2011.

Inioluwa Deborah Raji and Lydia T. Liu. Evaluating prediction-based interventions with human decision makers in mind. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1180–1188. PMLR, 5 2025. URL https://proceedings.mlr.press/v258/raji25a.html.

Amy Richardson, Michael G. Hudgens, Peter B. Gilbert, and Jason P. Fine. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical Science*, 29(4), 11 2014. ISSN 0883-4237. doi: 10.1214/14-sts499. URL http://dx.doi.org/10.1214/14-STS499.

Thomas S Richardson and James M Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

James M. Robins. The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley, editors, *Health Service Research Methodology: A Focus on AIDS*, pages 113–159. U.S. Public Health Service, National Center for Health Services Research, Washington, DC, 1989.

Jonas Rothfuss, Bhavya Sukhija, Tobias Birchler, Parnian Kassraie, and Andreas Krause. Hallucinated adversarial control for conservative offline policy evaluation. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1774–1784. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/rothfuss23a.html.

Mostafa Shaban, Yasmine M Osman, Nermen Abdelfatah Mohamed, and Marwa Mamdouh Shaban. Empowering breast cancer clients through AI chatbots: transforming knowledge and attitudes for enhanced nursing care. *BMC nursing*, 24(1):994, 2025.

Aasma Shaukat, David R. Lichtenstein, Samuel C. Somers, Daniel C. Chung, David G. Perdue, Murali Gopal, Daniel R. Colucci, Sloane A. Phillips, Nicholas A. Marka, Timothy R. Church, William R. Brugge, Robert Thompson, Robert Chehade, Burr Loew, Jackie Downing, James Vermillion, Lawrence Borges, Ruma Rajbhandari, Theodore Schafer, Sahin Coban, James Richter, Peter Carolan, Francis Colizzo, Tiffany Jeong, Marisa DelSignore, Shreya Asher, Robert McCabe, Daniel Van Handel, Birtukan Cinnor, Benjamin Mitlyng, Cynthia Sherman, S. David Feldshon, Amy Lounsbury, Ana Thompson, Anusha Duggirala, Irena Davies, Christopher Huang, Charles Bliss, Arpan Mohanty, Oltion Sina, Jean Mendez, Allison Iwan, Jennifer Stromberg, Jonathan Ng, Lavi Erisson, Polina Golland, Daniel Wang, Evan Wlodkowski, Joseph Carlin, Perikumar Javia, Neelima Chavali, Austin Wang, Janine Little, and Cara Hunsberger. Computer-aided detection improves adenomas per colonoscopy for screening and surveillance colonoscopy: A randomized trial. *Gastroenterology*, 163(3):732–741, 9 2022. doi: 10.1053/j.gastro.2022.05.028. URL http://dx.doi.org/10.1053/j.gastro.2022.05.028.

Xu Shi, Ziyang Pan, and Wang Miao. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386, 2011.

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9196–9205. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/su20d.html.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Guido Vittorio Travaini, Federico Pacchioni, Silvia Bellumore, Marta Bosia, and Francesco De Micco. Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *International journal of environmental research and public health*, 19(17):10594, 2022.

Ross Upton, Ashley P. Akerman, Thomas H. Marwick, Casey L. Johnson, Hania Piotrowska, Mamta Bajre, Maria Breen, Helen Dawes, Hakim-Moulay Dehbi, Tine Descamps, Victoria Harris, Will Hawkes, Samuel Krasner, Emily Sanderson, Natalie Savage, Ben Thompson, Victoria Williamson, William Woodward, Rizwan Sarwar, Jamie O'Driscoll, Rajan Sharma, Virginia Chiocchia, Steffen E. Petersen, Elena Frangou, Ged Ridgway, Sanjeev Bhattacharyya, David P. Ripley, Gary Woodward,

and Paul Leeson. PROTEUS: A prospective RCT evaluating use of AI in stress echocardiography. *NEJM AI*, 1(11), 10 2024. doi: 10.1056/aioa2400865. URL http://dx.doi.org/10.1056/aioa2400865.

Wouter A. C. van Amsterdam, Michael Oberst, Jean Feng, Jenna Wiens, Shengpu Tang, Shalmali Joshi, Rajesh Ranganath, Mark Sendak, Uri Shalit, Julia E. Vogt, Brett Beaulieu-Jones, Muhammad Mamdani, David Kent, Patrick J. Heagerty, Thomas R. Fleming, and Anna Goldenberg. Clinical trials for continuously monitored and updated AI systems. *Nature Medicine*, 32(6):1972–1974, 6 2026. doi: 10.1038/s41591-026-04368-9. URL https://doi.org/10.1038/s41591-026-04368-9.

Wouter AC Van Amsterdam, Nan Van Geloven, Jesse H Krijthe, Rajesh Ranganath, and Giovanni Ciná. When accurate prediction models yield harmful self-fulfilling prophecies. *Patterns*, 6(4):101229, 2025.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pages 22184–22202. PMLR, 2022.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Dl4LetuLdyK.

Runzhe Wu, Masatoshi Uehara, and Wen Sun. Distributional offline policy evaluation with predictive error guarantees. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37685–37712. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wu23s.html.

Xiaoxi Yao, David R. Rushlow, Jonathan W. Inselman, Rozalina G. McCoy, Thomas D. Thacher, Emma M. Behnken, Matthew E. Bernard, Steven L. Rosas, Abdulla Akfaly, Artika Misra, Paul E. Molling, Joseph S. Krien, Randy M. Foss, Barbara A. Barry, Konstantinos C. Siontis, Suraj Kapa, Patricia A. Pellikka, Francisco Lopez-Jimenez, Zachi I. Attia, Nilay D. Shah, Paul A. Friedman, and Peter A. Noseworthy. Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nature Medicine*, 27(5):815–819, 5 2021. doi: 10.1038/s41591-021-01335-4. URL http://dx.doi.org/10.1038/s41591-021-01335-4.

Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12207–12215, 2021a.

Junzhe Zhang and Elias Bareinboim. Non-parametric methods for partial identification of causal effects. *Columbia CausalAI Laboratory Technical Report*, 2021b.

Bounding the Causal Impact of ML-assisted Decision-Making via Counterfactual Correctness (Supplementary Material)

Jonathan Zhang[1] Erik Skalnes[1] Jacob M. Chen[1] Michael Oberst[1]

[1]Computer Science Dept., Johns Hopkins University, Baltimore, Maryland, USA

A APPENDIX OVERVIEW

The appendix contains the following sections:

- Appendix B: Proofs and Additional Materials.
  This section provides proofs for falsification tests for the core assumptions introduced in the main text. If these tests fail, then the assumptions do not hold in the data, and any results may not be valid.
- Appendix C: Proof of Theorem 4.1.
  This section provides a formal proof of the core theorem of the paper. To prove the validity of the bounds, we write the analytical form of the bounds and demonstrate the expected value of the untrialed model falls within the bounds.
- Appendix D: Proof of Theorem 5.1.
  This section provides a formal proof that the empirical estimator for our bounds is consistent and asymptotically normal, enabling us to construct confidence intervals for our bounds. Since some of the branches of the bounds contain discontinuous "max/min" operators, the techniques of Chen and Oberst [2025] cannot be directly applied to the entire estimator. We instead formulate an AIPW estimator for our bounds. By Theorem 2 of Byun et al. [2024], we can demonstrate that such an estimator is consistent and asymptotically normal if seven conditions are met. For each assumption, we verify the conditions are met and directly apply the theorem.
- Appendix E: Further Simulation Details.
  This section provides some brief additional information about the simulation study.
- Appendix F: Additional Related Work. This section provides an overview of works in related areas that are relevant to our work, but were not discussed in the main text.
- Appendix G: Missing Data.
  This section discusses the handling of cases in which Z is missing from our data.
- Appendix H: Flowchart.
  This section provides a flowchart visualizing the branch-selection logic (b_L) used to construct the bounds.

B PROOFS AND ADDITIONAL MATERIALS

The following assumptions are used in the proofs of the propositions and theorems in this paper.

Corollary B.1 (Consistency)**.** *Under Assump. 2.1, if* $\hat{Z} = \hat{z},\ M = m,\ Z = z$ *then*

$$Y = Y(\hat{Z} = \hat{z}, M = m, Z = z).$$

Corollary B.2 (Conditional Ignorability)**.** *Under Assump. 2.1, since* $Y(\hat{z}, m) \perp\!\!\!\perp (\hat{Z}, M) \mid (X, Z)$*, we have*

$$P\left(Y(\hat{z}, m) \mid X, Z\right) = P\left(Y \mid \hat{Z} = \hat{z}, M = m, X, Z\right). \tag{10}$$

B.1 FALSIFICATION OF ASSUMPTION 3.1

Assumption 3.1 (Counterfactual Correctness)**.** We assume that correct predictions cause no harm. That is, for all $m, m' \in \mathcal{M}$ and $z, z' \in \mathcal{Z}$ with $z' \neq z$,

$$Y(\hat{Z} = z, M = m, Z = z) \geq Y(\hat{Z} = z', M = m', Z = z)$$

Proposition 3.1 (Falsification of Assump. 3.1)**.** *Let* $\pi_c, \pi_w \in \mathcal{T}$ *be two trialed models (one of which will be "correct" and the other "wrong"), and let* $\mathcal{X}_{c \neq w} \coloneqq \{x \in \mathcal{X} \mid \pi_c(x) \neq \pi_w(x)\}$ *be the set of covariates on which their predictions disagree. Consider the subpopulation where* $X \in \mathcal{X}_{c \neq w}$ *and* $Z = \pi_c(X)$ *(i.e., where* π_c *is correct and* π_w *incorrect). Then, the observation that outcomes are better under* π_w*,*[1]

$$\begin{aligned}&\mathbb{E}[Y \mid X \in \mathcal{X}_{c \neq w},\, \Pi = \pi_c,\, Z = \pi_c(X)] \\ &< \mathbb{E}[Y \mid X \in \mathcal{X}_{c \neq w},\, \Pi = \pi_w,\, Z = \pi_c(X)]\end{aligned}$$

falsifies Assump. 3.1.

Proof. Suppose Assump. 2.1 holds. Choose $x \in \mathcal{X}_{c \neq w}$. Then we have the following:

$$\begin{aligned}&\mathbb{E}\left[Y \mid X = x, \Pi = \pi_c, Z = \pi_c(x)\right] - \mathbb{E}\left[Y \mid X = x, \Pi = \pi_w, Z = \pi_c(x)\right] \\ &= \mathbb{E}\left[Y \mid X = x, \hat{Z} = \pi_c(x), M = f_M(x, \pi_c), Z = \pi_c(x)\right] - \mathbb{E}\left[Y \mid X = x, \hat{Z} = \pi_w(x), M = f_M(x, \pi_w), Z = \pi_c(x)\right] && (11) \\ &= \mathbb{E}\left[Y(\hat{Z} = \pi_c(x), M = f_M(x, \pi_c), Z = \pi_c(x)) - Y(\hat{Z} = \pi_w(x), M = f_M(x, \pi_w), Z = \pi_c(x)) \mid X = x\right] && (12) \\ &\geq 0 && (13)\end{aligned}$$

Where Eq. (11) follows from Assump. 2.1, Eq. (12) follows from Corollary B.1, and Eq. (13) follows from Assump. 3.1. We now aggregate.

$$\begin{aligned}&\mathbb{E}[Y \mid X \in \mathcal{X}_{c \neq w}, \Pi = \pi_c, Z = \pi_c(X)] - \mathbb{E}[Y \mid X \in \mathcal{X}_{c \neq w}, \Pi = \pi_w, Z = \pi_c(X)] \\ &= \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_c, Z = \pi_c(x)]\, dP(x \mid X \in \mathcal{X}_{c \neq w}, \Pi = \pi_c, Z = \pi_c(X)) \\ &\quad - \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_w, Z = \pi_c(x)]\, dP(x \mid X \in \mathcal{X}_{c \neq w}, \Pi = \pi_w, Z = \pi_c(X)) && (14) \\ &= \int_x \Big(\mathbb{E}[Y \mid X = x, \Pi = \pi_c, Z = \pi_c(x)] - \mathbb{E}[Y \mid X = x, \Pi = \pi_w, Z = \pi_c(x)]\Big)\, dP(x \mid X \in \mathcal{X}_{c \neq w}, Z = \pi_c(X)) && (15) \\ &\geq 0 && (16)\end{aligned}$$

Where Eq. (14) follows by the tower property, Eq. (15) by $\Pi \perp\!\!\!\perp (X, Z)$ (dropping Π from both measures and combining), and Eq. (16) by Eq. (13). □

[1] All falsification tests require that the relevant conditioning sets have non-zero probability of occurring. For this case, if D is uniformly distributed and independent of X, Z, it suffices that $P\big(X \in \mathcal{X}_{c \neq w},\ Z = \pi_c(X)\big) > 0$.

B.2 FALSIFICATION OF ASSUMPTION 3.2

Assumption 3.2 (Conditional Performance monotonicity)**.** Potential outcomes are monotonic in model performance relative to correctness: If $m_i < m_j$ for $m_i, m_j \in \mathcal{M}$, then for all $z \in \mathcal{Z}$ and all $\hat{z} \in \mathcal{Z}$ with $\hat{z} \neq z$,

$$Y(\hat{Z} = z, M = m_i, Z = z) \leq Y(\hat{Z} = z, M = m_j, Z = z)$$
$$Y(\hat{Z} = \hat{z}, M = m_i, Z = z) \geq Y(\hat{Z} = \hat{z}, M = m_j, Z = z)$$

Proposition 3.2 (Falsification of Assump. 3.2)**.** *Let $\pi_+, \pi_- \in \mathcal{T}$ be two trialed models. Consider a sub-population with $\mathcal{X}_{better} \subseteq \mathcal{X}$ such that for all $x \in \mathcal{X}_{better}$, $f_M(x, \pi_+) > f_M(x, \pi_-)$, and a subpopulation with covariates $\mathcal{X}_{agree} \subseteq \mathcal{X}$ such that for all $x \in \mathcal{X}_{agree}$, $\pi_+(x) = \pi_-(x)$. Then, for any $\mathcal{X}_s = \mathcal{X}_{better} \cap \mathcal{X}_{agree}$*

$$\mathbb{E}[Y|X \in \mathcal{X}_s, \Pi = \pi_+, Z = \pi_+(X)]$$
$$< \mathbb{E}[Y|X \in \mathcal{X}_s, \Pi = \pi_-, Z = \pi_+(X)]$$

falsifies Assump. 3.2.

We will use the following lemma:

Lemma B.1 (Observational Equivalence)**.** *For any $\pi_i \in \mathcal{T}$ and any $x \in \mathcal{X}$, $z \in \mathcal{Z}$ such that $P(X = x,\ Z = z,\ \Pi = \pi_i) > 0$:*

$$\mathbb{E}\left[Y|X = x, \Pi = \pi_i, Z = z\right] = \mathbb{E}\left[Y(\hat{Z} = \pi_i(x), M = f_M(x, \pi_i), Z = z)|X = x, Z = z\right] \tag{17}$$

Proof.

$$\mathbb{E}\left[Y|X = x, \Pi = \pi_i, Z = z\right] = \mathbb{E}\left[Y|X = x, \hat{Z} = \pi_i(x), M = f_M(x, \pi_i), Z = z\right] \tag{18}$$
$$= \mathbb{E}[Y(\hat{Z} = \pi_i(x), M = f_M(x, \pi_i), Z = z)|X = x, \hat{Z} = \pi_i(x), M = f_M(x, \pi_i), Z = z] \tag{19}$$
$$= \mathbb{E}\left[Y(\hat{Z} = \pi_i(x), M = f_M(x, \pi_i), Z = z)|X = x, Z = z\right] \tag{20}$$

Where Eq. (18) follows because $\hat{Z} = \Pi(X)$ and $M = f_M(X, \Pi)$ are deterministic given (X, Π), and $Y \perp\!\!\!\perp \Pi \mid X, \hat{Z}, M, Z$ under Assump. 2.1, since Π affects Y only through $(\hat{Z}, M)$. Eq. (19) follows from Consistency (Corollary B.1), noting that given $Z = z$, we have $Y(\hat{z}, m, z) = Y(\hat{z}, m)$ under Assump. 2.1. Eq. (20) follows from Conditional Ignorability (Corollary B.2), i.e., $Y(\hat{z}, m) \perp\!\!\!\perp (\hat{Z}, M) \mid X, Z$. □

This lemma enables us to prove Proposition 3.2 and 3.3.

Proof. Choosing $x \in \mathcal{X}_s$ and setting $z = \pi_+(x)$, we have that:

$$\mathbb{E}\left[Y \mid X = x, \Pi = \pi_+, Z = \pi_+(x)\right] - \mathbb{E}\left[Y \mid X = x, \Pi = \pi_-, Z = \pi_+(x)\right]$$
$$= \mathbb{E}[Y(\hat{Z} = \pi_+(x), M = f_M(x, \pi_+), Z = \pi_+(x)) - \tag{21}$$
$$Y(\hat{Z} = \pi_+(x), M = f_M(x, \pi_-), Z = \pi_+(x)) \mid X = x, Z = \pi_+(x)] \tag{22}$$
$$\geq 0 \tag{23}$$

Where Eq. (22) follows from Lemma B.1 with $z = \pi_+(x)$, noting that since $x \in \mathcal{X}_s \subseteq \mathcal{X}_{agree}$ we have $\pi_-(x) = \pi_+(x)$, so both potential outcomes share the prediction $\hat{Z} = \pi_+(x)$ and differ only in M. Eq. (23) then follows from Assump. 3.2, which compares same-prediction outcomes that differ only in performance M. For this to hold over the entire $\mathcal{X}_s$, we have:

$$\mathbb{E}[Y \mid X \in \mathcal{X}_s, \Pi = \pi_+, Z = \pi_+(X)] - \mathbb{E}[Y \mid X \in \mathcal{X}_s, \Pi = \pi_-, Z = \pi_+(X)]$$
$$= \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_+, Z = \pi_+(x)]\, dP(x \mid X \in \mathcal{X}_s, \Pi = \pi_+, Z = \pi_+(X))$$

$$- \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_-, Z = \pi_+(x)]\, dP(x \mid X \in \mathcal{X}_s, \Pi = \pi_-, Z = \pi_+(X)) \tag{24}$$

$$= \int_x \Big(\mathbb{E}[Y \mid X = x, \Pi = \pi_+, Z = \pi_+(x)] - \mathbb{E}[Y \mid X = x, \Pi = \pi_-, Z = \pi_+(x)] \Big)\, dP(x \mid X \in \mathcal{X}_s, Z = \pi_+(X)) \tag{25}$$

$$\geq 0 \tag{26}$$

Where Eq. (24) follows from the tower property, Eq. (25) follows from $\Pi \perp\!\!\!\perp (X, Z)$ (dropping Π from both measures, which makes them identical and lets the integrals combine), and Eq. (26) follows from Eq. (23), thus contradicting our observation. □

This falsification test only tests for violations of Assump. 3.2 when the models give the correct prediction. The following falsification test also exists for the case when the models give the incorrect prediction.

Proposition B.1 (Falsification of Assump. 3.2)**.** *Let $\pi_+, \pi_- \in \mathcal{T}$ be two trialed models. Consider a sub-population with $\mathcal{X}_{better} \subseteq \mathcal{X}$ such that for all $x \in \mathcal{X}_{better}$, $f_M(x, \pi_+) > f_M(x, \pi_-)$, and a subpopulation with covariates $\mathcal{X}_{agree} \subseteq \mathcal{X}$ such that for all $x \in \mathcal{X}_{agree}$, $\pi_+(x) = \pi_-(x)$. Then, for any $\mathcal{X}_s = \mathcal{X}_{better} \cap \mathcal{X}_{agree}$*

$$\begin{aligned} &\mathbb{E}[Y|X \in \mathcal{X}_s, \Pi = \pi_+, Z \neq \pi_+(X)] \\ &> \mathbb{E}[Y|X \in \mathcal{X}_s, \Pi = \pi_-, Z \neq \pi_+(X)] \end{aligned}$$

falsifies Assump. 3.2.

The proof is identical to the above, save for changes in the inequality direction and the value of Z in the expectation.

B.3 FALSIFICATION OF ASSUMPTION 3.3

Assumption 3.3 (Neutral prediction)**.** There exists $\hat{z}_0 \in \mathcal{Z}$ such that for all $m_i, m_j \in \mathcal{M}$ and $z \in \mathcal{Z}$

$$Y(\hat{Z} = \hat{z}_0, M = m_i, Z = z) = Y(\hat{Z} = \hat{z}_0, M = m_j, Z = z)$$

Proposition 3.3 (Falsification of Assump. 3.3)**.** *Let $\pi_1, \pi_2 \in \mathcal{T}$ be two trialed models. Letting $\mathcal{X}_{\hat{z}_0} := \{x \in \mathcal{X} | \pi_1(x) = \pi_2(x) = \hat{z}_0\}$, the observation that for some $z \in \mathcal{Z}$:*

$$\begin{aligned} &\mathbb{E}\left[Y|X \in \mathcal{X}_{\hat{z}_0}, \Pi = \pi_1, Z = z\right] \\ &\neq \mathbb{E}\left[Y|X \in \mathcal{X}_{\hat{z}_0}, \Pi = \pi_2, Z = z\right] \end{aligned}$$

falsifies Assump. 3.3.

Proof. Suppose Assump. 3.3 holds. Choose $x \in \mathcal{X}_{\hat{z}_0}$ and $z \in \mathcal{Z}$, then we have the following:

$$\begin{aligned} &\mathbb{E}[Y|X = x, \Pi = \pi_2, Z = z] - \mathbb{E}[Y|X = x, \Pi = \pi_1, Z = z] \\ &= \mathbb{E}[Y(\hat{Z} = \pi_2(x), M = f_M(x, \pi_2), Z = z) - \\ &\quad Y(\hat{Z} = \pi_1(x), M = f_M(x, \pi_1), Z = z) \mid X = x, Z = z] && (27) \\ &= \mathbb{E}[Y(\hat{Z} = \hat{z}_0, M = f_M(x, \pi_2), Z = z) - Y(\hat{Z} = \hat{z}_0, M = f_M(x, \pi_1), Z = z) \mid X = x, Z = z] && (28) \\ &= 0 && (29) \end{aligned}$$

Where Eq. (27) follows from Lemma B.1. Eq. (28) follows from the definitions of π_1 and π_2. Eq. (29) follows from Assump. 3.3. We then aggregate:

$$\begin{aligned} &\mathbb{E}[Y|X \in \mathcal{X}_{\hat{z}_0}, \Pi = \pi_2, Z = z] - \mathbb{E}[Y|X \in \mathcal{X}_{\hat{z}_0}, \Pi = \pi_1, Z = z] && (30) \\ &= \int_x \mathbb{E}[Y|X = x, \Pi = \pi_2, Z = z] dP(x|\Pi = \pi_2, X \in \mathcal{X}_{\hat{z}_0}, Z = z) - \\ &\quad \int_x \mathbb{E}[Y|X = x, \Pi = \pi_1, Z = z] dP(x|\Pi = \pi_1, X \in \mathcal{X}_{\hat{z}_0}, Z = z) && (31) \end{aligned}$$

$$= \int_x \mathbb{E}[Y|X = x, \Pi = \pi_2, Z = z] dP(x|X \in \mathcal{X}_{\hat{z}_0}, Z = z) -$$
$$\int_x \mathbb{E}[Y|X = x, \Pi = \pi_1, Z = z] dP(x|X \in \mathcal{X}_{\hat{z}_0}, Z = z) \tag{32}$$
$$= \int_x \mathbb{E}[Y|X = x, \Pi = \pi_2, Z = z] - \mathbb{E}[Y|X = x, \Pi = \pi_1, Z = z] dP(x|X \in \mathcal{X}_{\hat{z}_0}, Z = z) \tag{33}$$
$$= 0 \tag{34}$$

Where the logic is the same as before, contradicting our observation. □

C PROOF OF THEOREM 4.1

Definition 4.2 (Branch selector)**.** For each (x, z) we define lower- and upper-bound branch selectors $b_L, b_U : \mathcal{X} \times \mathcal{Z} \to \{\mathsf{neu}, \mathsf{mon}, \mathsf{cor}, \mathsf{def}\}$ (neutral, monotonic, correct, default), evaluated by first match:

$$b_L(x,z) = \begin{cases} \mathsf{neu} & \pi_e(x) = \hat{z}_0,\ \mathcal{S}(x) \neq \varnothing \\ \mathsf{mon} & \pi_e(x) \neq \hat{z}_0,\ \mathcal{S}_{\downarrow}(x,z) \neq \varnothing \\ \mathsf{cor} & \pi_e(x) = z,\ \mathcal{S}_{\downarrow}(x,z) = \varnothing,\ \overline{\mathcal{C}}(x,z) \neq \varnothing \\ \mathsf{def} & \text{otherwise,} \end{cases}$$

$$b_U(x,z) = \begin{cases} \mathsf{neu} & \pi_e(x) = \hat{z}_0,\ \mathcal{S}(x) \neq \varnothing \\ \mathsf{mon} & \pi_e(x) \neq \hat{z}_0,\ \mathcal{S}_{\uparrow}(x,z) \neq \varnothing \\ \mathsf{cor} & \pi_e(x) \neq z,\ \mathcal{S}_{\uparrow}(x,z) = \varnothing,\ \mathcal{C}(x,z) \neq \varnothing \\ \mathsf{def} & \text{otherwise.} \end{cases}$$

Because each selector is a function evaluated by first match, it returns exactly one value, so the induced indicators $\mathbf{1}\{b_L = k\}, \mathbf{1}\{b_U = k\}$ partition the space of $\mathcal{X} \times \mathcal{Z}$.

Theorem 4.1 (Bounds)**.** *Under Assumptions 2.1 and 3.1 to 3.4, we have the following bounds on the impact of our untrialed model π_e:*

$$L(\pi_e) \leq \mathbb{E}[Y(\hat{Z} = \pi_e(X), M = f_M(X, \pi_e))] \leq U(\pi_e)$$

Where:

$$\begin{aligned} L(\pi_e) = \mathbb{E}[&\mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}(X), Z] \\ &+ \mathbf{1}\{b_L = \mathsf{mon}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}_{\Downarrow}(X,Z), Z] \\ &+ \mathbf{1}\{b_L = \mathsf{cor}\} \max_{i \in \overline{\mathcal{C}}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z] \end{aligned} \tag{2}$$

$$+ \mathbf{1}\{b_L = \mathsf{def}\}Y_{min}]. \tag{3}$$

$$\begin{aligned} U(\pi_e) = \mathbb{E}[&\mathbf{1}\{b_U = \mathsf{neu}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}(X), Z] \\ &+ \mathbf{1}\{b_U = \mathsf{mon}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}_{\Uparrow}(X,Z), Z] \\ &+ \mathbf{1}\{b_U = \mathsf{cor}\} \min_{i \in \mathcal{C}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z] \end{aligned}$$

$$+ \mathbf{1}\{b_U = \mathsf{def}\}Y_{max}]. \tag{4}$$

C.1 PROOF OF THEOREM 4.1 (LOWER BOUND)

Since b_L is a function defined by first-match `cases` (Def. 4.2), exactly one of the indicators $\mathbf{1}\{b_L = k\}$, $k \in \{\mathsf{neu}, \mathsf{mon}, \mathsf{cor}, \mathsf{def}\}$, equals 1, so they are mutually exclusive and exhaustive. We will sometimes omit the arguments for brevity.

Thus the equality follows:

$$1 = \mathbf{1}\{b_L = \mathsf{neu}\} + \mathbf{1}\{b_L = \mathsf{mon}\} + \mathbf{1}\{b_L = \mathsf{cor}\} + \mathbf{1}\{b_L = \mathsf{def}\} \tag{35}$$

We have the following:

$$\begin{aligned} &\mathbb{E}[Y(\hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z)] \\ &= \mathbb{E}[\mathbb{E}[Y(\hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z) \mid X]] && (36) \\ &= \mathbb{E}[\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z]] && (37) \\ &= \mathbb{E}[\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z](\mathbf{1}\{b_L = \mathsf{neu}\}+ \\ &\quad \mathbf{1}\{b_L = \mathsf{mon}\}+ \end{aligned}$$

$$
\begin{aligned}
&\mathbf{1}\{b_L = \mathsf{cor}\}+ \\
&\mathbf{1}\{b_L = \mathsf{def}\})] && (38)\\
&= \mathbb{E}[\mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z]+ && (39)\\
&\mathbf{1}\{b_L = \mathsf{mon}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z]+ && (40)\\
&\mathbf{1}\{b_L = \mathsf{cor}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z]+ && (41)\\
&\mathbf{1}\{b_L = \mathsf{def}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z]] && (42)
\end{aligned}
$$

Where Eq. (36) follows from the law of iterated expectations. Eq. (37) follows from conditional ignorability (Corollary B.2) and the fact that under Assump. 2.1, $Y(\hat{z}, m, z) = Y(\hat{z}, m)$ under the conditioning. Eq. (38) follows from Eq. (35). Eq. (42)-Eq. (40) follow from distributing the expectation over the sum.

We will now demonstrate that each line is greater than or equal to the corresponding line in the lower bound.

Eq. (39) We use the analogous proof from Chen and Oberst [2025].
We first must demonstrate $Y \perp\!\!\!\perp M \mid X, \hat{Z} = \hat{z}_0$:

$$
\begin{aligned}
P(Y \mid X, \hat{Z} = \hat{z}_0, M = m, Z = z) &= P(Y(\hat{Z} = \hat{z}_0, M = m, Z = z) \mid X = x, M = m, \hat{Z} = \hat{z}_0, Z = z) && (43)\\
&= P(Y(\hat{Z} = \hat{z}_0, M = m', Z = z) \mid X = x, M = m, \hat{Z} = \hat{z}_0, Z = z) && (44)\\
&= P(Y(\hat{Z} = \hat{z}_0, M = m', Z = z) \mid X = x, M = m', \hat{Z} = \hat{z}_0, Z = z) && (45)\\
&= P(Y \mid X = x, \hat{Z} = \hat{z}_0, M = m', Z = z) && (46)
\end{aligned}
$$

Where Eq. (43) follows from Corollary B.1. Eq. (44) follows from Assump. 3.3. Eq. (45) follows from the structure of Assump. 2.1 implying $Y(\hat{z}, m) \perp\!\!\!\perp M, \hat{Z} \mid X$. Eq. (46) follows from Corollary B.1. Thus we have demonstrated $Y \perp\!\!\!\perp M \mid X, \hat{Z} = \hat{z}_0$.
Next, we recall that the lower selector takes the value neu exactly when

$$
b_L = \mathsf{neu}: \quad \pi_e(x) = \hat{z}_0,\ \mathcal{S}(x) \neq \varnothing.
$$

Thus we have:

$$
\begin{aligned}
&\mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z] \\
&= \mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y \mid X, \hat{Z} = \hat{z}_0, M = f_M(X, \pi_e), Z = z] && (47)\\
&= \mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y \mid X, \hat{Z} = \hat{z}_0, Z = z] && (48)\\
&= \mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[\mathbb{E}[Y \mid X = x, \hat{Z} = \hat{z}_0, Z = z] \mid X = x, \hat{Z} = \pi_e(x), \Pi \in \mathcal{S}(x)] && (49)\\
&= \mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[\mathbb{E}[Y \mid X = x, \hat{Z} = \hat{z}_0, M, Z = z] \mid X = x, \hat{Z} = \pi_e(x), \Pi \in \mathcal{S}(x)] && (50)\\
&= \mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[\mathbb{E}[Y \mid X = x, \hat{Z} = \pi_e(x), M, Z = z] \mid X = x, \hat{Z} = \pi_e(x), \Pi \in \mathcal{S}(x)] && (51)\\
&= \mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[\mathbb{E}[Y \mid X = x, \hat{Z} = \pi_e(x), M, \Pi \in \mathcal{S}(x), Z = z] \mid X = x, \hat{Z} = \pi_e(x), \Pi \in \mathcal{S}(x)] && (52)\\
&= \mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y \mid X = x, \hat{Z} = \pi_e(x), \Pi \in \mathcal{S}(x), Z = z] && (53)\\
&= \mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y \mid X = x, \Pi \in \mathcal{S}(x), Z = z] && (54)
\end{aligned}
$$

Where Eq. (47) follows from implication of the indicators. Eq. (48) follows from $Y \perp\!\!\!\perp M \mid X, \hat{Z} = \hat{z}_0$ demonstrated above. Eq. (49) follows from the law of total expectation. Eq. (50) follows from Assump. 3.3. Eq. (51) follows from implication of the indicators. Eq. (52) follows from $Y \perp\!\!\!\perp \Pi \mid X, \hat{Z}, M$ implied by the structure of Assump. 2.1. Eq. (53) follows from the law of total expectation. Eq. (54) follows from implication of the indicators.

Eq. (40) We have the following: First we recall that the lower selector takes the value mon exactly when

$$
b_L = \mathsf{mon}: \quad \pi_e(x) \neq \hat{z}_0,\ \mathcal{S}_{\downarrow}(x, z) \neq \varnothing.
$$

Thus we have:

$$
\mathbf{1}\{b_L = \mathsf{mon}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z]
$$

$$\geq \mathbf{1}\{b_L = \mathsf{mon}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}_{\Downarrow}(X, Z), Z = z] \tag{55}$$

Where Eq. (55) follows from Assump. 3.2 and the definition of $\mathcal{S}_{\Downarrow}(x)$ and the fact that it takes only the next-worst policies and not all worse policies. Thus, we have that the expected value of the outcome under the untrialed model is greater than or equal to the lower bound.

Eq. (41) We have the following:
We recall that the lower selector takes the value cor exactly when

$$b_L = \mathsf{cor}: \quad \pi_e(x) = z,\ \mathcal{S}_{\downarrow}(x, z) = \varnothing,\ \overline{\mathcal{C}}(x, z) \neq \varnothing.$$

For all $i \in \mathcal{T}$, we have the following:

$$\mathbf{1}\{b_L = \mathsf{cor}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z] = \mathbf{1}\{b_L = \mathsf{cor}\}\mathbb{E}[Y|X, \hat{Z} = z, M = f_M(X, \pi_e), Z = z] \tag{56}$$

$$\geq \mathbf{1}\{b_L = \mathsf{cor}\} \max_{i \in \overline{\mathcal{C}}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z = z] \tag{57}$$

Eq. (56) holds by implication of the indicators. Eq. (57) follows by Assump. 3.1, since $\pi_e(x) = z$, while every $\pi_i \in \overline{\mathcal{C}}(X, Z)$ has $\pi_i(x) \neq z$ by definition, so each such model takes an incorrect prediction.

Eq. (42) We have the following:

$$\mathbf{1}\{b_L = \mathsf{def}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z] \geq Y_{min} \tag{58}$$

Which holds by Assump. 3.4.

C.2 PROOF OF THEOREM 4.1 (UPPER BOUND)

Analogous to the lower bound, b_U is a function defined by first-match `cases` (Def. 4.2), so exactly one of the indicators $\mathbf{1}\{b_U = k\}$, $k \in \{\mathsf{neu}, \mathsf{mon}, \mathsf{cor}, \mathsf{def}\}$, equals 1, and they are mutually exclusive and exhaustive.

$$1 = \mathbf{1}\{b_U = \mathsf{neu}\} + \mathbf{1}\{b_U = \mathsf{mon}\} + \mathbf{1}\{b_U = \mathsf{cor}\} + \mathbf{1}\{b_U = \mathsf{def}\} \tag{59}$$

From the previous computation, we have the following:

$$\mathbb{E}[Y(\hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z)] = \mathbb{E}[\mathbf{1}\{b_U = \mathsf{neu}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z] + \tag{60}$$

$$\mathbf{1}\{b_U = \mathsf{mon}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z] + \tag{61}$$

$$\mathbf{1}\{b_U = \mathsf{cor}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z] + \tag{62}$$

$$\mathbf{1}\{b_U = \mathsf{def}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z]] \tag{63}$$

Using the equivalent indicator statements established in Eq. (59), we have that each line is less than or equal to the corresponding line in the upper bound.

Eq. (60) Identical to the proof in the lower bound.

Eq. (61) We recall that the upper selector takes the value mon exactly when

$$b_U = \mathsf{mon}: \quad \pi_e(x) \neq \hat{z}_0,\ \mathcal{S}_{\uparrow}(x, z) \neq \varnothing.$$

Thus we have:

$$\mathbf{1}\{b_U = \mathsf{mon}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z]$$

$$\leq \mathbf{1}\{b_U = \mathsf{mon}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}_{\Uparrow}(X, Z), Z = z] \tag{64}$$

Eq. (64) holds by Assump. 3.2 and the definition of $\mathcal{S}_{\Uparrow}(x, z)$ and conditional independence.

Eq. (62) We recall that the upper selector takes the value cor exactly when

$$b_U = \mathsf{cor}: \quad \pi_e(x) \neq z,\ \mathcal{S}_{\uparrow}(x, z) = \varnothing,\ \mathcal{C}(x, z) \neq \varnothing.$$

Thus, for all $i \in \mathcal{T}$, we have the following:

$$\mathbf{1}\{b_U = \mathsf{cor}\}\mathbb{E}[Y \mid X, \hat{Z} = \pi_e(X), M = f_M(X, \pi_e), Z = z] = \mathbf{1}\{b_U = \mathsf{cor}\}\mathbb{E}[Y|X, \hat{Z} \neq z, M = f_M(X, \pi_e), Z = z] \tag{65}$$

$$\leq \mathbf{1}\{b_U = \mathsf{cor}\} \min_{i \in \mathcal{C}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z = z] \tag{66}$$

Eq. (65) holds by implication of the indicators. Eq. (66) follows from Assump. 3.1, since $\pi_e(x) \neq z$ while every $\pi_i \in \mathcal{C}(X, Z)$, we have $\pi_i(x) = z$.

Eq. (63) Holds by the definition of Y_{max}.

Thus, we have that the expected value of the outcome under the untrialed model is less than or equal to the upper bound.

D PROOF OF THEOREM 5.1

Theorem 5.1 (Consistency and asymptotic normality of $\widehat{L}(\pi_e), \widehat{U}(\pi_e)$)**.** *If Assumptions 5.1 to 5.3 hold for some $\alpha, \beta > 0$, then*

$$\widehat{L}(\pi_e) = n^{-1} \sum_{i=1}^{n} \widehat{\psi}_L(Y_i, X_i, \Pi_i, Z_i; \widehat{f}_M) \xrightarrow{P} L(\pi_e)$$

$$\widehat{U}(\pi_e) = n^{-1} \sum_{i=1}^{n} \widehat{\psi}_U(Y_i, X_i, \Pi_i, Z_i; \widehat{f}_M) \xrightarrow{P} U(\pi_e)$$

Where $\xrightarrow{P}$ denotes convergence in probability. Furthermore,

$$\sqrt{n}(\widehat{L}(\pi_e) - L(\pi_e)) \xrightarrow{d} N(0, \sigma_L^2)$$

$$\sqrt{n}(\widehat{U}(\pi_e) - U(\pi_e)) \xrightarrow{d} N(0, \sigma_U^2)$$

Where $\xrightarrow{d}$ denotes convergence in distribution, and $\sigma_L^2 = Var(\psi_L), \sigma_U^2 = Var(\psi_U)$, with ψ_L, ψ_U denoting the pseudo-outcomes of Def. 5.2 evaluated with the true nuisance functions μ_j, f_M in place of their estimates.

Proof. We first investigate the lower bound. Recall the following lines (Eq. (6), Eq. (7), Eq. (8), Eq. (9)). We also define $\hat{\mu}_S(X, Z)$ as the estimator for $E[Y|X, \Pi \in S, Z] = \mu_S(X, Z)$, and $\overline{\mathcal{C}}(x, z)$ is the set of policies, indexed by i, whose outputs do not agree with the true label at x.

Recall that conceptually, $d_L(X, Z)$ is the index of the model with the highest expected outcome among those that do not agree with π_e at X. $\mathcal{S}_{\Downarrow}(x, z)$ is the best-performing set of agreeing models with performance less than or equal to the untrialed model. $\varphi_L\big(X, Z; \hat{d}_L\big)$ is the estimator for the outcome under the model with index $d_L(X, Z)$.

First, we will demonstrate that $\hat{L}(\pi_e)$ is a consistent estimator of $L(\pi_e)$, as follows. Recall our definition of the estimator from Def. 5.2, using the sets defined in Def. 4.1 and Theorem 4.1:

$$\begin{aligned}\widehat{L}(\pi_e) = \mathbb{E}_n\left[\widehat{\psi}_L(Y_i, X_i, \Pi_i, Z_i; \widehat{f}_M)\right] = \mathbb{E}_n\Big[&\mathbf{1}\{\hat{b}_L = \mathsf{neu}\}\, \hat{h}(X, Y, \mathcal{S}) && (67)\\ &+ \mathbf{1}\{\hat{b}_L = \mathsf{mon}\}\, \hat{h}(X, Y, \widehat{\mathcal{S}}_{\Downarrow}) && (68)\\ &+ \mathbf{1}\{\hat{b}_L = \mathsf{cor}\}\, \varphi_L(X, Z; \hat{d})\\ &+ \mathbf{1}\{\hat{b}_L = \mathsf{def}\}\, Y_{\min}\Big].\end{aligned}$$

Where we use the following substitution for brevity:

$$\hat{h}(X, Y, S) := \left(\widehat{\mu}_S(X, Z) + \frac{\mathbf{1}\{\Pi \in S\}}{P(\Pi \in S \mid X)}\big(Y - \widehat{\mu}_S(X, Z)\big)\right)$$

Where we use Π to denote the random variable and $\mathcal{T}$ as its support. S is used to denote a subset of $\mathcal{T}$, and π to denote an instance of Π. (See Def. 4.1 where we defined notation).

We first demonstrate the following Lemma, in which h, φ_L, and d_L denote the functionals of Definitions 5.1 and 5.2 evaluated with the true nuisance functions μ_j and true performance f_M in place of their estimates:

Lemma D.1 (Unbiased Estimator)**.**

$$\begin{aligned}L(\pi_e) = \mathbb{E}\left[\psi_L(Y_i, X_i, \Pi_i, Z_i; f_M)\right] = \mathbb{E}\Big[&\mathbf{1}\{b_L = \mathsf{neu}\}\, h(X, Y, \mathcal{S})\\ &+ \mathbf{1}\{b_L = \mathsf{mon}\}\, h(X, Y, \mathcal{S}_{\Downarrow})\\ &+ \mathbf{1}\{b_L = \mathsf{cor}\}\, \varphi_L(X, Z; d)\\ &+ \mathbf{1}\{b_L = \mathsf{def}\}\, Y_{\min}\Big].\end{aligned}$$

By linearity of expectation, we have:

$$= \mathbb{E}\Big[\mathbf{1}\{b_L = \mathsf{neu}\}\, h(X, Y, \mathcal{S})\Big] \tag{69}$$

$$+ \mathbb{E}\Big[\mathbf{1}\{b_L = \mathsf{mon}\}\, h(X, Y, \mathcal{S}_{\Downarrow})\Big] \tag{70}$$

$$+ \mathbb{E}\Big[\mathbf{1}\{b_L = \mathsf{cor}\}\, \varphi_L(X, Z; d)\Big] \tag{71}$$

$$+ \mathbb{E}\Big[\mathbf{1}\{b_L = \mathsf{def}\}\, Y_{\min}\Big]. \tag{72}$$

For Eq. (69)-Eq. (71), each term consists of a regression term and a correction term as follows:

$$\mathbb{E}[\mathbf{1}\{\cdot\}(\mu(X, Z) + \lambda(X, Z))] \tag{73}$$

μ is defined as the corresponding conditional expectation term in Theorem 4.1. For λ:

$$\begin{aligned}
\mathbb{E}[\mathbf{1}\{\cdot\}\lambda(X, Z)] &= \mathbb{E}[\mathbb{E}[\mathbf{1}\{\cdot\}\lambda(X, Z) \mid X, Z]] \\
&= \mathbb{E}[\mathbf{1}\{\cdot\}\mathbb{E}[\lambda(X, Z) \mid X, Z]] \\
&= \mathbb{E}[\mathbf{1}\{\cdot\}0] \\
&= 0
\end{aligned}$$

Where the first line follows by the tower property and the fact that the branch selector $\mathbf{1}\{\cdot\}$ is measurable with respect to (X, Z). The second line follows from condition 2 of Lemma D.2. Thus, we have:

$$\begin{aligned}
L(\pi_e) = \mathbb{E}\left[\psi_L(Y_i, X_i, \Pi_i, Z_i; f_M)\right] = \mathbb{E}\Big[&\mathbf{1}\{b_L = \mathsf{neu}\}\, \mu_{\mathcal{S}}(X, Z) \\
&+ \mathbf{1}\{b_L = \mathsf{mon}\}\, \mu_{\mathcal{S}_{\Downarrow}}(X, Z) \\
&+ \mathbf{1}\{b_L = \mathsf{cor}\}\, \mu_{d_L(X,Z)}(X, Z) \\
&+ \mathbf{1}\{b_L = \mathsf{def}\}\, Y_{\min}\Big].
\end{aligned}$$

Which is precisely $L(\pi_e)$. The proof for the upper bound is analogous.

We have (omitting the arguments (Y_i, X_i, Π_i, Z_i) for brevity, retaining only the performance function after the semicolon; the unhatted $\psi_L(\,\cdot\,; f_M)$ denotes the pseudo-outcome computed with the true nuisance functions and true performance):

$$\begin{aligned}
\widehat{L}(\pi_e) - L(\pi_e) &= \mathbb{E}_n\left[\widehat{\psi}_L(Y_i, X_i, \Pi_i, Z_i; \widehat{f}_M)\right] - \mathbb{E}\left[\psi_L(Y, X, \Pi, Z; f_M)\right] \qquad (74) \\
&= \underbrace{\mathbb{E}_n\left[\widehat{\psi}_L(\,\cdot\,; \widehat{f}_M)\right] - \mathbb{E}_n\left[\widehat{\psi}_L(\,\cdot\,; f_M)\right]}_{R_1} + \underbrace{\mathbb{E}_n\left[\widehat{\psi}_L(\,\cdot\,; f_M)\right] - \mathbb{E}\left[\psi_L(\,\cdot\,; f_M)\right]}_{R_2}
\end{aligned}$$

Where Eq. (74) follows from Lemma D.1.

We will demonstrate

$$\begin{aligned}
\widehat{L}(\pi_e) - L(\pi_e) &= R_1 + R_2 \\
&= O_P(n^{-1/2})
\end{aligned}$$

We will focus on R_2 first.

$$\begin{aligned}
R_2 = &\underbrace{(\mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{neu}\}\hat{h}(X, Y, \mathcal{S})] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}, Z]])}_{R_{2,1}} \\
&+ \underbrace{(\mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{mon}\}\hat{h}(X, Y, \mathcal{S}_{\Downarrow})] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{mon}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}_{\Downarrow}, Z]])}_{R_{2,2}}
\end{aligned}$$

$$+ \underbrace{(\mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{cor}\}\varphi_L(X, Z; \hat{d})] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{cor}\} \max_{i \in \overline{\mathcal{C}}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z]])}_{R_{2,3}}$$
$$+ \underbrace{Y_{min}(\mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{def}\}] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{def}\}])}_{R_{2,4}}$$
$$= R_{2,1} + R_{2,2} + R_{2,3} + R_{2,4}$$

We will leave $R_{2,4}$ as is. We will consider $R_{2,3}$ first. We note that the proofs for $R_{2,1}, R_{2,2}$ are identical except for the sets S being considered, as they can be interpreted as being a maximum over a subset of $\mathcal{T}$ that contains one element. For example:

$$\mathbb{E}[\mathbf{1}\{b_L = \mathsf{neu}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}, Z]] = \mathbb{E}[\mathbf{1}\{b_L = \mathsf{neu}\} \max_{j \in \{1\}} \mathbb{E}[Y|X, \Pi \in \mathcal{S}, Z]] \tag{75}$$

The term $R_{2,3}$ corresponds to the following branch in the lower bound:

$$\mathbb{E}[\mathbf{1}\{b_L = \mathsf{cor}\} \max_{i \in \overline{\mathcal{C}}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z]]$$

We note this term has an "expectation of max" form. This structure has non-smooth properties that prevent straightforward estimation via plugins or other methods. As a result, we can utilize an augmented inverse probability-weighted estimator. Previous literature exists on similar topics. For example, Byun et al. [2024] considers a problem of using AIPW to estimate the expected maximum of a set of functions. We will adapt their results to our setting.

Definition D.1. Consider the following class of estimators. Let $\theta_j(W; P), \lambda_j(W; P)$ for all $j \in \{1, \ldots, J\}$ represent some arbitrary set of smooth functions and their associated influence functions, where W is the observed data and P is the true data distribution. Use of $\widehat{P}$ indicates the parameter was taken with respect to an estimated distribution. We further define the following functionals:

$$\theta_j := \mathbb{E}[\theta_j(W; P)] \tag{76}$$
$$\widehat{\theta}_j := \mathbb{E}_n[\theta_j(W; \widehat{P})] \tag{77}$$
$$\psi := \mathbb{E}\left[\max_{j \in \{1,\ldots,J\}} \theta_j(W; P)\right] = \mathbb{E}\left[\theta_{d(W)}(W; P)\right] \tag{78}$$
$$d(W) := \arg\max_{j \in \{1,\ldots,J\}} \theta_j(W; P) \tag{79}$$
$$\widehat{\psi}_j := \mathbb{E}_n[\theta_j(W; \widehat{P}) + \lambda_j(W; \widehat{P})] \tag{80}$$
$$\widehat{\psi} := \mathbb{E}_n[\varphi(W; \widehat{P}, \widehat{d})] \tag{81}$$
$$\varphi(W; \widehat{P}, \widehat{d}) := \theta_{\widehat{d}(W)}(W; \widehat{P}) + \lambda_{\widehat{d}(W)}(W; \widehat{P}) \tag{82}$$
$$\widehat{d}(W) := \arg\max_{j \in \{1,\ldots,J\}} \theta_j(W; \widehat{P}) \tag{83}$$

We note that these definitions are copied verbatim from Byun et al. [2024]. We can now state their main lemma:

Lemma D.2 (Lemma 10 and 11 of Byun et al. [2024])**.** *For the class of estimators defined in Def. D.1, if the following conditions hold:*

1. *For every $j \in \{1, \ldots, J\}$, $\lambda_j(W; P)$ and $\theta_j(W; P)$ are both uniformly bounded by constants with respect to n.*
2. *For every $j \in \{1, \ldots, J\}$, $\mathbb{E}[\lambda_j(W; P)] = 0$*
3. *For every $j \in \{1, \ldots, J\}$, $\|\widehat{\theta}_j - \theta_j\| = o_P(1)$*
4. *There exists $\alpha > 0$ such that: $P\left[\min_{j \neq d(W)} |\theta_{d(W)}(W) - \theta_j(W)| \leq t\right] \lesssim t^\alpha$*
5. *In Eq. (81), the expectation is taken with respect to $\widehat{P}_1$, while the estimator $\varphi(W; \widehat{P}_2, \widehat{d})$ uses an independent sample $\widehat{P}_2$.*

Then the following is true:

$$\widehat{\psi} - \psi = \mathbb{E}_n[\varphi(W;\ P, d)] - \mathbb{E}[\varphi(W; P, d)] \tag{84}$$

$$+ O_P\left(\|\widehat{\theta}_j - \theta_j\|_\infty^{1+\alpha} + \max_{j\in\{1,\dots,J\}} \mathbb{E}\left[\theta_j(W;\widehat{P}) + \lambda_j(W;\widehat{P}) - \theta_j(W;P)\right]\right) \tag{85}$$

$$+ o_P(n^{-1/2}) \tag{86}$$

If two additional conditions hold:

6. $\|\widehat{\theta}_j - \theta_j\|_\infty^{1+\alpha} = o_P(n^{-1/2})$
7. *For each* $j \in \{1,\dots,J\}$, $\mathbb{E}[\theta_j(W;\widehat{P}) + \lambda_j(W;\widehat{P}) - \theta_j(W;P)] = o_P(n^{-1/2})$

Then the following simplification holds:

$$\widehat{\psi} - \psi = \mathbb{E}_n[\varphi(W;\ P,d)] - \mathbb{E}[\varphi(W;P,d)] + o_P(n^{-1/2}) \tag{87}$$

And the error is asymptotically normal:

$$\sqrt{n}(\widehat{\psi} - \psi) \xrightarrow{d} N(0, Var(\varphi(W;P,d))) \tag{88}$$

We recognize that our problem fits into this framework with the following identifications:

$$\begin{aligned}
W &\leftrightarrow (X,Z)\\
\theta_j(W;P) &\leftrightarrow \mu_j(X,Z)\\
\widehat{\theta}_j(W;\widehat{P}) &\leftrightarrow \hat{\mu}_j(X,Z)\\
\theta_j &\leftrightarrow \mathbb{E}[\mu_j(X,Z)]\\
\widehat{\theta}_j &\leftrightarrow \mathbb{E}_n[\hat{\mu}_j(X,Z)]\\
\lambda_j(W;P) &\leftrightarrow \lambda_j(X,Z)\\
d(W) &\leftrightarrow d_L(X,Z)\\
\widehat{d}(W) &\leftrightarrow \hat{d}(X,Z)\\
J &\leftrightarrow \overline{\mathcal{C}}(x,z)
\end{aligned}$$

Where any unspecified functionals are defined analogously to their counterparts in Eq. (76) - Eq. (83). Here the conditioning argument W of Byun et al. [2024] is identified with the pair (X,Z), since each branch of Theorem 4.1 conditions on (X,Z). Because model assignment is randomized, $\Pi \perp\!\!\!\perp Z \mid X$ and hence $P(\Pi = j \mid X,Z) = P(\Pi = j \mid X)$, so the propensity nuisance is unchanged and only the outcome regression μ_j acquires the Z argument. Thus, we can apply Lemma D.2 to our setting,

Lemma D.3. *Suppose the following conditions hold:*

1. *For every* $j \in \overline{\mathcal{C}}(x,z)$, $\lambda_j(X,Z)$ *and* $\hat{\mu}_j(X,Z)$ *are both uniformly bounded by constants.*
2. $\mathbb{E}[\lambda_j(X,Z) \mid X,Z] = 0$ *for all* $j \in \overline{\mathcal{C}}(x,z)$.
3. *For every* $j \in \overline{\mathcal{C}}(x,z)$,
$$|\mathbb{E}_n[\hat{\mu}_j(X,Z)] - \mathbb{E}[\mu_j(X,Z)]| = o_P(1)$$
4. *There exists* $\alpha > 0$ *such that:*
$$P\left[\min_{j\neq d_L(X,Z)} \left|\mu_{d_L(X,Z)}(X,Z) - \mu_j(X,Z)\right| \le t\right] \le t^\alpha$$
5. *Independent samples are used for* $\mathbb{E}_n\left[\widehat{\psi}_L(Y_i,X_i,\Pi_i,Z_i;\widehat{f}_M)\right]$ *and* $\varphi_L(X,Z;\hat{d})$.

Then the difference between the estimator and the true parameter can be expressed as:

$$\begin{aligned}
\widehat{\psi}^l - \psi^l &= \mathbb{E}_n[\varphi_L(X,Z;d)] - \mathbb{E}[\varphi_L(X,Z;d)]\\
&+ O_P\left(\|\mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}[\mu_j(X,Z)]\|_\infty^{1+\alpha} + \max_{j\in\overline{\mathcal{C}}(x,z)} \mathbb{E}\left[\widehat{\mu}_j(X,Z) + \hat{\lambda}_j(X,Z) - \mu_j(X,Z)\right]\right)\\
&+ o_P(n^{-1/2})
\end{aligned}$$

We can further simplify this with the following 2 additional conditions:

6. *The following holds:*
$$\|\mathbb{E}_n[\hat{\mu}_j(X,Z)] - \mathbb{E}[\mu_j(X,Z)]\|_\infty^{1+\alpha} = o_P(n^{-1/2})$$
where α is the same as in Assump. 5.2.
7. *For each $j \in \overline{\mathcal{C}}(x,z)$,*
$$\mathbb{E}[\hat{\mu}_j(X,Z) + \hat{\lambda}_j(X,Z) - \mu_j(X,Z)] = o_P(n^{-1/2}).$$

Thus the following holds:

$$\widehat{\psi}^l - \psi^l = \mathbb{E}_n[\varphi_L(X,Z;d)] - \mathbb{E}[\varphi_L(X,Z;d)] + o_P(n^{-1/2})$$

We will demonstrate the conditions hold for $R_{2,3}$.

1. (Boundedness): For every $j \in \overline{\mathcal{C}}(x,z)$, $\lambda_j(X,Z)$ and $\hat{\mu}_j(X,Z)$ are both uniformly bounded by constants.

   This is true by Assump. 3.4 and can be enforced by clipping the value of the nuisance function.

2. (Zero-mean Correction Term): $\mathbb{E}[\lambda_j(X,Z)|X,Z] = 0$ for all $j \in \overline{\mathcal{C}}(x,z)$.

$$\begin{aligned}
\mathbb{E}[\lambda_j(X,Z)|X,Z] &= \mathbb{E}\left[\frac{\mathbf{1}\{\Pi = j\}}{P(\Pi = j|X)}(Y - g_j(X,Z))\Big|X,Z\right] && (89)\\
&= \frac{\mathbb{E}\left[\mathbf{1}\{\Pi = j\}(Y - g_j(X,Z))|X,Z\right]}{P(\Pi = j|X)} && (90)\\
&= \frac{P\left(\Pi = j|X,Z\right)\left(\mathbb{E}\left[Y|X,\Pi = j,Z\right] - g_j(X,Z)\right)}{P(\Pi = j|X)} && (91)\\
&= \left(\mathbb{E}\left[Y|X,\Pi = j,Z\right] - g_j(X,Z)\right) && (92)\\
&= 0 && (93)
\end{aligned}$$

   Eq. (89) follows by the definition of $\lambda_j(X,Z)$. Eq. (90) follows because the propensity is constant under the conditioning. Eq. (91) follows because the expectation of the indicator is the probability of the event, and $g_j(X,Z)$ is a constant under the conditions. Eq. (92) follows because randomized model assignment gives $\Pi \perp\!\!\!\perp Z \mid X$, so $P(\Pi = j|X,Z) = P(\Pi = j|X)$ and the ratio of propensities is one. Eq. (93) follows by the definition of $g_j(X,Z)$ as a plug-in estimator for $\mathbb{E}[Y|X,\Pi = j,Z]$.

3. (Consistent Plug-in Estimator): For every $j \in \overline{\mathcal{C}}(x,z)$,
$$\|\mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}[\mu_j(X,Z)]\| = o_P(1)$$

   This will be satisfied by the validation of Condition 7, so discussion is deferred to there.

4. (Margin Condition): There exists $\alpha > 0$ such that:
$$P\left[\min_{j \neq d_L(X,Z)} \left|\mu_{d_L(X,Z)}(X,Z) - \mu_j(X,Z)\right| \leq t\right] \leq t^\alpha$$

   This condition asserts that the probability of a small gap between the outcome regressions μ_j of the best two models is bounded polynomially. This holds by the first margin condition of Assump. 5.2.

5. (Independent Samples):

   This holds by construction. As described in Appendix E, data are split into $K \geq 2$ folds, with all nuisance functions ($\widehat{\mu}_j$ and $\widehat{f}_M$, etc.) fit on the first and the bounds estimated on the second.

6. (Convergence of plug-in estimators)

$$\|\mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}[\mu_j(X,Z)]\|_\infty^{1+\alpha} = o_P(n^{-\frac{1}{2}}), \text{where } \alpha \text{ is the same as in Assump. 5.2.}$$

This is stating that the maximum difference between the empirical expectation of our machine learning-derived estimate from the true expected outcome, across all j, converges at $o_P(n^{-\frac{1}{2}})$ rate. We will prove this by demonstrating that all plug-in estimators converge sufficiently fast.

We want to show:

$$\forall j, \|\mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}[\mu_j(X,Z)]\|^{1+\alpha} = o_P(n^{-\frac{1}{2}})$$

Consider the following:
For each $j \in \overline{\mathcal{C}}(x,z)$, we have:

$$\begin{aligned}
\mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}[\mu_j(X,Z)] &= \mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}\left[\hat{\mu}_j(X,Z)\right] + \mathbb{E}\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}[\mu_j(X,Z)] && (94)\\
&= \left(\mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}\left[\hat{\mu}_j(X,Z)\right]\right) + \left(\mathbb{E}\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}[\mu_j(X,Z)]\right) && (95)
\end{aligned}$$

Looking solely at the first term, we have:

$$\begin{aligned}
\mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}\left[\hat{\mu}_j(X,Z)\right] &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}_j(X_i,Z_i) - \mathbb{E}\left[\hat{g}_j(X_i,Z_i)\right]\right) && (96)\\
\sqrt{n}\left(\mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}\left[\hat{\mu}_j(X,Z)\right]\right) &\xrightarrow{d} N(0,\sigma^2) && (97)\\
\implies \mathbb{E}_n\left[\hat{\mu}_j(X,Z)\right] - \mathbb{E}\left[\hat{\mu}_j(X,Z)\right] &= O_p(n^{-\frac{1}{2}}) && (98)
\end{aligned}$$

Which follows from the CLT.

The second term is the bias of the machine learning estimator. In order for the whole norm to converge at $o_P(n^{-\frac{1}{2}})$, we require this term to converge at a rate of $o_P(n^{-\frac{1}{2(1+\alpha)}})$. For a reasonable α, such as $\alpha = 1$, this is satisfied by a convergence rate of $o_P(n^{-\frac{1}{4}})$. It is reasonable to assume our model converges at this rate.
We now have:

$$\begin{aligned}
\|O(n^{-\frac{1}{2}}) + o_P(n^{-\frac{1}{2(1+\alpha)}})\|^{1+\alpha} &\leq \left(\|O(n^{-\frac{1}{2}})\| + \|o_P(n^{-\frac{1}{2(1+\alpha)}})\|\right)^{1+\alpha} && (99)\\
&\leq 2^\alpha\left(\|O(n^{-\frac{1}{2}})\|^{1+\alpha} + \|o_P(n^{-\frac{1}{2(1+\alpha)}})\|^{1+\alpha}\right) && (100)\\
&= O(n^{-\frac{1+\alpha}{2}}) + o_P(n^{-\frac{1}{2}}) && (101)\\
&= o_P(n^{-\frac{1}{2}}) && (102)
\end{aligned}$$

Where Eq. (99) follows from the triangle inequality. Eq. (100) follows from the C_r inequality with $r = 1+\alpha$ and the constant gets absorbed. Eq. (101) evaluates the rates. Eq. (102) since $\alpha > 0$. Since this holds for all $j \in \overline{\mathcal{C}}(x,z)$, we have shown our result.

7. (Convergence of bias-corrected estimator) For each

$$j \in \overline{\mathcal{C}}(x,z), \mathbb{E}\left[\widehat{\mu}_j(X,Z) + \hat{\lambda}_j(X,Z) - \mu_j(X,Z)\right] = o_P(n^{-\frac{1}{2}}).$$

Where $\overline{\mathcal{C}}(x,z)$ is the set of models in the branch.

For clarity, we rewrite the full bias-corrected estimator as:

$$\varphi_L(X,Z;\hat{d}) = \widehat{\mu}_{\hat{d}}(X,Z) + \frac{\mathbf{1}\left\{\pi = \hat{d}\right\}}{P(\Pi = \hat{d}|X)}(Y - \widehat{\mu}_{\hat{d}}(X,Z))$$

Thus, we have the following nuisance functions:

$$\mu_j(X,Z), \quad P(\Pi = j|X)$$

We will assume that $\mu_j(X,Z)$ converges at a rate of $o_P(n^{-\frac{1}{4}})$. We also assume that $P(\Pi = j|X)$ is known.
We have, for arbitrary $j \in \overline{\mathcal{C}}(x,z)$:

$$\mathbb{E}\left[\widehat{\mu}_j(X,Z) + \hat{\lambda}_j(X,Z) - \mu_j(X,Z)\right] = \mathbb{E}\left[\widehat{\mu}_j(X,Z) - \mu_j(X,Z) + \frac{\mathbf{1}\{\Pi = j\}}{P(\Pi = j|X)}(Y - \widehat{\mu}_j(X,Z))\right] \tag{103}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\widehat{\mu}_j(X,Z) - \mu_j(X,Z) + \frac{\mathbf{1}\{\Pi = j\}}{P(\Pi = j|X)}(Y - \widehat{\mu}_j(X,Z)) \mid X, Z\right]\right] \tag{104}$$

$$= \mathbb{E}\left[\widehat{\mu}_j(X,Z) - \mu_j(X,Z) + \frac{1}{P(\Pi = j|X)}\mathbb{E}\left[\mathbf{1}\{\Pi = j\}(Y - \widehat{\mu}_j(X,Z)) \mid X, Z\right]\right] \tag{105}$$

$$= \mathbb{E}\left[\widehat{\mu}_j(X,Z) - \mu_j(X,Z) + \frac{P(\Pi = j|X,Z)}{P(\Pi = j|X)}\mathbb{E}\left[(Y - \widehat{\mu}_j(X,Z)) \mid X, \Pi = j, Z\right]\right] \tag{106}$$

$$= \mathbb{E}\left[\widehat{\mu}_j(X,Z) - \mu_j(X,Z) + \frac{P(\Pi = j|X,Z)}{P(\Pi = j|X)}(\mu_j(X,Z) - \widehat{\mu}_j(X,Z))\right] \tag{107}$$

$$= \mathbb{E}\left[(\widehat{\mu}_j(X,Z) - \mu_j(X,Z))\left(1 - \frac{P(\Pi = j|X,Z)}{P(\Pi = j|X)}\right)\right] \tag{108}$$

$$= \mathbb{E}\left[(\widehat{\mu}_j(X,Z) - \mu_j(X,Z))(1-1)\right] \tag{109}$$

$$= 0 \tag{110}$$

$$= o_P(n^{-\frac{1}{2}}) \tag{111}$$

Eq. (103) follows from Assump. 5.1. Eq. (104) follows from the law of iterated expectation. Eq. (105) follows because the relevant expressions are constants under the conditioning. Eq. (106) follows from removing the indicator from the definition of conditional expectation. Eq. (107) follows by the definition of $\mu_j(X,Z)$. Eq. (108) is an algebraic rearrangement; Eq. (109) uses $P(\Pi = j|X,Z) = P(\Pi = j|X)$, which holds because randomized model assignment gives $\Pi \perp\!\!\!\perp Z \mid X$. Eq. (111) follows vacuously from the definition of convergence in probability.
Thus, we have verified Condition 7. This also verifies Condition 3. Since the bias corrected estimator converges at a rate of $o_P(n^{-\frac{1}{2}})$ to the true value, it follows that the uncorrected plug-in estimator is at least consistent, since the correction term has mean 0.

We have thus verified all necessary conditions. As a result, the following equality holds:

$$R_{2,3} = \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{cor}\}\varphi(X,Z;\hat{d})] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{cor}\}\varphi(X,Z;\hat{d})] + o_P(n^{-1/2})$$

It follows that the same applies to $R_{2,1}, R_{2,2}, R_{2,4}$, since the $\hat{h}(X,Y,S)$ functionals are identical in structure to $\varphi(x,z;\hat{d})$, except for the subscript of $\hat{\mu}$, which instead is a single nuisance function that converges to the true value at $o_P(n^{-\frac{1}{2}})$ under Assump. 5.3, allowing the whole function $\hat{h}(X,Y,S)$ to converge at $o_P(n^{-\frac{1}{2}})$ rate under analogous logic.

Thus we have:

$$\begin{aligned} R_2 =& \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{neu}\}h(X,Y,\mathcal{S})] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{neu}\}h(X,Y,\mathcal{S})] + \\ & \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{mon}\}h(X,Y,\mathcal{S}_\Downarrow)] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{mon}\}h(X,Y,\mathcal{S}_\Downarrow)] + \\ & \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{cor}\}\varphi(X,Z;d)] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{cor}\}\varphi(X,Z;d)] + \\ & \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{def}\}Y_{min}] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{def}\}Y_{min}] + o_P(n^{-1/2}) \end{aligned}$$

We will now consider $R_1 = \mathbb{E}_n\left[\widehat{\psi}_L(\cdot\,;\widehat{f}_M)\right] - \mathbb{E}_n\left[\widehat{\psi}_L(\cdot\,;f_M)\right]$:

The theoretical construction of ψ_L involves the construction of the set $\mathcal{S}_\downarrow(x,z)$, defined in Def. 4.1 as:

$$\mathcal{S}_\downarrow(x,z) = \{\pi_i \in \mathcal{S}(x) \mid f_M(x,\pi_i) \leq f_M(x,\pi_e),\ \pi_e(x) = z\}$$

$$\cup \{\pi_i \in \mathcal{S}(x) \mid f_M(x, \pi_i) \geq f_M(x, \pi_e),\ \pi_e(x) \neq z\}$$

Where $f_M(x, \pi_i)$ is the true performance of model π_i at covariate x. If we instead use the estimated performance $\widehat{f}_M$ to construct the set, as in Def. 5.1, we have:

$$\begin{aligned}\widehat{\mathcal{S}}_{\downarrow}(x, z) = &\{\pi_i \in \mathcal{S}(x) \mid \widehat{f}_M(x, \pi_i) \leq \widehat{f}_M(x, \pi_e),\ \pi_e(x) = z\} \\ &\cup \{\pi_i \in \mathcal{S}(x) \mid \widehat{f}_M(x, \pi_i) \geq \widehat{f}_M(x, \pi_e),\ \pi_e(x) \neq z\}\end{aligned}$$

However, we only concern ourselves with the model(s) whose performance is closest to that of the untrialed model, per Definitions 4.1 and 5.1:

$$\mathcal{S}_{\Downarrow}(x, z) = \underset{\pi_i \in \mathcal{S}_{\downarrow}(x,z)}{\arg\min} |f_M(x, \pi_i) - f_M(x, \pi_e)|$$

$$\widehat{\mathcal{S}}_{\Downarrow}(x, z) = \underset{\pi_i \in \widehat{\mathcal{S}}_{\downarrow}(x,z)}{\arg\min} |\widehat{f}_M(x, \pi_i) - \widehat{f}_M(x, \pi_e)|$$

However, error in our estimation may permute the ordering of the models relative to the untrialed model, in terms of their estimated performance, resulting in:

$$\widehat{\mathcal{S}}_{\Downarrow}(x, z) \neq \mathcal{S}_{\Downarrow}(x, z)$$

This leads to two distinct cases:

1. Case 1: $\widehat{\mathcal{S}}_{\Downarrow}(x, z) = \mathcal{S}_{\Downarrow}(x, z)$
2. Case 2: $\widehat{\mathcal{S}}_{\Downarrow}(x, z) \neq \mathcal{S}_{\Downarrow}(x, z)$

In case 1, we have that:

$$\widehat{\psi}_L(\cdot\,; \widehat{f}_M) = \widehat{\psi}_L(\cdot\,; f_M)$$

Since $\widehat{\psi}_L$ depends on the performance function only through the set $\mathcal{S}_{\Downarrow}(x, z)$. If the estimates have no impact on the set, then the estimates have no impact on the value of $\widehat{\psi}_L$. In case 2, we know that the error is bounded: $\left|\widehat{\psi}_L(\cdot\,; f_M) - \widehat{\psi}_L(\cdot\,; \widehat{f}_M)\right| \leq |Y_{max} - Y_{min}| = B$, by Assump. 3.4.[2]

To prove convergence, we utilize the second margin condition of Assump. 5.2: there exists $\beta > 0$ such that for any $t \geq 0$, we have:

$$P\left[\min_{i \neq j} |f_M(x; \pi_j) - f_M(x; \pi_i)| \leq t\right] \leq t^{\beta}$$

For all $x \in \mathcal{X}$ and $\pi_i, \pi_j \in \mathcal{T} \cup \{\pi_e\}$.

This condition states that the probability that the minimum performance gap between any two models is less than t decays at least polynomially in t. Under this assumption, we can show that the first term converges to 0 as $n \to \infty$ by demonstrating the indicator converges to 0:

We first define:

$$\delta_n := 2 \max_i \left|\left|f_M(x; \pi_i) - \widehat{f}_M(x; \pi_i)\right|\right|_{\infty}$$

$$g(X) := \min_{i,j} |f_M(X; \pi_i) - f_M(X; \pi_j)|$$

It follows that:

$$|R_1| \leq \mathbb{E}_n\left[\left|\widehat{\psi}_L(\cdot\,; \widehat{f}_M) - \widehat{\psi}_L(\cdot\,; f_M)\right|\right] \tag{112}$$

$$= \mathbb{E}_n\left[\left|\mathbf{1}\left\{\widehat{\mathcal{S}}_{\Downarrow}(X, Z) \neq \mathcal{S}_{\Downarrow}(X, Z)\right\}\left(\widehat{\psi}_L(\cdot\,; \widehat{f}_M) - \widehat{\psi}_L(\cdot\,; f_M)\right)\right|\right] \tag{113}$$

$$\leq \mathbb{E}_n\left[\left|\mathbf{1}\left\{\min_{i,j} |f_M(X; \pi_i) - f_M(X; \pi_j)| \leq 2 \max_i \left|\left|f_M(x; \pi_i) - \widehat{f}_M(x; \pi_i)\right|\right|_{\infty}\right\}\right.$$

[2]This assumption is technically stronger than necessary, as only sufficient separation between the performance of the two closest models with π_e on both sides is necessary.

$$\left(\widehat{\psi}_L(\cdot\,;\widehat{f}_M) - \widehat{\psi}_L(\cdot\,;f_M)\right)|\Bigg] \tag{114}$$

$$= \mathbb{E}_n\left[\left|\mathbf{1}\left\{g(X) \le \delta_n\right\}\left(\widehat{\psi}_L(\cdot\,;\widehat{f}_M) - \widehat{\psi}_L(\cdot\,;f_M)\right)\right|\right] \tag{115}$$

$$\le \mathbb{E}_n\left[\left|\mathbf{1}\left\{g(X) \le \delta_n\right\} B\right|\right] \tag{116}$$

$$= \mathbb{E}_n\left[\mathbf{1}\left\{g(X) \le \delta_n\right\}\right] B \tag{117}$$

Where B is an arbitrary constant, $\pi_i, \pi_j \in \mathcal{T} \cup \{\pi_e\}$. Eq. (112) is Jensen's inequality. Eq. (113) follows because the difference is only non-zero when the sets differ. Eq. (114) follows because the sets differ only when model error causes permutations in the ordering of π_e and the 2 closest performing models on either side. This occurs when the estimation error is larger than half of the minimum distance between the untrialed model and any other model, as maximum error in both values can "bridge the gap". Eq. (115) is by the definitions of δ_n and $g(X)$. Eq. (116) follows from the boundedness of the difference from Assump. 3.4, where $B = Y_{max} - Y_{min}$. Eq. (117) follows because the indicator is non-negative. From here, we cannot directly apply the margin condition of Assump. 5.2, because the expectation is empirical; instead, we have:

$$\begin{aligned}
P\left(\sqrt{n}\,|R_1| > \epsilon\right) &\le P\left(\mathbb{E}_n\left[\mathbf{1}\{g(X_i) \le \delta_n\}\right] \ge \tfrac{\epsilon}{B\sqrt{n}}\right) && (118)\\
&\le \mathbb{E}\left(\mathbb{E}_n\left[\mathbf{1}\{g(X_i) \le \delta_n\}\right]\right)\tfrac{B\sqrt{n}}{\epsilon} && (119)\\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}\{g(X_i) \le \delta_n\}\right]\right)\tfrac{B\sqrt{n}}{\epsilon} && (120)\\
&= P(g(X_i) \le \delta_n)\,\tfrac{B\sqrt{n}}{\epsilon} && (121)\\
&\le \delta_n^{\beta}\tfrac{B\sqrt{n}}{\epsilon} && (122)\\
&\longrightarrow 0. && (123)
\end{aligned}$$

Eq. (118) uses $|R_1| \le B\mathbb{E}_n[\mathbf{1}\{g(X_i) \le \delta_n\}]$. Eq. (119) follows from Markov's inequality. Equation (120) is by definition. Eq. (121) is by linearity of expectation. Eq. (122) follows from the margin condition (Assump. 5.2), and Eq. (123) follows since $\delta_n = o_P(n^{-\frac{1}{2\beta}})$ by Assump. 5.3, so $\delta_n^{\beta}\sqrt{n} \to 0$ in probability. Because this holds for every $\epsilon > 0$, $R_1 = o_P(n^{-1/2})$.

We note that *this is different* than the required convergence rate for the plug-in estimators for $\mu_j(X, Z)$, which was $o_P(n^{-\frac{1}{2(1+\alpha)}})$. This is because in Byun et al. [2024], or our 6th condition of Lemma D.3, at the step analogous to Eq. (114), the term in the right hand side inside of the indicator is the same as the term outside of the indicator. After application of the margin condition, this results in a squared convergence rate. In our case, the term outside of the indicator is B, a constant, which does not converge. We are unable to improve this, since the convergence of $\widehat{f}_M$ is not necessarily related to the convergence of $\widehat{\psi}_L$, whose error is bounded by B. We have now shown:

$$\begin{aligned}
\widehat{L}(\pi_e) - L(\pi_e) =& \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{neu}\}h(X,Y,\mathcal{S})] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{neu}\}h(X,Y,\mathcal{S})] +\\
& \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{mon}\}h(X,Y,\mathcal{S}_{\Downarrow})] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{mon}\}h(X,Y,\mathcal{S}_{\Downarrow})] +\\
& \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{cor}\}\varphi(X,Z;d)] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{cor}\}\varphi(X,Z;d)] +\\
& \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{def}\}Y_{min}] - \mathbb{E}[\mathbf{1}\{b_L = \mathsf{def}\}Y_{min}]\\
& + o_P(n^{-1/2}) + \underbrace{o_P(n^{-1/2})}_{R_1} && (124)\\
=& \mathbb{E}_n[\mathbf{1}\{b_L = \mathsf{neu}\}h(X,Y,\mathcal{S}) + \mathbf{1}\{b_L = \mathsf{mon}\}h(X,Y,\mathcal{S}_{\Downarrow}) + && (125)\\
& \mathbf{1}\{b_L = \mathsf{cor}\}\varphi(X,Z;d) + \mathbf{1}\{b_L = \mathsf{def}\}Y_{min}] - && (126)\\
& \mathbb{E}[\mathbf{1}\{b_L = \mathsf{neu}\}h(X,Y,\mathcal{S}) + \mathbf{1}\{b_L = \mathsf{mon}\}h(X,Y,\mathcal{S}_{\Downarrow}) + && (127)\\
& \mathbf{1}\{b_L = \mathsf{cor}\}\varphi(X,Z;d) + \mathbf{1}\{b_L = \mathsf{def}\}Y_{min}] && (128)\\
& + o_P(n^{-1/2}) && (129)
\end{aligned}$$

In Eq. (124), the term outside the underbrace corresponds to R_2. In Eq. (129), we have the difference of an empirical mean and the corresponding true mean. By the CLT, this is a random variable that converges to a normal distribution. By Slutsky's, the $o_P(n^{-1/2})$ terms vanish asymptotically.

It directly follows that since $\widehat{L}(\pi_e)$ is an empirical function of i.i.d. random variables, we can apply the CLT to show that:

$$\sqrt{n}(\widehat{L}(\pi_e) - L(\pi_e)) \sim N\left(0, \sigma_L^2\right) \tag{130}$$

By the CLT, where $\sigma_L^2 = Var(\psi_L)$. It follows:

Corollary 5.1 (Confidence Intervals for $L(\pi_e), U(\pi_e)$)**.** *Under the conditions of Theorem 5.1, we can construct the following confidence intervals for $L(\pi_e), U(\pi_e)$:*

$$\widehat{L}(\pi_e) \pm z_{1-\gamma/2}\sqrt{\widehat{Var}(\widehat{\psi}_L)/n}$$
$$\widehat{U}(\pi_e) \pm z_{1-\gamma/2}\sqrt{\widehat{Var}(\widehat{\psi}_U)/n}$$

Where $z_{1-\gamma/2}$ is the $1-\gamma/2$ quantile of the standard normal distribution, and $\widehat{Var}(\widehat{\psi}_L), \widehat{Var}(\widehat{\psi}_U)$ are the empirical variance of $\widehat{\psi}_L, \widehat{\psi}_U$, respectively.

Proof. By Theorem 5.1, $\sqrt{n}(\widehat{L}(\pi_e) - L(\pi_e)) \xrightarrow{d} N(0, \sigma_L^2)$ with $\sigma_L^2 = \text{Var}(\psi_L)$. We must show $\widehat{\text{Var}}(\widehat{\psi}_L) \xrightarrow{P} \sigma_L^2$. Decompose

$$\widehat{\text{Var}}(\widehat{\psi}_L) - \text{Var}(\psi_L) = \underbrace{\widehat{\text{Var}}(\widehat{\psi}_L) - \text{Var}(\widehat{\psi}_L)}_{(A)} + \underbrace{\text{Var}(\widehat{\psi}_L) - \text{Var}(\psi_L)}_{(B)}.$$

For term (A), $\widehat{\psi}_L$ is uniformly bounded by a constant independent of n: Assump. 3.4 bounds $Y \in [Y_{\min}, Y_{\max}]$, and the influence terms $\widehat{\lambda}_j$ are bounded by Condition 1 of Lemma D.3. Under cross-fitting, the evaluation-fold observations are i.i.d. conditional on the training fold, on which $\widehat{\psi}_L$ is a fixed bounded function; a law of large numbers for this bounded conditionally-i.i.d. sample therefore gives $\widehat{\text{Var}}(\widehat{\psi}_L) - \text{Var}(\widehat{\psi}_L) = o_P(1)$, so $(A) = o_P(1)$.

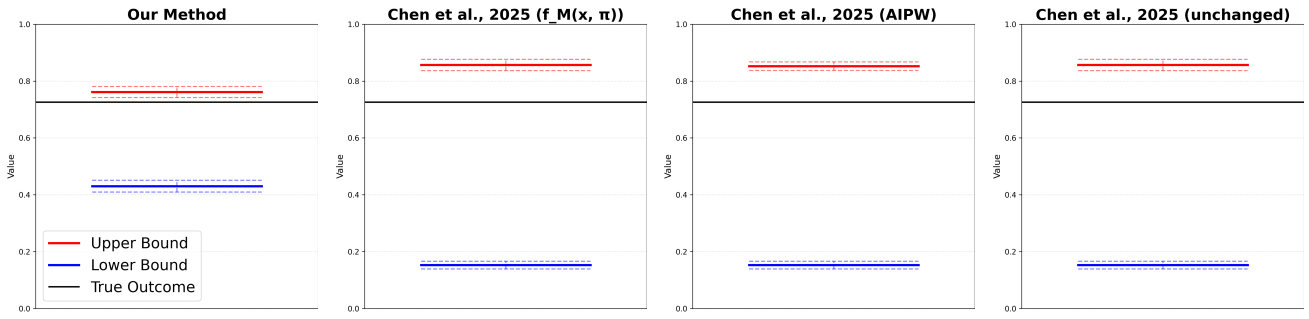
For term (B), the rate condition Assump. 5.3 and margin condition Assump. 5.2 give $\widehat{\psi}_L \xrightarrow{P} \psi_L$ (as established in the proof of Theorem 5.1). Since $\widehat{\psi}_L$ is uniformly bounded, the bounded convergence theorem upgrades convergence in probability to moment convergence, $\mathbb{E}[\widehat{\psi}_L] \to \mathbb{E}[\psi_L]$ and $\mathbb{E}[\widehat{\psi}_L^2] \to \mathbb{E}[\psi_L^2]$, hence $\text{Var}(\widehat{\psi}_L) \to \text{Var}(\psi_L)$ and $(B) = o_P(1)$.

Thus, $\widehat{\text{Var}}(\widehat{\psi}_L) \xrightarrow{P} \sigma_L^2$, and by Slutsky's theorem the interval attains asymptotic coverage $1-\gamma$. □

We now consider the upper bound.

$$\begin{aligned} U(\pi_e) = \mathbb{E}[&\mathbf{1}\{b_U = \textsf{neu}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}(X), Z] \\ &+ \mathbf{1}\{b_U = \textsf{mon}\}\mathbb{E}[Y|X, \Pi \in \mathcal{S}_{\Uparrow}(X, Z), Z] \\ &+ \mathbf{1}\{b_U = \textsf{cor}\} \min_{i \in \mathcal{C}(X,Z)} \mathbb{E}[Y|X, \Pi = i, Z] \\ &+ \mathbf{1}\{b_U = \textsf{def}\}Y_{max}]. \end{aligned}$$

Where the sets in the indicators are defined as they are in Theorem 4.1. We note that the structure of this expression is identical to that of the lower bound, except for the use of minima instead of maxima, and the different definitions of the sets in the indicators. Since the proof was agnostic to the specific definitions of the sets, and a minimum can be interpreted as a maximum over negative values, the proof follows identically. □



(a) (b) (c) (d)

Figure A.1: Ablation study of generated bounds on downstream outcomes. The figure displays the lower bound and upper bound computed using our proposed method in Fig. A.1a, the existing method in Chen and Oberst [2025] using performance stratified across age groups in Fig. A.1b, the existing method using AIPW estimation in Fig. A.1c, and the existing method without any modifications in Fig. A.1d. 90% two-sided confidence intervals are shown.

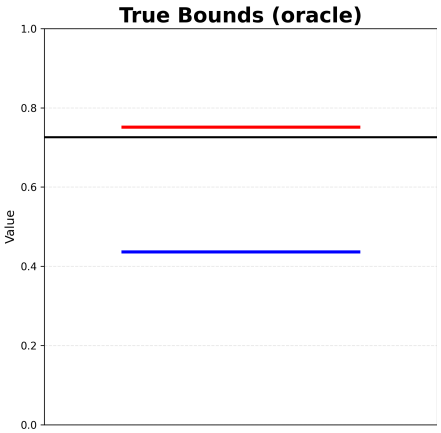


Figure A.2: The figure displays the lower and upper bound computed with our method using the true theoretical values, as if we had access to the counterfactual outcomes for all models.

E FURTHER SIMULATION DETAILS

Method	Bound width	Lower (mean)	Lower 90% CI	Upper (mean)	Upper 90% CI	True value
Our Method	0.331	0.430	[0.409, 0.451]	0.761	[0.742, 0.781]	0.726
True Bounds (oracle)	0.315	0.437	–	0.751	–	0.726
Chen et al., 2025 ($f_M(x, \pi)$ stratified)	0.704	0.153	[0.139, 0.166]	0.857	[0.837, 0.877]	0.726
Chen et al., 2025 (AIPW)	0.700	0.153	[0.139, 0.166]	0.852	[0.837, 0.867]	0.726
Chen et al., 2025 (unchanged)	0.704	0.153	[0.139, 0.166]	0.857	[0.837, 0.877]	0.726

Table A.1: (Experiment 1) Numerical summary of the ablation study results.

The experimental trial consisted of 1,891 defendants from the dataset whose various covariates, such as demographics, charge history, and staff recommendations were recorded. The original dataset also contained an indicator of whether the judge was shown the risk score predicted by the Public Safety Assessment (PSA) model, the judge's detention decision, and the outcome of whether the defendant failed to appear in court (FTA). We use these data to construct all the variables in our RCT, as outlined in the experiments section of the main text.

In the original setting of Imai et al. [2023], judges were able to give high or low bail, and models provided integer risk scores on a scale of 1-6 on not only FTA, but also new criminal activity and new violent criminal activity. The score is also always shown to the judge. Our method is still applicable to these settings, however we choose to artificially binarize the model

Experiment	Bounds	Bound width	Lower (mean)	Lower 90% CI	Upper (mean)	Upper 90% CI	True value
Experiment 3	Estimated	0.140	0.624	[0.607, 0.641]	0.764	[0.749, 0.779]	0.696
Experiment 3	Ground truth	0.144	0.615	–	0.760	–	0.696
Experiment 5	Estimated	0.204	0.586	[0.568, 0.604]	0.790	[0.776, 0.804]	0.697
Experiment 5	Ground truth	0.209	0.578	–	0.787	–	0.697

Table A.2: Semi-synthetic experiment results comparison between experiment 3 and experiment 5, which differ only in the presence of missing data.

prediction and modify the setting to an alerting context so that a meaningful neutral action ($\hat{z}_0$) exists. We chose to binarize the judge's decision and omit the other scores largely for simplicity.

Van Amsterdam et al. [2025] discusses a setting in which performance is dependent on the data distribution on which it is measured. Thus, under a performance-dependent intervention, the dataset will shift, causing confounding that may lead to orthogonal changes between potential outcomes and performance. We highlight this due to its similarity to our setting, but emphasize that it does not apply. This is because performance is not measured relative to the potential outcomes, but instead relative to counterfactual outcomes that do not change with any interventions.

The nuisance models used to learn the relationships in the data to simulate judge decisions and outcomes were simple logistic regressions, without any feature engineering outside of the existing features in the dataset. The nuisance models used in the AIPW estimation were lasso regularized linear regressions with cross-validated L_1 penalty.

E.1 ADDITIONAL EXPERIMENT DETAILS

We now present additional details to some of the experiments discussed. We present some of the figures for each experiment, but any graphs and tables presented have analogous versions for every experiment. They are available in the GitHub repository. In no experiment do the falsification tests fail.

Experiment 1 Suppose the alerting model discussed in Experiment 3 already exists and a two-arm RCT was conducted comparing the FTA of the alerting model against a control arm (never alert). We are considering lowering the threshold of the alerting model so that it alerts when the same risk score is >1 instead, without running an RCT. We use our method to develop bounds on the downstream FTA from the data in the two-arm RCT.

We use the same experimental setup as in Experiment 3, except that we now have an additional trialed model. Thus, we first randomly assign each defendant to either the control arm (never alert) or the trialed model (alert if risk score > 3). The values of A, Y are generated only for the appropriate model per the random assignment.

We show our results of this experiment in Fig. A.3. We see that our bounds remain tighter than the existing methods, though both are significantly wider than in Experiment 4. The difference between the two methods is also less pronounced. This is because the additional trialed model decreases the likelihood that the untrialed model disagrees with all trialed models, which activates the $\mathbf{1}\{b_L = \text{cor}\}, \mathbf{1}\{b_U = \text{cor}\}$ terms less often. This demonstrates that our method is more effective in settings with more limited data, and has diminishing returns as the number of trialed models increases.

Experiment 2 Suppose a model that alerts for offenders with a Imai et al. [2023] risk score greater than 3 already exists, and a two-arm RCT was conducted comparing the FTA of the alerting model against a control arm (never alert). We develop a new model that still alerts when the risk score is > 3; however, it generates those risk scores in a fundamentally different way. We want to evaluate the effectiveness of the new risk score algorithm without running an RCT. We use our method to develop bounds on the downstream FTA from the data in the two-arm RCT.

We first randomly assign each defendant to either the control arm (never alert) or the trialed model (alert if risk score > 3). The values of A, Y are generated only for the appropriate model per the random assignment. The untrialed model doesn't use the risk scores from the Imai et al. [2023] data. Instead, it finds the top 5 non-demographic features most correlated with FTA in the data. It then does a simple tally of the number of those features that are present for a given defendant, and uses that as the risk score, where 0 features present corresponds to a risk score of 1, and 5 features present corresponds to a risk score of 6.

We show the results of this experiment in Fig. A.4. We see that our bounds are somewhat looser but still comparable to the

bounds in Fig. A.3. This demonstrates that our method still remains more effective than existing methods even when the trialed model is fundamentally different from the untrialed model.

Experiment 5 This experiment is identical to Experiment 3, but with missing values of Z. Z is defined as the counterfactual outcome under no bail from the judge. Thus, it is only observed for individuals where $A = 0$, but never observed for individuals where $A = 1$. In this experiment, we set all values of Z for such individuals to $Z = -1$, which prevents the logic in $\mathbf{1}\{b_L = \mathsf{cor}\}, \mathbf{1}\{b_U = \mathsf{cor}\}, \mathbf{1}\{b_L = \mathsf{mon}\}, \mathbf{1}\{b_U = \mathsf{mon}\}$ from activating. As discussed in Appendix G.1, this is a possibly misspecified experiment.

We see that our bounds are somewhat looser but still comparable to the bounds in Fig. 2. They remain tighter than the existing methods, which are not affected by the missingness of Z. We see that in this case, the violation of MAR does not significantly affect the bounds. This is because the fallback to $\mathbf{1}\{b_L = \mathsf{def}\}, \mathbf{1}\{b_U = \mathsf{def}\}$ branches remain valid, the $\mathbf{1}\{b_L = \mathsf{neu}\}, \mathbf{1}\{b_U = \mathsf{neu}\}$ branches for missing Z values are only impacted through the estimation of $\mu(X, Z)$, and the $\mathbf{1}\{b_L = \mathsf{cor}\}, \mathbf{1}\{b_U = \mathsf{cor}\}, \mathbf{1}\{b_L = \mathsf{mon}\}, \mathbf{1}\{b_U = \mathsf{mon}\}$ branches do not get activated sufficiently to significantly affect the bounds.

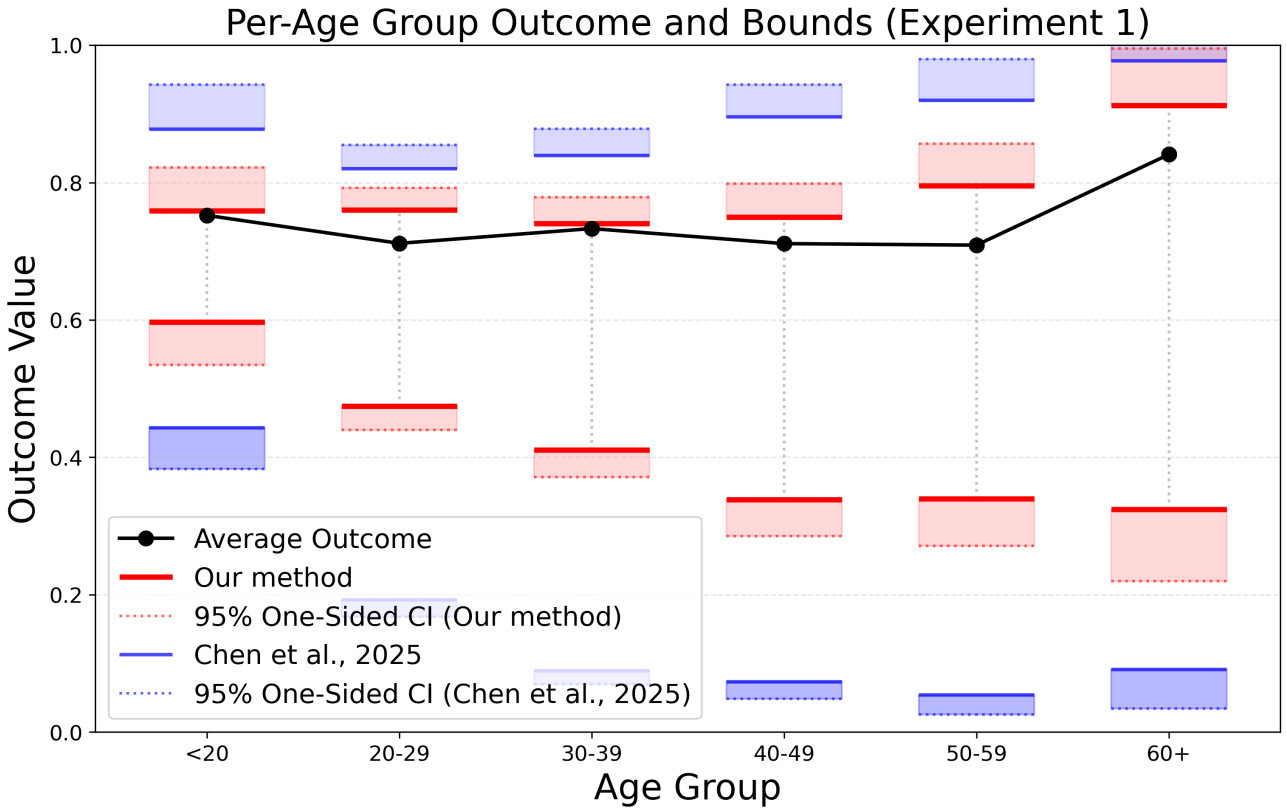


Figure A.3: (Experiment 1) Bounds on downstream outcomes for an untrialed model with two trialed models in the RCT, plotted as functions of individual age.

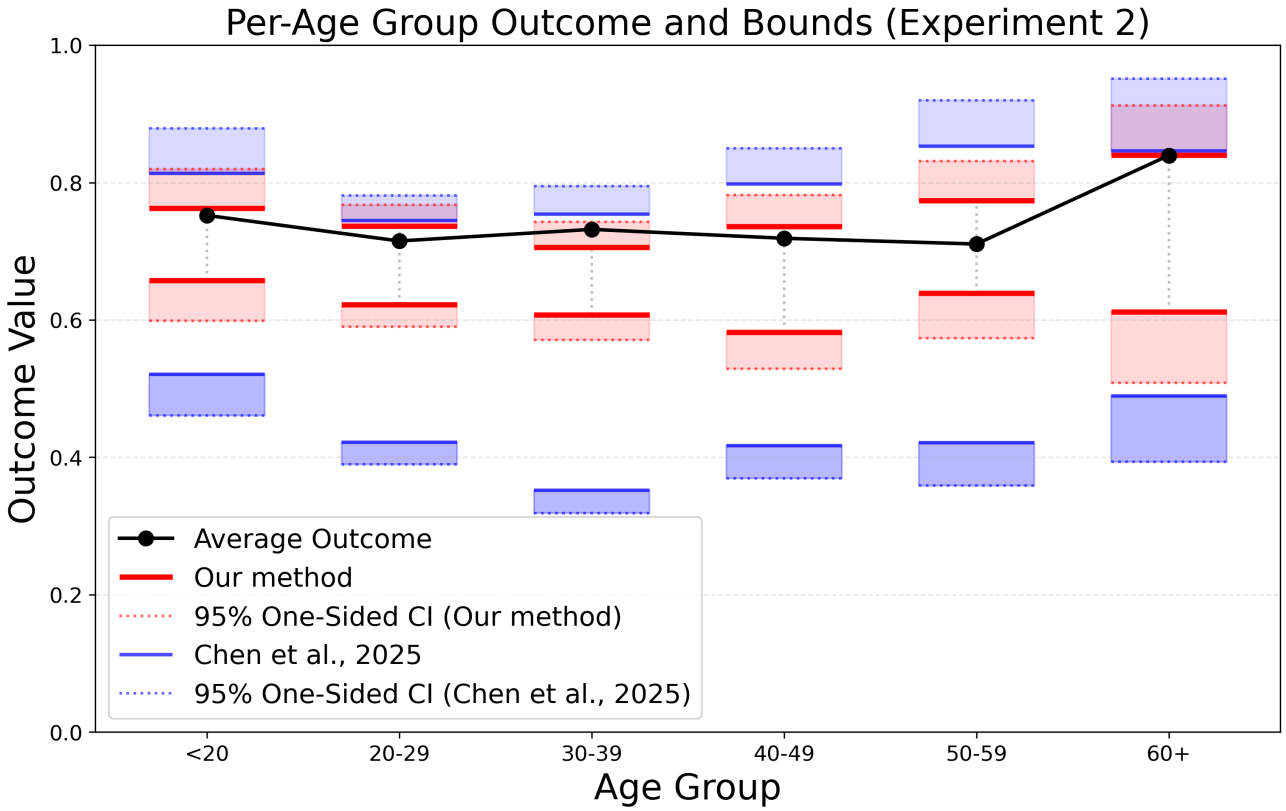


Figure A.4: (Experiment 2) Bounds on downstream outcomes for an untrialed model under a covariate tally alerting rule, plotted as functions of age.

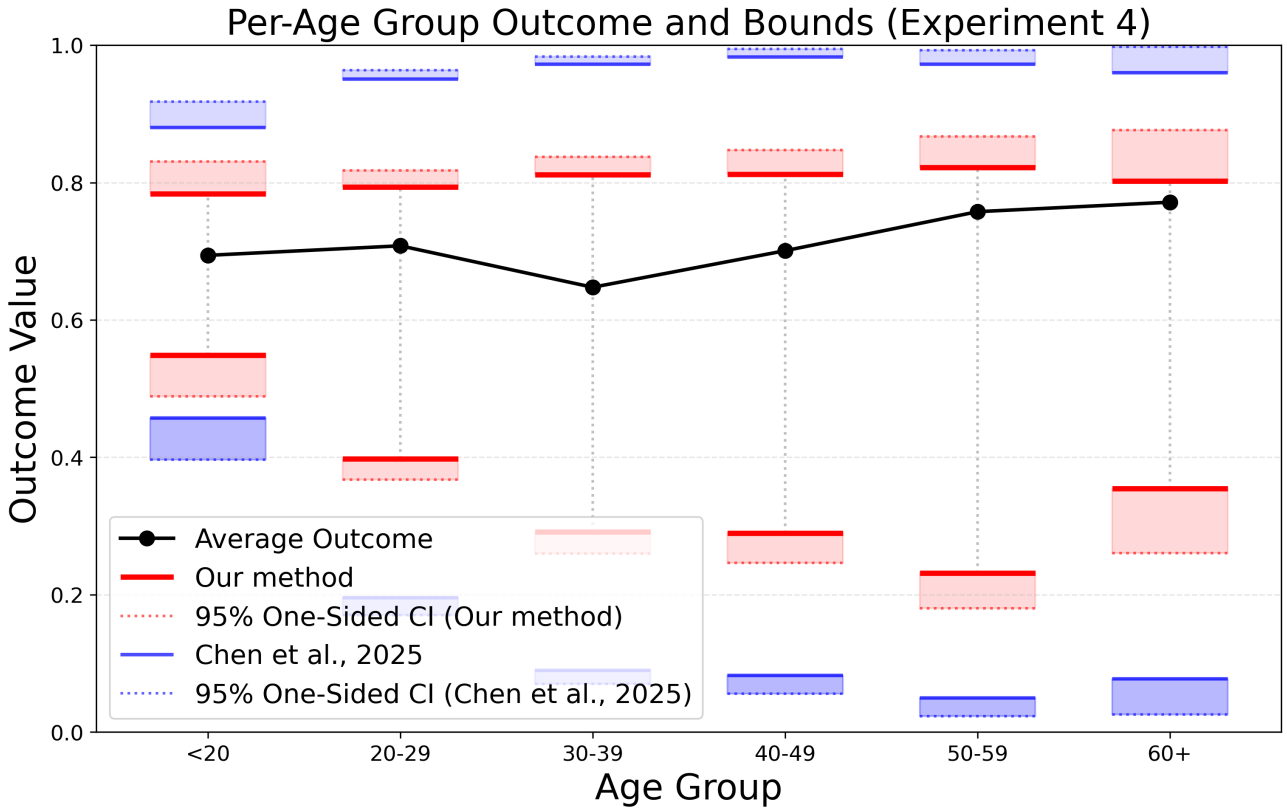


Figure A.5: (Experiment 4) Bounds on downstream outcomes for an untrialed model with a single trialed model in the RCT, plotted as functions of age.

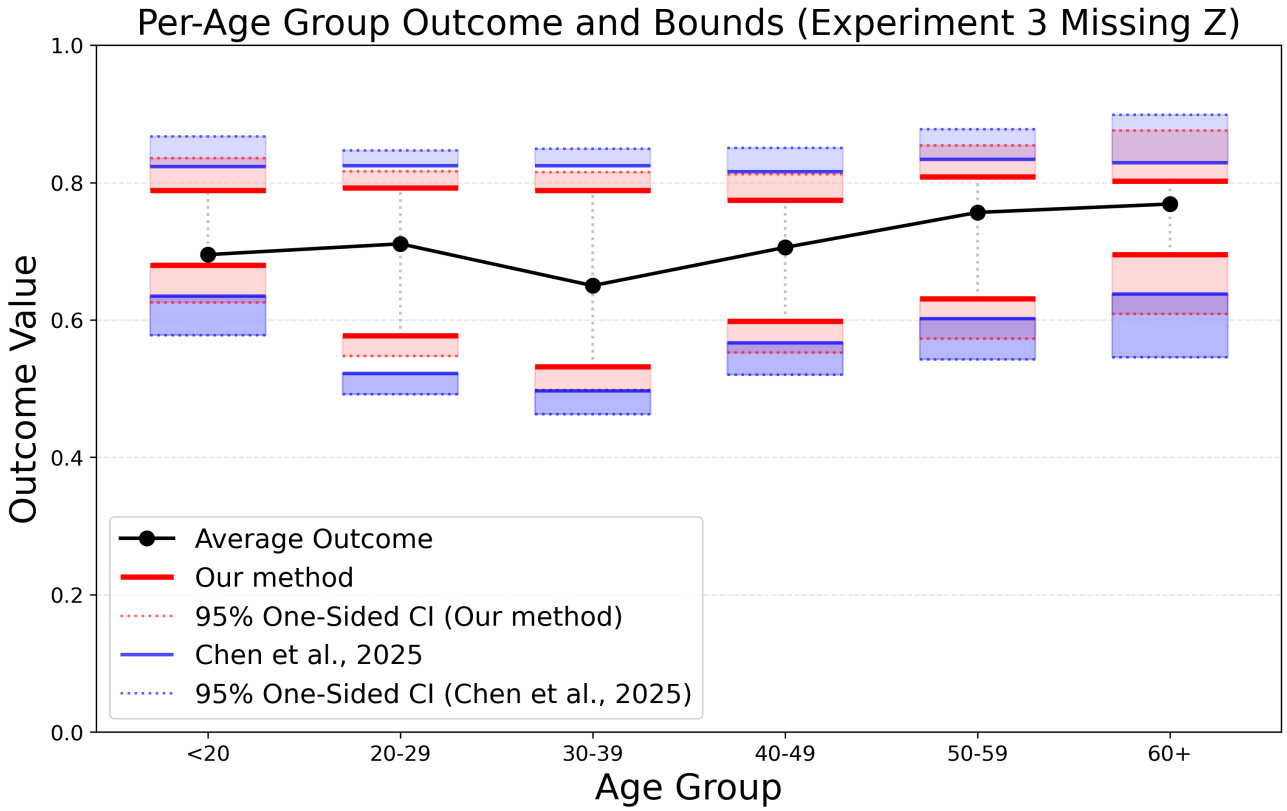


Figure A.6: (Experiment 5) Bounds on downstream outcomes for an untrialed model with a single trialed model in the RCT, plotted as functions of age, with missing data.

Patient_ID	age_group	UPPER	LOWER	UPPER (ground-truth)	LOWER (ground-truth)	TRUE_EXPECTED
0	4	1.000	1.000	0.681	0.681	0.681
1	4	0.909	0.000	0.651	0.000	0.652
2	3	1.000	1.000	0.818	0.818	0.818
3	2	1.000	0.909	1.000	0.472	0.476
4	2	1.000	1.000	0.777	0.777	0.777
5	4	1.000	1.000	0.715	0.715	0.715
6	3	1.000	0.000	0.612	0.000	0.613
7	2	0.091	0.000	0.672	0.000	0.672
8	3	1.000	1.000	0.687	0.687	0.687
9	4	1.000	0.000	0.676	0.000	0.677
10	2	0.091	0.000	0.665	0.000	0.666
11	4	1.000	0.000	1.000	0.570	0.572
12	2	1.000	1.000	1.000	0.647	0.649

Table A.3: The bound values generated for the first 13 individuals in the dataset (Experiment 3). The UPPER and LOWER columns refer to the values of the upper and lower bounds generated using this paper's method. These sometimes fall outside $[0, 1]$ or fail to capture the true value due to estimation error. The ground truth columns are the bounds generated using the true theoretical values. The TRUE_EXPECTED column is the true expected outcome for the individual, which was unknown to the methods.

Method	Outcome cov.	Interval width	Lower CI cov.	Lower CI width	Upper CI cov.	Upper CI width
Our Method	1.00	0.313	0.97	0.042	0.91	0.039
True Bounds (oracle)	1.00	0.315	–	–	–	–
Chen et al. ($f_M(x, \pi)$)	1.00	0.692	0.00	0.027	0.00	0.039
Chen et al. (unchanged)	1.00	0.692	0.00	0.027	0.00	0.039
Chen et al. (AIPW)	1.00	0.693	0.00	0.027	0.00	0.031

Table A.4: Monte Carlo coverage and width across 100 seeds (Experiment 1). The Outcome cov. column is how often the final bound interval [lower, upper] covers the true outcome. The Lower/Upper CI cov.: how often the two-sided 90% CI around the lower (resp. upper) bound covers the true lower (resp. upper) bound. Widths are means.

Policy	Precision	Alert fraction
Control	–	0.000
$\mathfrak{R} > 1$	0.349	0.802
$\mathfrak{R} > 3$	0.456	0.281
CustomRiskScore	0.413	0.456

Table A.5: Overall precision and alert frequency of each alerting policy used across the experiments, evaluated on the full dataset ($n = 1891$). The alert fraction is the proportion of defendants for whom the policy raises an alert. Precision is the fraction of raised alerts for which the defendant would fail to appear if released. The Control policy never alerts, so its precision is undefined (–). "$\mathfrak{R} > 1$" and "$\mathfrak{R} > 3$" are the models that threshold the Imai et al. [2023] PSA failure-to-appear risk score, and CustomRiskScore is the custom alerting system introduced in Experiment 2.

FTAScore	Count	Proportion
1	375	0.198
2	531	0.281
3	454	0.240
4	286	0.151
5	189	0.100
6	56	0.030

Table A.6: Baseline distribution of the Imai et al. [2023] PSA failure-to-appear risk score (integer values 1–6) across all defendants in the dataset ($n = 1891$). This is the score thresholded by the $\mathfrak{R} > 1$ and $\mathfrak{R} > 3$ policies in Table A.5.

F RELATED WORK

Offline Policy Evaluation Our work is related to offline policy evaluation [Su et al., 2020, Wu et al., 2023, Rothfuss et al., 2023, Namkoong et al., 2020]. If we interpret a model as a deterministic policy that dictates treatments, we are seeking to evaluate the effectiveness of a new decision-making policy, using recorded outcomes under different policies. However, our work differs due to the presence of a human decision-maker, whose compliance with the policy is dependent on many factors such as their perception of the policy's effectiveness. Furthermore, in offline policy evaluation, if the unknown policy makes a prediction not present in data, we cannot estimate its effect due to a positivity violation. However, our setting specifically focuses on using counterfactual correctness to resolve this situation.

Transportability and Z-Identification Transportability studies compute the causal effect of an intervention in a "target" environment from data on the same intervention in a "source" environment. The two environments differ structurally, often due to a selection node [Pearl, 2015, Mao et al., 2022, Shi et al., 2023, Lal et al., 2023, Stuart et al., 2011, Pearl and Bareinboim, 2011]. Conversely, Z-identification studies compute the causal effect of an intervention based on data of a different intervention in the same environment [Correa and Bareinboim, 2020b, Bareinboim and Pearl, 2012, Lee et al., 2019, Correa and Bareinboim, 2020a]. We are related to both fields, but more closely to z-identification, as we are estimating the causal effect of an untrialed model (different intervention) deployed in the existing environment of the RCT. We differ from both approaches in that we make parametric assumptions, like monotonicity, and our goal is not to give conditions for unbiased point estimation from infinite data. Instead, we give an interval containing the effect from finite data when the above algorithms preclude unbiased point estimation.

Partial Identification Partial identification is concerned with bounding causal effects when point estimation is impossible. It begins with Manski's assumption-free natural bounds and progresses to sharper bounds under graphical assumptions via linear and polynomial programming [Manski, 1990, Robins, 1989, Chickering and Pearl, 1996, Balke and Pearl, 1994, 1997, Richardson et al., 2014, Zhang and Bareinboim, 2021a,b, Bellot, 2024, Bellot and Chiappa, 2024, Duarte et al., 2024]. We differ from these approaches because our bounds use non-graphical assumptions pertinent to the application (as well as graphical assumptions), and we use experimental (interventional) data instead of observational data. Furthermore, the causal effect we bound occurs in a structurally different environment.

Evaluation of Machine Learning Models under Distribution Shift Distribution shift occurs when the environment in which a model is deployed differs structurally (not randomly) from the environment in which it is trained, violating the assumption that train and test sets are identically distributed [Taori et al., 2020, Hendrycks et al., 2021, Wiles et al., 2022, Koh et al., 2021, Miller et al., 2021]. In our setting, the presence of a human decision-maker causes the deployment

Bound	Branch	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5
Lower	def	698 (36.91%)	238 (12.59%)	289 (15.28%)	987 (52.19%)	406 (21.47%)
Lower	cor	529 (27.97%)	356 (18.83%)	242 (12.80%)	529 (27.97%)	125 (6.61%)
Lower	neu	375 (19.83%)	1028 (54.36%)	1360 (71.92%)	375 (19.83%)	1360 (71.92%)
Lower	mon	289 (15.28%)	269 (14.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Upper	def	287 (15.18%)	124 (6.56%)	242 (12.80%)	529 (27.97%)	372 (19.67%)
Upper	cor	987 (52.19%)	507 (26.81%)	289 (15.28%)	987 (52.19%)	159 (8.41%)
Upper	neu	375 (19.83%)	1028 (54.36%)	1360 (71.92%)	375 (19.83%)	1360 (71.92%)
Upper	mon	242 (12.80%)	232 (12.27%)	0 (0.00%)	0 (0.00%)	0 (0.00%)

Table A.7: Frequency of each bound branch across all individuals ($n = 1891$) for Experiments 1–4 (fully observed Z) and Experiment 5 (Experiment 3 with missing Z). Each cell reports the count of individuals and, in parentheses, the percentage; each individual activates exactly one lower and one upper branch.

environment to be differently distributed than the training environment, because the human can choose to implement a different action from the machine's suggestion. We overcome this challenge by making assumptions about how the human will act, and the RCT data that we learn from is unconfounded.

Learning to Defer In our work, we mention deferral as a possible example of a neutral outcome. Our setting is closely related to learning to defer, in which an AI classifier may abstain on a given input and pass the decision to a human expert, with the goal of training the classifier to complement rather than duplicate the human [Verma and Nalisnick, 2022, Mozannar et al., 2023]. Like our work, L2D recognizes that the human decision-maker, not the model, is the final arbiter of the outcome, and that the two should be modeled jointly. However, L2D differs from our setting in several respects. First, L2D is concerned more with specifically how to train models, while our work focuses on the evaluation of already trained models. Second, L2D models an explicit, learnable rejector that routes each input to either the model or the human, while in our setting the human's response to a model prediction is unobserved and mediated by their trust in the model's perceived performance. Finally, L2D typically assumes the human prediction is available on deferred examples, whereas our framework must reason about counterfactual outcomes that are never directly observed.

G MISSING DATA

In the main text, we assume that Z is observed for all subjects, and derive our results in that context. In many practical settings, however, Z may be missing for some subjects, due to limitations of data collection and/or due to being defined in counterfactual terms (e.g., in our motivating example for the semi-synthetic experiments, Z is a counterfactual FTA indicator that is unobserved when bail is granted).

In this section, we discuss an example missingness structure under which our method remains valid. We limit our discussion to a simple missingness mechanism, to demonstrate how our method can hold under missingness. There may exist more interesting and unique structures in which our method remains valid. However, we leave the exploration of these structures to future work.

Let $R \in \{0, 1\}$ be an observation indicator with $R = 1$ when Z is observed and $R = 0$ otherwise, so the analyst observes $(X, \Pi, \hat{Z}, M, Y, R, Z)$ rather than fully observing Z.

Suppose we have the augmented DAG in Fig. A.7, which is a simple extension of the DAG in Fig. 1a to include the missingness indicator R, where:

$$R = f_R(X, \epsilon_R), \qquad \epsilon_R \perp (\epsilon_{XZ}, \epsilon_Y, \dots).$$

The analyst observes $(X, \Pi, \hat{Z}, M, Y, R)$ and observes Z only when $R = 1$.

We define augmented versions of our bounds, $L'(\pi_e)$ and $U'(\pi_e)$, where $L'(\pi_e) \leq L(\pi_e)$ and $U'(\pi_e) \geq U(\pi_e)$. In these augmented versions, we modify the definitions of $\mathbf{1}\{b_L = \mathsf{cor}\}, \mathbf{1}\{b_U = \mathsf{cor}\}, \mathbf{1}\{b_L = \mathsf{mon}\}, \mathbf{1}\{b_U = \mathsf{mon}\}$ to additionally require $R = 1$, so that these branches are only activated when Z is observed. Thus, for individuals with $R = 0$, the only branches that can activate are $\mathbf{1}\{b_L = \mathsf{def}\}, \mathbf{1}\{b_U = \mathsf{def}\}, \mathbf{1}\{b_L = \mathsf{neu}\}, \mathbf{1}\{b_U = \mathsf{neu}\}$. The branches $\mathbf{1}\{b_L = \mathsf{def}\}, \mathbf{1}\{b_U = \mathsf{def}\}$ take the constant values $Y_{\min}, Y_{\max}$; under Assump. 3.4 they are always valid, and can only loosen the inequality, never violate it, regardless of the missingness mechanism. The branches $\mathbf{1}\{b_L = \mathsf{neu}\}, \mathbf{1}\{b_U = \mathsf{neu}\}$,

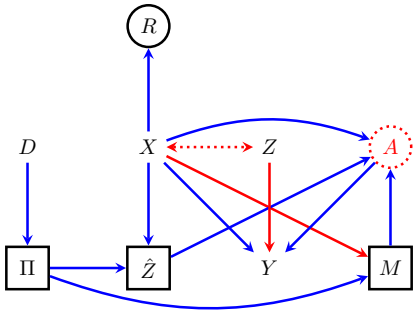


Figure A.7: Augmented version of Fig. 1a including missingness. Node styles and edge colors follow Fig. 1a.

however, require estimation of a nuisance function:

$$\hat{h}(X, Y, Z, \mathcal{S}) = \widehat{\mu}_{\mathcal{S}}(X, Z) + \frac{\mathbf{1}\{\Pi \in \mathcal{S}\}}{P(\Pi \in \mathcal{S} \mid X)}\big(Y - \widehat{\mu}_{\mathcal{S}}(X, Z)\big),$$

which references the individual's Z through $\widehat{\mu}_{\mathcal{S}}(X, Z)$ and is therefore not directly computable when Z is missing.

To resolve this, we consider the neutral contribution to the population bound, noting that the neutral selector is a function of X only. Let

- $b(X) := \mathbf{1}\{b_L = \mathsf{neu}\}(X)$ — a function of X only,
- $g(X, Z) := \mathbb{E}[Y \mid X, \Pi \in \mathcal{S}, Z]$ — a function of (X, Z).

Thus, the neutral contribution to the lower bound is $\mathbb{E}\big[b(X)\, g(X, Z)\big]$, where the expectation is taken over the joint distribution of (X, Z). We consider only the lower bound, but the upper bound is identical. Because $b(X)$ depends on X alone, the law of iterated expectations gives

$$\begin{aligned}\mathbb{E}\big[b(X)\, g(X, Z)\big] &= \mathbb{E}\big[\,\mathbb{E}[b(X)\, g(X, Z) \mid X]\,\big] \\ &= \mathbb{E}\big[\,b(X)\,\mathbb{E}[g(X, Z) \mid X]\,\big].\end{aligned}$$

From the randomization we have $\Pi \perp Z \mid X$, so the inner term satisfies

$$\begin{aligned}\mathbb{E}[g(X, Z) \mid X] &= \mathbb{E}\big[\mathbb{E}[Y \mid X, \Pi \in \mathcal{S}, Z] \mid X\big] \\ &= \mathbb{E}\big[\mathbb{E}[Y \mid X, \Pi \in \mathcal{S}, Z] \mid X, \Pi \in \mathcal{S}\big] \\ &= \mathbb{E}[Y \mid X, \Pi \in \mathcal{S}].\end{aligned}$$

We then define $\nu(X) := \mathbb{E}[Y \mid X, \Pi \in \mathcal{S}]$. Since $\nu(X)$ depends only on X, it can be estimated regardless of whether Z is observed, and we estimate $\widehat{\nu}(X)$ in place of $\widehat{\mu}_{\mathcal{S}}(X, Z)$. Thus the neutral term of our estimator becomes

$$\hat{h}_{\text{marg}}(X, Y, \mathcal{S}) = \widehat{\nu}(X) + \frac{\mathbf{1}\{\Pi \in \mathcal{S}\}}{P(\Pi \in \mathcal{S} \mid X)}\big(Y - \widehat{\nu}(X)\big).$$

We use that estimator for all individuals. For the other branches on the individuals with $R = 1$, we must show that the nuisance functions in Theorem 5.1 do not become inconsistent under missingness. To estimate our bounds, we require nuisance functions of the form $\mathbb{E}[Y \mid X, \Pi = \pi_j, Z]$. This cannot be estimated directly, since Z is missing for $R = 0$ subjects; instead, we can only estimate

$$\mathbb{E}[Y \mid X, \Pi = \pi_j, Z, R = 1].$$

Because in Fig. A.7 we have $R = f_R(X, \epsilon_R)$ with ϵ_R independent of the rest, it follows that

$$R \perp (Z, Y) \mid X \;\Rightarrow\; R \perp Y \mid X, \Pi, Z,$$

and therefore

$$\mathbb{E}[Y \mid X, \Pi = \pi_j, Z, R = 1] \;=\; \mathbb{E}[Y \mid X, \Pi = \pi_j, Z] \;=\; \mu_j(X, Z).$$

For the fitted $\widehat{\mu}_j$ to be consistent, and for the margin and rate conditions of Theorem 5.1 to carry over, the observed-Z subsample must be representative of the population X-support. This requires the following assumption:

Assumption G.1 (Missingness Positivity)**.** There exists $\eta > 0$ such that $\pi_R(x) := P(R = 1 \mid X = x) \geq \eta$ for all $x \in \mathcal{X}$.

This prevents missingness structures in which Z is never observed for some values of X. For example, in a diagnostic context, if Z were never observed for patients above age 80, the assumption would be violated. Under these conditions, the nuisance functions estimated on the $R = 1$ subset converge to the original nuisance functions, and our estimator remains consistent. The model-assignment propensity $P(\Pi \mid X)$ is fixed by RCT design and does not involve R, so Assump. 5.1 is unaffected; under Assump. G.1 the observed-Z subsample shares the population X-support, so the margin conditions and nuisance rates transfer, with the observed fraction affecting only the constant in the rate. Therefore, under this missingness structure and Assump. G.1, $\widehat{L}'(\pi_e)$ and $\widehat{U}'(\pi_e)$ are consistent and asymptotically normal (Theorem 5.1) for the valid bounds $L'(\pi_e) \leq \mathbb{E}[Y(\pi_e)] \leq U'(\pi_e)$.

G.1 RELATION TO EXPERIMENT 5

In Experiment 5, these conditions are clearly not satisfied. The missingness of Z depends directly on the value of A. In this experiment, Z is defined as the counterfactual outcome under no judge intervention, when it is available. When $A = 1$, the judge gives bail to the defendant, and Z is never observed. Conversely, when $A = 0$, the judge does not give bail to the defendant, and Z is always observed.

This clearly does not match the missingness structure described in Fig. A.7 and violates Assump. G.1. Regardless, as mentioned, we run it for demonstration purposes.

G.2 SPECIFIC VIOLATED ASSUMPTIONS

We also discuss how our method can handle violations of specific assumptions posited in our method. Since each branch corresponds to a specific assumption, if any of the assumptions is violated, the corresponding indicator and its branch can be removed from the bounds. This would cause any values of x that would have activated the branch to activate $b_L/b_U = \text{def}$ instead, resulting in a valid albeit looser bound. For example, if we are in a diagnostic setting, and there is no adequate neutral action (model cannot defer to physician), then we have violated Assump. 3.3, but we can run the method as is, since the neutral branches ($b_U = \text{neu}$, $b_L = \text{neu}$) will never activate regardless, and the bounds will be constructed using the other assumptions. Under violations of Assump. 3.1 or Assump. 3.2, the corresponding branches (1 or 2, respectively) can instead default to $Y_{\min}, Y_{\max}$ bounds, maintaining validity, which trivially follows from Assump. 3.4.

H FLOWCHART

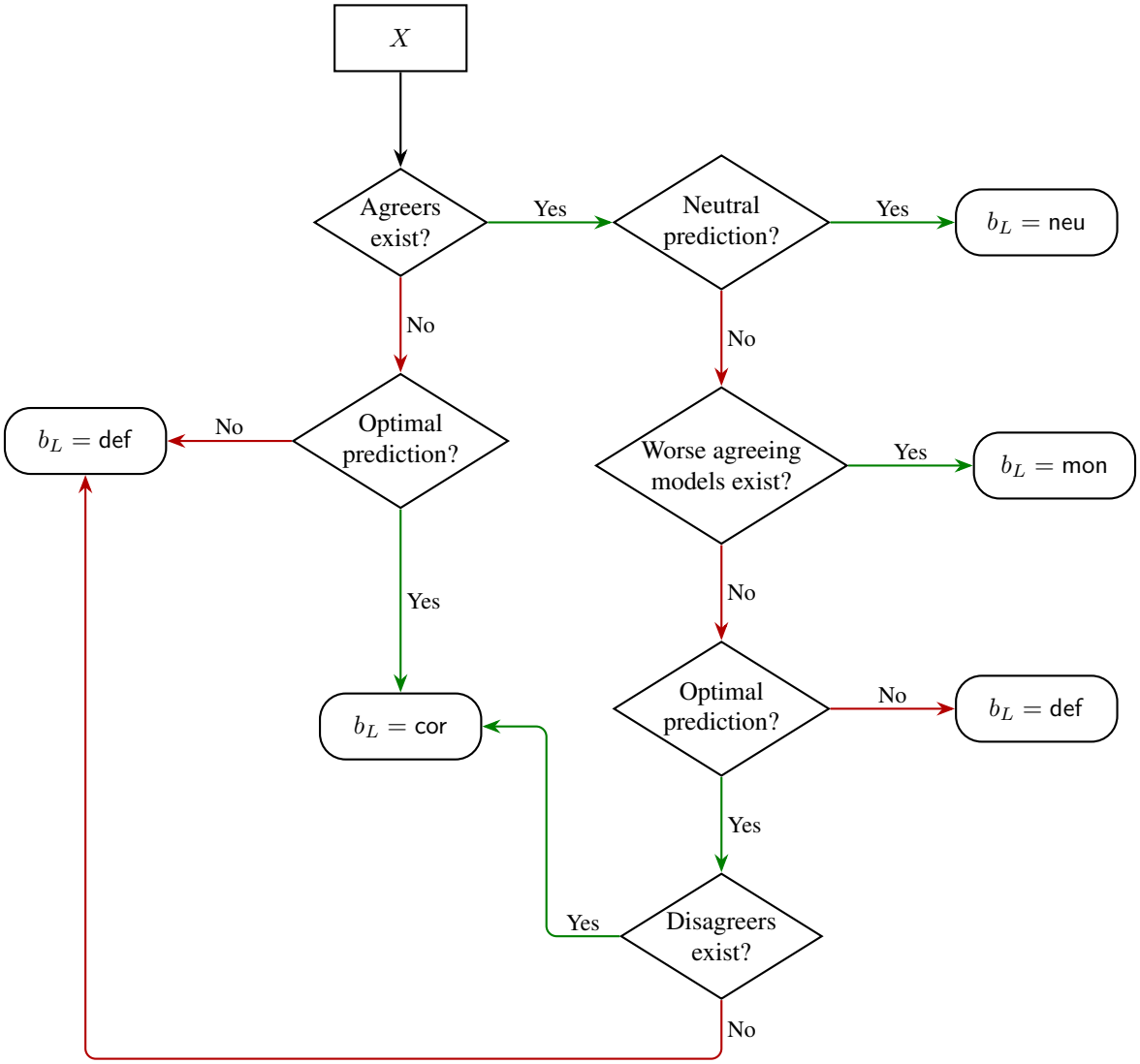


Figure A.8: Flowchart with logic equivalent to that of $L(\pi_e)$