Just Trial Once: Ongoing Causal Validation of Machine Learning Models

Jacob M. Chen[1] **Michael Oberst**[1]

[1]Department of Computer Science, Johns Hopkins University

Abstract

Machine learning (ML) models are increasingly used as decision-support tools in high-risk domains. Evaluating the causal impact of deploying such models can be done with a randomized controlled trial (RCT) that randomizes users to ML vs. control groups and assesses the effect on relevant outcomes. However, ML models are inevitably updated over time, and we often lack evidence for the causal impact of these updates. While the causal effect could be repeatedly validated with ongoing RCTs, such experiments are expensive and time-consuming to run. In this work, we present an alternative solution: using only data from a prior RCT, we give conditions under which the causal impact of a new ML model can be precisely bounded or estimated, even if it was not included in the RCT. Our assumptions incorporate two realistic constraints: ML predictions are often deterministic, and their impacts depend on user trust in the model. Based on our analysis, we give recommendations for trial designs that maximize our ability to assess future versions of an ML model. Our hope is that our trial design recommendations will save practitioners time and resources while allowing for quicker deployments of updates to ML models.

1 INTRODUCTION

Machine learning (ML) models are increasingly deployed in high-risk domains like healthcare and criminal justice as tools to support human decision-makers. For instance, in healthcare, ML-powered decision-support tools (ML-DSTs) are widespread, including early warning systems for sepsis [Adams et al., 2022, Sendak et al., 2020, Boussina et al., 2024], computer-assisted decision-support for antibiotic treatment decisions [Gohil et al., 2024a,b], and a variety of tools for computer-aided diagnostics in radiology and pathology, with the FDA having cleared or approved over 1,000 AI/ML-enabled devices to date [FDA, 2024]. Although these models often exhibit high accuracy, it is not always clear whether their deployment actually leads to better decisions, and thus, better downstream outcomes. In healthcare, for instance, we are interested not only in model accuracy, but also whether deployment of an ML-DST improves health outcomes for patients.

The gold standard evaluation of ML-enabled decision-support is to assess impact in a randomized controlled trial (RCT), typically structured as a cluster RCT, where decision-makers (e.g., clinicians in a given hospital) are randomized to an ML-DST or no ML-DST. Such trials are becoming more common in healthcare [Han et al., 2024] and criminal justice [Imai et al., 2023]. Examples include recent "failed trials" like the PROTEUS trial of ML-assisted diagnosis of stress echocardiography [Upton et al., 2024] and trials with more positive results, such as the INSPIRE trials for antibiotic recommendations powered by ML predictions of resistance likelihood [Gohil et al., 2024a,b]. These trials provide rigorous evidence for the impact of deploying specific ML-enabled systems (and their underlying models), and the broader research community recognizes the need for more randomized trials [Ouyang and Hogan, 2024] and evaluation of ML systems as interventions [Joshi et al., 2025].

However, the traditional RCT framework is not designed for ML-enabled systems, which (unlike drugs) are often updated frequently to handle performance degradation. Even when RCT data is available for a single version of an ML-DST, it is not obvious whether those results apply to later models, and running additional RCTs to verify continued effectiveness is both time-consuming and costly.

Our work addresses this challenge from a methodological perspective, as illustrated in Fig. 1: We formalize conditions under which data from an existing RCT can be used to precisely infer or bound the causal impact of deploying models that were not included in the original RCT. We take

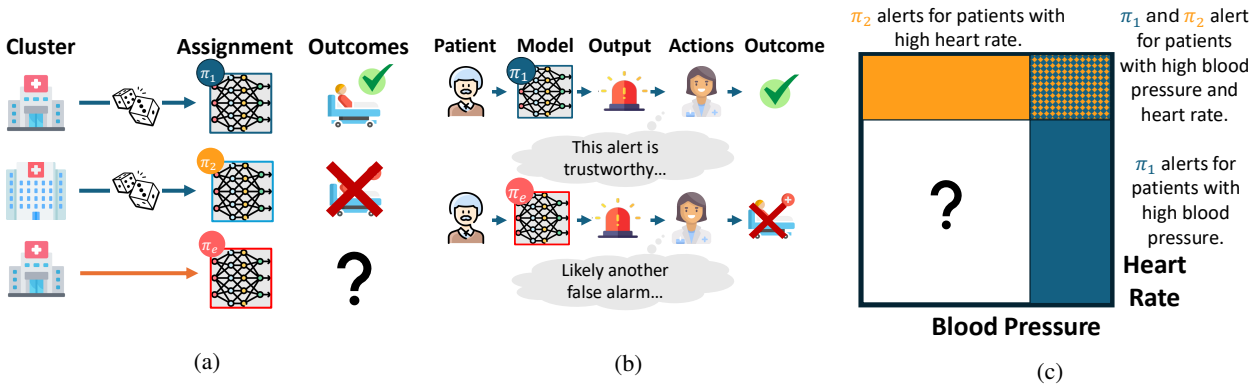

Figure 1: (a) The goal of this paper is to predict the causal impact of deploying a new model π_e, given data from a cluster randomized trial that randomizes sets of users (e.g., hospitals) to one of a set of trial models that does not include π_e. (b) The first challenge: Relevant outcomes (e.g., of patients) are not only influenced by model outputs, but also by how users actually respond to the outputs of model-based decision-support, which may itself be affected by the perceived reliability / performance of the model. (c) The second challenge: There may exist some subset of cases (e.g., patients) for whom we never observe certain model outputs, making it impossible to give precise predictions for outcomes of patients in that group. In this example, π_1 alerts for patients with high blood pressure, and π_2 alerts for patients with high heart rate; thus, patients with high blood pressure and high heart rate receive an alert from both π_1 and π_2. However, patients in the white region – those who have both low to medium blood pressure and heart rate – never receive an alert from either trialed model.

into account two important practical considerations: First, model performance (e.g., accuracy at diagnosing disease) will influence user trust in the system, and thereby indirectly influence outcomes (Fig. 1b). Second, while the deployment of DSTs is often randomized, the predictions themselves are not typically randomized (Fig. 1c), since doing so would undermine trust (e.g., by raising alerts randomly). Hence, there may be some combinations of model outputs (e.g., diagnoses) and inputs (e.g., patients) that we never observe.

Under limited assumptions that incorporate these considerations, we derive bounds on the causal impact of deploying a new model. Crucially, we show that both of our main assumptions can be checked using RCT data that includes at least two models with differing performance characteristics. In a simulation study, we show how our framework yields more rigorous conclusions about the value of model updates, as compared to naive approaches that only judge models based on their raw performance.

Our results have practical implications for post-trial analysis and pre-trial design. First, evaluating new models using historical trial data is possible under reasonably limited assumptions, but not all alternative models can be precisely evaluated in this way. Second, our results suggest a benefit to running RCTs with multiple ML models to maximize the ability to estimate causal impacts in future model updates.

To summarize, our contributions are as follows:

- We provide assumptions (Assumptions 2.1 and 3.1 to 3.3) under which we derive bounds (Theorem 3.1) on the effect of deploying a new ML model, given data from a prior RCT. Our bounds are tight, and cannot be improved without further assumptions (Theorem 3.2).
- We provide a simple estimator for these bounds and a procedure for generating asymptotically valid confidence intervals (Proposition 3.4). We also show that our core assumptions can be falsified via hypothesis tests constructed from RCT data trialing multiple models (Propositions 3.1 and 3.2).
- We provide recommendations for pre-trial design and post-trial analysis in light of our results (Section 4), and demonstrate in a simulation study (Section 5) that our bounds provide a more informative tool to select among model updates as compared to using the raw performance (e.g., accuracy) of updated models.

Related literature: Our work is related to *off-policy policy evaluation* in causal inference and reinforcement learning [Uehara et al., 2022]. An ML-DST can be viewed as a deterministic policy that chooses actions (i.e., predictions or alerts to raise) based on context (i.e., inputs to the model) with the goal of obtaining some reward (i.e., positively influencing outcomes for patients). Two critical distinctions arise in our work versus the standard setting: First, the policies that are present in retrospective data (in our case, from a trial) are deterministic rather than random, leading to violation of the common assumption that, for a given context, there is a positive probability of seeing any action. Second, we allow for the fact that actions taken for one patient can influence outcomes for other patients.

Our work is also related to causal evaluations of AI-assisted

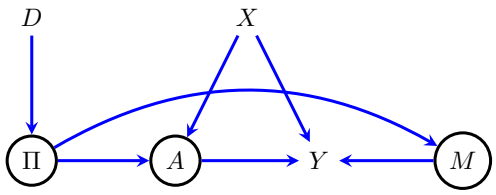

Figure 2: The directed acyclic graph (DAG) $\mathcal{G}$ depicting the causal relationships in our problem setup (Assumption 2.1). We draw circles around the nodes Π, A, and M to represent that these variables are deterministic given their parents (the nodes that have a direct edge to them).

decision-making in criminal justice settings [Imai et al., 2023, Ben-Michael et al., 2024, 2025], but our goal differs: Rather than evaluating the impact of AI-assistance on the accuracy of (observable) predictions made by a human, we are interested in the total effect of model deployment on downstream outcomes. Finally, our work is connected to the study of causal transportability where the goal is to infer the effect of a known intervention from an RCT onto a new target population where randomization is difficult, expensive, or impossible [Pearl and Bareinboim, 2011, Stuart et al., 2011]. This is similar to our setting as we are also attempting to draw inference from RCT data. However, our problem differs in that we would like to infer the effect of an unseen intervention (i.e., a new model with no historical trial data) on the same population as in the original RCT.

2 MODEL AND PROBLEM SETUP

Notation: In the rest of this paper, we use the terms *model* and *policy* interchangeably. We use upper case letters X to denote a random variable, calligraphic font $\mathcal{X}$ to denote the space of possible values, and lower-case letters x to denote a specific realization of a random variable. We assume that the causal structure of an RCT is modeled by a directed acyclic graph (DAG) $\mathcal{G}$ over a set of vertices $\mathbf{V} = \{A, Y, D, X, \Pi, M\}$, where $A \in \mathcal{A}$ represents the output (or "action") of the deployed model, $Y \in \mathbb{R}$ represents an outcome of interest, $D \in \mathcal{D}$ represents the cluster to which a user is assigned, $X \in \mathcal{X}$ represents covariates used as inputs to the ML model, $\Pi \in \mathit{\Pi}$ represents the specific ML model that was deployed, and $M \in \mathbb{R}$ represents model performance, which we represent as a real number. We assume that model performance is computable for any model via some functional $f_M(\pi)$ (e.g., the accuracy, precision, recall, sensitivity, specificity, or some combination, computed on a held-out dataset where $\pi(X)$ is considered the model prediction). We also use the indicator function $\mathbf{1}\{S\}$ that is equal to 1 if the event S is true, and 0 otherwise.

Example 1 (Alerting Systems)**.** Suppose we are interested in the effect of deploying a DST that monitors patient vital signs and selectively raises an "alert". A common application of these systems is detecting the onset of sepsis and alerting clinicians to facilitate timely intervention [Adams et al., 2022, Sendak et al., 2020, Boussina et al., 2024]. Here, the inputs X to the model are typically vital signs, the outcome Y may be long-term patient survival, and the outputs $\mathcal{A}$ include raising an alert ($A = 1$) or not ($A = 0$). The variable M in this setting could correspond to the false alarm rate of the alerting policy Π when it comes to predicting the onset of disease within the next hour. Note that the label used for computing performance here (onset of disease) differs from the patient outcome of interest Y (survival). A control arm of "no assistance" can be represented as a deterministic policy that never raises an alert.

Example 2 (Computer Assisted Diagnosis)**.** Suppose we are interested in the effect of deploying a diagnostic model that assists with screening for some disease. Here, the outcome of interest Y may be long-term patient survival, X would include inputs to the model (e.g., medical imaging, past medical history), and the set of actions $\mathcal{A}$ could include a set of K possible diagnostic labels as well as the option of deferring to a human expert, such that $\mathcal{A} = \{\varnothing, 1, \ldots, K\}$, where $\varnothing$ denotes deferral. The variable M in this setting could represent the overall accuracy of the diagnostic model at predicting some true diagnostic label or some combination of its sensitivity and specificity when it does not defer. In a randomized trial where the control arm consists of "no assistance", the resulting "policy" in the control arm could be viewed as a deterministic policy that always defers.

For concreteness in the remainder of this paper, we will primarily use the language of healthcare applications (e.g., patients, likelihood of disease onset, clinical outcomes, etc). Our assumed causal structure can be represented by the structural causal model (SCM) [Pearl, 2009] that we define below, which is consistent with the DAG shown in Fig. 2.

Assumption 2.1 (Data Generating Process)**.** *The random variables $D \in \mathcal{D}$, $\Pi \in \mathit{\Pi}$, $X \in \mathcal{X}$, $A \in \mathcal{A}$, and $Y \in \mathbb{R}$ are generated according to the SCM*

$$\begin{aligned} D &= f_D(\epsilon_D), & X &= f_X(\epsilon_X), \\ \Pi &= \pi_D, & M &= f_M(\Pi), \\ A &= \Pi(X) & Y &= f_Y(A, X, M, \epsilon_Y), \end{aligned}$$

where ϵ_Y, ϵ_D, and ϵ_X are mutually independent.

We make a few notes regarding Assumption 2.1. First, the randomization into a specific policy (signified by D) is independent of covariates X. Second, the policy Π is entirely determined by D, model performance M is entirely determined by Π (and observable), and the output A is a deterministic function of X, based on Π. This deterministic nature of model outputs can create difficulties in evaluating new models; in particular, we are unlikely to see all possible outputs $a \in \mathcal{A}$ for all types of patients X. Finally, we assume that outcomes Y are not only a function of covariates

X and the model output A, but also the performance M of the model[1]. Note that we assume that M and A are sufficient to capture the impact of a deployed model on outcomes.

3 IDENTIFICATION AND BOUNDS

Goal: We adopt potential outcomes notation [Richardson and Robins, 2013] where we use $Y(A = a, M = m) := f_Y(a, X, m, \epsilon_Y)$ to denote counterfactual outcomes, representing the value of Y that would be observed if we had taken action $A = a$ with a model whose performance is given by $M = m$[2]. Using this notation, our goal is to infer expected outcomes if we had deployed a new model / policy π_e not trialed in the original RCT, i.e.

$$\mathbb{E}[Y(\pi_e)] = \mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e))]^3 \quad (1)$$

We refer to $\mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e))]$ as our *target estimand* or *policy value*. Once this value is inferred, one could compute the causal effect of deploying π_e as opposed to any other trialed model π_i by evaluating $\mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e))] - \mathbb{E}[Y(A = \pi_i, M = f_M(\pi_i))]$.

Example 1 (continued)**.** Suppose that the trialed model alerts based on thresholding a pre-defined risk score $r(x)$ that is a function of vital signs (e.g., systolic blood pressure, respiratory rate, etc). Suppose that an initial RCT assigns patients to a control arm, $D = 0$ where $\pi_0 = 0$ (alerts are never raised), and a treatment arm, $D = 1$ where the model raises alerts using the threshold T^*, i.e., $\pi_1(x) := \mathbf{1}\{r(x) > T^*\}$. Suppose we want to use this RCT data to evaluate the impact of an alternative model with a lower threshold, $\pi_l(x) := \mathbf{1}\{r(x) > T^l\}$ where $T^l < T^*$. Fig. 3 visually demonstrates the challenges of this inference task for π_l as we never observe alerts for patients with $r(x) \in [T^l, T^*]$.

In order to estimate the policy value in Eq. (1), we introduce a few key assumptions that relate outcomes under different hypothetical models / policies. First, since it is unlikely that our target policy π_e has exactly the same performance M as policies trialed during the RCT, we need to assume a relationship between outcomes and model performance.

Assumption 3.1 (Performance Monotonicity)**.** *Potential outcomes are non-decreasing in model performance, i.e., if $m_i < m_j$, then for all $a \in \mathcal{A}$,*

$$Y(A = a, M = m_i) \leq Y(A = a, M = m_j)$$

Assumption 3.1 says that improvements in model performance do not harm patient outcomes, given a fixed action.

[1]We discuss defining M for the control arm in Section 3.

[2]We defer a more detailed discussion of potential outcomes and other causal inference background to Appendix A.

[3]In the rest of this paper, we use π_e as shorthand for $\pi_e(X)$.

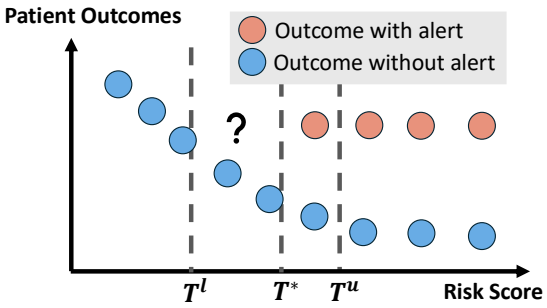

Figure 3: Illustration from Example 1, demonstrating the challenge of our task. We can use data from both the control arm, that does not raise alerts, and the treatment arm, that raises alerts for patients with risk scores greater than T^*, to infer patient outcomes when π_l does not raise alerts for patients with risk scores less than T^l. Next, we can use data from π_1 to infer patient outcomes when π_l raises alerts for patients with risk scores greater than T^*. However, we do not know what patient outcomes are when π_l raises alerts for patients with risk scores between T^l and T^*.

For instance, in the context of Example 1, we might expect that, for a given patient, having an alarm raised by a high-performance model would not lead to worse outcomes than if that alert had been raised by a model with frequent false alarms. Note that this assumption is stated with a fixed action $A = a$ and does *not* imply that improving performance alone is guaranteed to improve outcomes – a change in model performance is generally associated with a change in outputs, which may have its own effect on outcomes. In Section 5, we give a case where improved overall performance (accuracy) is associated with worse outcomes. Assumption 3.1, however, may not always hold; for instance, clinicians may begin paying less attention to patients that receive a low risk score from the DST, even if it is wrong, as their trust in the system increases. Such a scenario would violate Assumption 3.1. To address this, we propose a method for falsifying Assumption 3.1 below.

Proposition 3.1 (Falsification of Assumption 3.1)**.** *Let $\mathcal{X}$ denote the full space of possible covariate values. Under Assumption 2.1, given data from an RCT that includes at least two trialed models π_1 and π_2 with different levels of performance $f_M(\pi_1) < f_M(\pi_2)$, and whose actions agree on a non-empty set of patients $\mathcal{X}_{agree} := \{x \in \mathcal{X} \mid \pi_1(x) = \pi_2(x)\}$ such that $P(X \in \mathcal{X}_{agree}) > 0$, the observation that*

$$\mathbb{E}[Y \mid X \in \mathcal{X}_{agree}, \Pi = \pi_2] < \mathbb{E}[Y \mid X \in \mathcal{X}_{agree}, \Pi = \pi_1],$$

implies that Assumption 3.1 is false.

The proof for Proposition 3.1, along with all other proofs, is given in Appendix D. While it is not possible to guarantee that Assumption 3.1 is true in general (over all possible models), it has observable implications in an RCT that we

can check. In particular, this result suggests a simple hypothesis test that we can use to falsify Assumption 3.1: compare two empirical means in the data and check if outcomes are lower under π_2 than under π_1 on those cases where π_1 and π_2 agree on their actions.

In our discussions thus far, we have considered the control arm to be just another policy. However, this framework creates practical difficulties when considering the model performance of a control arm. For instance, suppose the relevant metric for model performance is the false positivity rate (as in Example 1); then, M is not clearly defined because the control arm never raises alerts. Alternatively, if the relevant metric were model accuracy under no deferral (as in Example 2), then the performance of the control arm would be similarly undefined. One way of resolving this tension is to presume the existence of a "neutral" action (e.g., not raising an alert, or deferring to clinicians).

Assumption 3.2 (Neutral Actions)**.** *There exists a "neutral action" $a_0 \in \mathcal{A}$ such that the potential outcome of Y under a_0 does not depend on model performance M. That is, for any two values m_1, m_2, including when $m_1 \neq m_2$,*

$$Y(A = a_0, M = m_1) = Y(A = a_0, M = m_2), \qquad (2)$$

and in these cases we use the shorthand $Y(A = a_0)$ to reflect the fact that the outcome does not depend on M.

Assumption 3.2 is a sufficient condition for leveraging data from the control arm of an RCT in our setting, and it implies that, when a model output is "neutral" (e.g., no alert in Example 1, or deferral in Example 2), decision-makers act as they would if no model were deployed. However, note that there may not always exist a "neutral" output, especially if decision-makers tend to pay attention to model performance for all possible model outputs[4]. Again, we propose a method for falsifying Assumption 3.2 below.

Proposition 3.2 (Falsification of Assumption 3.2)**.** *Under Assumption 2.1, given data from an RCT that includes at least two trialed models π_1 and π_2 with different levels of performance $f_M(\pi_1) < f_M(\pi_2)$, and which both models take the neutral action a_0 on a non-empty set of patients $\mathcal{X}_{a_0} := \{x \in \mathcal{X} \mid \pi_1(x) = \pi_2(x) = a_0\}$ such that $P(X \in \mathcal{X}_{a_0}) > 0$, the observation that*

$$\mathbb{E}[Y \mid X \in \mathcal{X}_{a_0}, \Pi = \pi_2] \neq \mathbb{E}[Y \mid X \in \mathcal{X}_{a_0}, \Pi = \pi_1],$$

implies that Assumption 3.2 is false.

Similar to Proposition 3.1, Proposition 3.2 suggests a simple hypothesis test that can be used to falsify Assumption 3.2: compare two empirical means in the data to test if outcomes under π_1 and π_2 are significantly different on the cases where they both choose a_0 as the model output. Of particular interest is the scenario where a control arm exists, and we are interested in checking whether outcomes under the control arm (e.g., not alerting in Example 1) coincide with outcomes in a treatment arm where the trialed model agrees with the control arm (e.g., does not raise alerts in Example 1).

Example 1 (continued)**.** Consider an evaluation policy $\pi_u(x) = \mathbf{1}\{r(x) > T^u\}$ where the threshold for alerting $T^u > T^*$ is greater than the threshold used in the original trial and where the performance of $\pi_u(x)$ (e.g., the precision) is greater than that of the original trialed policy π_1. Fig. 3 gives a visual illustration of such a policy. In this scenario, under Assumptions 3.1 and 3.2, we can intuitively infer a lower bound on the policy value of π_u using the outcomes of both (a) patients with $r(x) \leq T^u$ who did not receive alerts in the trial (either because they were in the control arm or because π_1 did not raise alerts), and (b) those patients with $r(x) > T^u$ who did receive alerts under π_1.

While these assumptions are sufficient in some scenarios, they do not yield meaningful bounds when a new policy takes actions (i.e., a new model produces outputs) on a given case that was never seen for similar cases in the RCT.

Example 1 (continued)**.** Consider the evaluation policy $\pi_l(x) = \mathbf{1}\{r(x) > T^l\}$ where the threshold for alerting $T^l < T^*$ is **less** than the threshold used in the original trial. Regardless of the performance of π_l in this scenario, even under Assumptions 3.1 and 3.2, we have no way to infer outcomes under π_l for the individuals where $r(x) \in [T^l, T^*]$. These correspond to a set of "never alerted" individuals where π_l raises an alert but where neither the control arm nor the trialed policy π_1 raised an alert.

To resolve this fundamental uncertainty, it is sufficient to know that outcomes Y are bounded, such that we can provide some bounds on expected outcomes in the evaluation of policies that take never-before-seen actions.

Assumption 3.3 (Bounded Outcomes)**.** *There exists constants Y_{min}, Y_{max} such that $Y_{min} \leq Y \leq Y_{max}$.*

Aside: Why require the performance assumption? We pause to reflect on the importance of the assumption (implicit in Assumption 2.1) that implies that our choice of model Π impacts outcomes, not only through the outputs A, but also through model performance M. Broadly speaking, this assumption is not only intuitive from a real-world perspective, but it also has the welcome side-effect of ruling out nonsensical conclusions about trial design. For instance, there are trivial ways to satisfy the condition that, for every value of $\mathcal{X}$, there exists some model in the trial that matches the output of π_e. In the context of Example 1, for instance, one could trial an alerting system that simply *always raises*

[4] While all of our results make use of Assumption 3.2, they can also be re-written to hold if Assumption 3.2 is false by re-defining a_0 as some placeholder model output that is never observed under any model (including π_e), such that indicators like $\mathbf{1}\{\pi_e(x) = a_0\}$ are always zero.

alerts for every patient, alongside a control arm that never raises alerts. Then, the requirement that we observe what happens to patients both under no alerts and under alerts for each $x \in \mathcal{X}$ would be satisfied, eliminating any challenges related to coverage. The assumption that model accuracy M impacts outcomes gives a formal rationale for why this type of trial design is nonsensical: The observed impact of this "always alert" policy would likely be minimal, or even harmful, compared to never raising alerts due to the negative impact of extremely poor accuracy.

We will shortly present our main result: Under our data-generating process (Assumption 2.1) and the assumptions above (Assumptions 3.1 to 3.3), we can compute tight bounds on expected outcomes under any proposed model π_e. First, however, we will define some useful notation for conveying our results, which builds upon the intuition above.

Definition 3.1 (Policy/Model Sets)**.** For each value of $x \in \mathcal{X}$, we define the sets of trialed policies/models (possibly none) that agree with $\pi_e(x)$ and subsets of this set based on the performance characteristics of those trialed models[5].

$$\begin{aligned}
\mathbf{\Pi}^e(x) &\coloneqq \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x)\} \\
\mathbf{\Pi}^e_{\leq}(x) &\coloneqq \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x), f_M(\pi) \leq f_M(\pi_e)\} \\
\mathbf{\Pi}^e_{\geq}(x) &\coloneqq \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x), f_M(\pi) \geq f_M(\pi_e)\}
\end{aligned}$$

We also further define subsets of $\mathbf{\Pi}^e_{\leq}$ and $\mathbf{\Pi}^e_{\geq}$ that contain only the next-worst or next-best performing model[6].

$$\begin{aligned}
\tilde{\mathbf{\Pi}}^e_{\leq}(x) &\coloneqq \underset{\pi \in \mathbf{\Pi}^e_{\leq}(x)}{\arg\max}\, f_M(\pi), \\
\tilde{\mathbf{\Pi}}^e_{\geq}(x) &\coloneqq \underset{\pi \in \mathbf{\Pi}^e_{\geq}(x)}{\arg\min}\, f_M(\pi)
\end{aligned}$$

Remark 3.1. To summarize our notation related to deployed and new models, we use $\mathit{\Pi}$ to denote the space of all deployed models in the RCT, Π to refer to the variable representing the deployed model, π to denote a specific deployed model, π_e to denote the new model we are evaluating, and Π in boldface with the superscript e, such as $\mathbf{\Pi}^e(x), \mathbf{\Pi}^e_{\leq}(x), \tilde{\mathbf{\Pi}}^e_{\leq}(x), \mathbf{\Pi}^e_{\geq}(x)$, and $\tilde{\mathbf{\Pi}}^e_{\geq}(x)$, to denote functions that take a value of $x \in \mathcal{X}$ as input and return a set of models satisfying various criteria.

Using Definition 3.1, we can give precise lower and upper bounds on the performance of a new model π_e.

Theorem 3.1. *Given the data generating process in Assumption 2.1, and under Assumptions 3.1 to 3.3, the policy value of a model / policy π_e is bounded as*

$$L(\pi_e) \leq \mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e))] \leq U(\pi_e), \quad (3)$$

where

$$\begin{aligned}
L(\pi_e) = \mathbb{E}\big[&\mathbf{1}\{\pi_e \neq a_0\}\big(\\
&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)] && (4) \\
&+\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}Y_{min}\big) && (5) \\
&+\mathbf{1}\{\pi_e = a_0\}\big(\\
&\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)] && (6) \\
&+\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}Y_{min}\big)\big] && (7) \\
U(\pi_e) = \mathbb{E}\big[&\mathbf{1}\{\pi_e \neq a_0\}\big(\\
&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)] \\
&+\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}Y_{max}\big) \\
&+\mathbf{1}\{\pi_e = a_0\}\big(\\
&\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)] \\
&+\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}Y_{max}\big)\big]
\end{aligned}$$

These bounds are still valid if we replace $\tilde{\mathbf{\Pi}}^e_{\leq}(X)$ with $\mathbf{\Pi}^e_{\leq}(X)$ and $\tilde{\mathbf{\Pi}}^e_{\geq}(X)$ with $\mathbf{\Pi}^e_{\geq}(X)$.

We give intuition for the construction of the lower bound. First, note that each value of $x \in \mathcal{X}$ makes a contribution to the construction of the lower bound based on which green and which purple indicator it activates. In Eq. (4), we consider model outputs that are not the neutral action a_0 and values of X at which $\tilde{\mathbf{\Pi}}^e_{\leq}(X)$ is non-empty. That is, there is at least one trialed model agreeing in output with π_e that also has less than or equal performance. Here, we use outcome data from trial arms with the next-worst performance to infer outcomes. Next, Eq. (5) represents values of X where π_e does not output the neutral action and there are no agreeing trialed models with less than or equal performance. In this case, we use Y_{min} to lower bound outcomes as we have no trial data on outcomes under such model output. In Eq. (6), the new model outputs the neutral action a_0, and there is at least one trial model agreeing in output. We can then use data from any trial models that agreed in output to infer outcomes as model performance does not influence outcomes under the neutral action. Finally, Eq. (7) represents no trial models agreeing in an output of the neutral action; here, we lower bound by Y_{min}. The intuition for the upper bound follows similarly.

The lower and upper bounds in Theorem 3.1 can be constructed by iterating over all possible values of X, determining whether the model output is a neutral action, checking whether there are agreeing trial models with appropriate levels of performance, and taking the weighted average of the appropriate bounds given the observations above over X. In Appendix B, we give an algorithm for constructing the bounds proposed in Theorem 3.1 in this manner.

Theorem 3.2 (Tightness of bounds in Theorem 3.1)**.** *For any observational distribution $P(X, Y, A, M, \Pi, D)$ consistent with the assumptions of Theorem 3.1, there exist two*

[5]All these sets are defined with respect to the model π_e and could be written more precisely with π_e as an argument instead of in the superscript (e.g., $\mathbf{\Pi}(x, \pi_e)$) but we use the superscript notation for conciseness.

[6]Where relevant, we use the convention that $\arg\min_{\pi \in \varnothing}(f_M(\pi)) = \varnothing$.

structural causal models $\mathcal{M}_L, \mathcal{M}_U$ such that both are consistent with Assumptions 2.1 and 3.1 to 3.3, both give rise to that same observational distribution P, and where the policy value of any policy π_e under $\mathcal{M}_L, \mathcal{M}_U$ is given by $L(\pi_e), U(\pi_e)$ from Theorem 3.1, respectively. Hence, these bounds cannot be improved without further assumptions.

Note that Theorem 3.2 implies that the tightest possible bounds require the use of $\tilde{\mathbf{\Pi}}^e_{\leq}(x)$ and $\tilde{\mathbf{\Pi}}^e_{\geq}(x)$ in place of $\mathbf{\Pi}^e_{\leq}(x)$ and $\mathbf{\Pi}^e_{\geq}(x)$, respectively. This requirement arises because we may get tighter lower-bounds (and similarly, tighter upper-bounds) by using only outcomes under the "next-worst/best" performing model, rather than averaging over outcomes under all worse/better-performing models. Nonetheless, it may be useful to use $\mathbf{\Pi}^e_{\leq}(x)$ instead of $\tilde{\mathbf{\Pi}}^e_{\leq}(x)$ in some scenarios due to sample-size concerns, especially if outcomes Y do not vary significantly with M.

When is exact identification possible? To better understand conditions for agreement of upper and lower bounds, we directly consider the width of these bounds.

Proposition 3.3 (Bound Decomposition)**.** *The gap between the bounds in Theorem 3.1 can be written as*

$$U(\pi_e) - L(\pi_e) = \mathbb{E}[\delta(X, Y, \Pi)]$$

where

$$\begin{aligned}
&\delta(X, Y, \Pi) = \\
&\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}(Y_{max} - Y_{min}) && (8)\\
&+\mathbf{1}\{\pi_e \neq a_0\}\big[\\
&\quad \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\cdot && (9)\\
&\quad (\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)] - \mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)]) \\
&+\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}\cdot && (10)\\
&\quad (Y_{max} - \mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)]) \\
&+\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\cdot && (11)\\
&\quad (\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)] - Y_{min})\big]
\end{aligned}$$

Moreover, $\delta(X, Y, \Pi) \geq 0$ almost surely under the assumptions of Theorem 3.1.

Note that the only way to achieve point identification (a gap of zero in Proposition 3.3) is if each component (Eqs. (8) to (11)) is equal to zero. Here, a value of $x \in \mathcal{X}$ only contributes to the bound decomposition if x satisfies a blue indicator or both a green and purple indicator. Eq. (8) captures uncertainty that arises in the subset of the population, indexed by $x \in \mathcal{X}$, for which $\{\mathbf{\Pi}^e(x) = \varnothing\}$ where *no trialed model agrees with the output of the model* π_e regardless of whether the new model outputs a_0. For this term to be zero, there must exist some trialed model that agrees with the action taken by π_e for every value of $x \in \mathcal{X}$. Second, Eq. (9) reflects differences in the bounds for the subpopulation where the evaluation policy π_e takes a non-neutral action ($\pi_e(X) \neq a_0$), and the trialed models that agree with π_e have at least one of equal performance[7] or include both better-performing and worse-performing models. Here, the outcomes under the better/worse-performing models give an upper/lower bound on the outcomes under π_e. For this term to be zero, either there exists an agreeing trial model with equal performance, this subpopulation $\{\tilde{\mathbf{\Pi}}^e_{\leq}(x) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(x) \neq \varnothing\}$ is empty, or the outcomes under the better and worse-performing models coincide[8]. Equations (10) and (11) capture subpopulations where the only models that agree with π_e are either worse-performing (giving a lower bound, but no meaningful upper bound) or better-performing (giving an upper bound, but no meaningful lower bound) when the new model does not output a_0, and these sets must be empty for these terms to be zero.

How can we estimate these bounds from data? Proposition 3.4 below implies a simple estimator that can be used to estimate bounds (and provide asymptotically valid confidence intervals) on the policy value for a new policy π_e without the need for training auxiliary models.

Proposition 3.4. *The bounds in Theorem 3.1 can be written $L(\pi_e) = \mathbb{E}[\psi_L(Y, X, \Pi)]$ and $U(\pi_e) = \mathbb{E}[\psi_U(Y, X, \Pi)]$, where ψ_L and ψ_U are defined as follows*

$$\psi_L(Y, X, \Pi) := \begin{cases} Y \cdot \frac{\mathbf{1}\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)\}}{P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X))}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0 \\ Y_{min}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0 \\ Y \cdot \frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}^e(X)\}}{P(\Pi \in \mathbf{\Pi}^e(X))}, & \text{if } \mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0 \\ Y_{min}, & \text{if } \mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0 \end{cases}$$

$$\psi_U(Y, X, \Pi) := \begin{cases} Y \cdot \frac{\mathbf{1}\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)\}}{P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X))}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \pi_e(X) \neq a_0 \\ Y_{max}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \pi_e(X) \neq a_0 \\ Y \cdot \frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}^e(X)\}}{P(\Pi \in \mathbf{\Pi}^e(X))}, & \text{if } \mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0 \\ Y_{max}, & \text{if } \mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0 \end{cases}$$

Moreover, since ψ_L, ψ_U are known functions of the data, these bounds can be estimated as

$$\hat{L}(\pi_e) := n^{-1} \sum_i \psi_L(Y_i, X_i, \Pi_i)$$

$$\hat{U}(\pi_e) := n^{-1} \sum_i \psi_U(Y_i, X_i, \Pi_i)$$

[7] When there exists an agreeing trial model with equal performance, both $\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}$ and $\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}$ are true according to their definitions.

[8] Equivalence of conditional outcomes could occur if differences in performance do not impact outcomes for the range of performances tested. Note that Assumption 3.1 allows for this occurrence, as it does not assume a strict inequality.

where $\sqrt{n}(L - \hat{L}) \xrightarrow{d} N(0, \sigma^2(\psi_L))$ *where* $\sigma^2(\psi_L)$ *is the variance of* ψ_L *and* $\xrightarrow{d}$ *denotes convergence in distribution, with similar convergence of* $\hat{U}$*, and hence*

$$\left[\hat{L}(\pi_e) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\hat{\sigma}(\psi_L)}{\sqrt{n}}, \right.$$
$$\left. \hat{U}(\pi_e) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\hat{\sigma}(\psi_U)}{\sqrt{n}}\right]$$

is an asymptotically valid $(1-\alpha)$*-confidence interval, where* $\hat{\sigma}(\psi)$ *is the empirical standard deviation of* ψ *and* Φ^{-1} *is the inverse of the standard normal CDF.*

Proposition 3.4 gives a straightforward way of estimating bounds from data as empirical means over the RCT dataset. To give intuition, terms like $P(\Pi \in \mathbf{\Pi}^e_{\geq}(X))$ are known by design (though they depend on X), since $P(\Pi)$ is assumed to be known and the trialed policies Π are known, and so asymptotic normality is straightforward to demonstrate[9]. In Section 5, we use Proposition 3.4 to estimate the bounds for the effect of deploying a new model.

4 RECOMMENDATIONS FOR PRE-TRIAL DESIGN

Recommendation: Conduct trials with multiple models. Our results suggest the utility of trialing multiple models in a cluster RCT that vary in their outputs on different patient populations and which exhibit a range of reasonable performance characteristics. First of all, doing so gives more flexibility to estimate the effect of new models if there are sizeable populations where trialed models raise different outputs. Second, falsification of our main assumptions (Assumptions 3.1 and 3.2) can be done using data from patient populations where the outputs of different models agree.

Recommendation: Use previous trial data to inform deployment of new models. Given a set of models or policies included in a trial, it may be tempting to conclude that an updated model that is more accurate (on average) should be deployed. However, this conclusion may be flawed given the goal of improving patient outcomes. For instance, a model that is more accurate, but achieves higher accuracy by sacrificing performance on some important subpopulation, may ultimately lead to worse outcomes. Indeed, in Section 5, we give a simple simulated example where the optimal model is not the model with the best performance. Given the potential impacts of deploying ML models in practice, it is paramount that we use previous trial data to draw inference on outcomes of interest prior to deploying a new model.

[9]The statistical efficiency of these bounds could potentially be improved by using a doubly-robust-style estimator that incorporates an estimate of terms, such as $\mu_=(X) := E[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)]$, but we present this simpler estimator for ease of exposition and understanding.

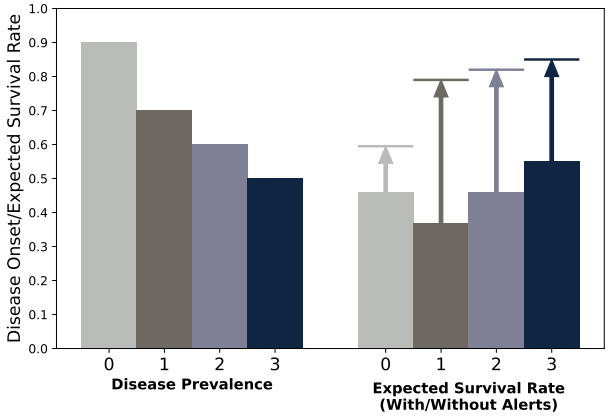

Figure 4: Bar graph giving a visual representation of the data generating process in the simulation study. The bars on the left depict the likelihood of disease onset for varying levels of baseline health X, and the bars on the right depict the likelihood of survival, assuming model performance is fixed at $m = 0.5$. The arrows above the bars on the right show how survival rates change for each level of X when patients receive an alert from an alerting model.

Table 1: Probability of developing disease ($\mathbb{E}[O]$) and expected survival rate ($\mathbb{E}[Y(a, m)]$) for the simulation study.

	$X = 0$	$X = 1$
$\mathbb{E}[O]$	0.9	0.7
$\mathbb{E}[Y(a,m)]$	$0.46 + ((1+m)/2) \cdot 0.18a$	$0.37 + ((1+m)/2) \cdot 0.56a$

	$X = 2$	$X = 3$
$\mathbb{E}[O]$	0.6	0.5
$\mathbb{E}[Y(a,m)]$	$0.46 + ((1+m)/2) \cdot 0.48a$	$0.55 + ((1+m)/2) \cdot 0.4a$

5 SIMULATION STUDY

We now describe a simple simulated example, inspired by Example 1, that demonstrates the results derived in Section 3 and how our proposed method allows for more robust comparisons between models. We consider machine learning models that alert clinicians to the near-term onset of some disease, denoted by $O \in \{0, 1\}$. We simulate a cluster RCT with three arms: A control arm, denoted as a policy π_0 that never alerts, and two arms where models π_1 and π_2 are deployed, respectively. For simplicity, we consider model performance M to be the ground-truth accuracy in predicting disease onset. Our outcome of interest, denoted by $Y \in \{0, 1\}$, is patient survival, and X represents a baseline health characteristic. For the sake of an interpretable simulation, X takes on values uniformly in $\{0, 1, 2, 3\}$, and O and Y are Bernoulli random variables. The probability of $O = 1$ depends only on X whereas the probability of $Y = 1$ depends on X, A, and M. The trialed models are defined as $\pi_1 = \mathbf{1}\{X = 1\}$ and $\pi_2 = \mathbf{1}\{X \in \{2, 3\}\}$.

Table 1 gives expected values of O and $Y(a, m)$ over all possible values of X, satisfying Assumptions 2.1, 3.2 and 3.3;

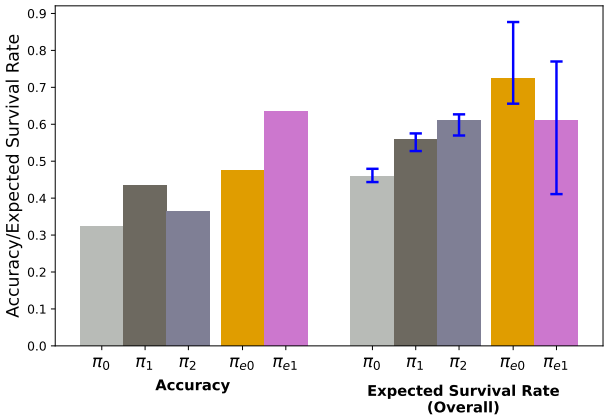

Figure 5: Bar graph showing accuracy at predicting disease and expected patient survival rates of different simulated alerting models. Bars denote ground-truth values, and blue intervals denote estimated bounds using our approach. Because π_{e0} and π_{e1} were not trialed in the RCT, their bounds are wider. Despite π_{e1} having the greatest raw accuracy, π_{e0} has the most positive effect on patient survival rates.

Fig. 4 visualizes the same information for fixed model performance. The aggregate patterns reflect the following process: First, we assume that all patients who do not develop disease will survive. Then, among the sickest individuals ($X = 0$), the survival rate among patients with disease is 40% without alerts[10] and between 50-60% with alerts, as influenced by m, where 50% corresponds to a model with zero accuracy, and 60% is achieved by a model with perfect accuracy. These patterns reflect the intuition that survival is improved (when alerts are raised) under a model perceived to be more accurate. However, the survival rate (among those with disease) is much improved for other groups, going from 10% without alerts to 50-90% with alerts[11]. Since the actual prevalence of disease varies across these groups, and since alerts only help those with disease, the overall causal effect is largest among those where $X = 1$. Appendix C describes and explains this data generating process in detail.

We consider the effect of deploying two new models, $\pi_{e0} = \mathbf{1}\{X \in \{1, 2, 3\}\}$ and $\pi_{e1} = \mathbf{1}\{X \in \{0, 1\}\}$, whose accuracies at predicting onset are computable using Table 1. We then estimate bounds on $\mathbb{E}[Y(A = \pi, M = f_M(\pi))]$ for each model (including confidence intervals to incorporate finite-sample uncertainty, as described under Proposition 3.4) using data from a simulation with $n = 5,000$. Fig. 5 shows the simulation results and illustrates that **the most accurate model is not always the best**: π_{e1} has the greatest accuracy in predicting the onset of disease, but π_{e0} has the largest causal impact on patient outcomes. This reversal occurs since π_{e0} raises alerts for patients who stand to benefit the most, whereas π_{e1} tends to alert for patients who have little to gain from them. Moreover, our bounds reflect greater confidence in the (positive) impact of π_{e0}, since the lower bound for patient outcomes under π_{e0} is greater than patient outcomes under all trialed models. Python code implementing this simulation study, implementing the estimation procedure proposed in Proposition 3.4, and for generating Figs. 4 and 5 are publicly available online at: https://github.com/jacobmchen/just_trial_once.

[10] To align this explanation with Table 1, note that all of the 10% without disease and 40% of the 90% with disease survive, yielding the overall survival rate of 46% without alerts.

[11] For explicit computations of survival rates with and without alerts, see Appendix C.

6 CONCLUSION AND LIMITATIONS

In this paper, we discussed methods for estimating the causal impact of new or updated ML and AI models not previously trialed in an RCT. Under the important considerations that ML predictions are deterministic and that clinician trust in ML models play a role in determining their impacts on patient outcomes, we demonstrated how one could estimate lower and upper bounds on the effect of deploying a new model. We further proved tightness of our proposed bounds (Theorem 3.2) and gave inverse probability weighted-style estimators for them (Proposition 3.4). Given the possibility that key assumptions for employing our methods may not be fulfilled, we also proposed simple strategies for testing and falsifying them. Finally, we concluded with a simulation study to illustrate the application of our method and to highlight its benefits when selecting among model updates.

However, our work is not without limitations: First, our derived bounds are naturally pessimistic, and, while they cannot be tightened without further assumptions (Theorem 3.2), some additional assumptions may be warranted in certain cases. For instance, we implicitly assume that "anything can happen" when a model raises an alert on patients who never received alerts in the past. One could instead assume that alerts are not harmful (except for their impact on performance), or that they are not harmful for a specific patient if the alert is correct. Second, we assume that model performance can be summarized in a single real number, but a more complex representation (e.g., involving subgroup-specific performance, performance adaptation over time, and clinician experience) may be warranted in some applications. Finally, a core limitation of our approach is that we still require an RCT. Using RCT data comes with many benefits: It allows for greater confidence that core assumptions (e.g., randomization of policies) hold by design, and even allows for checking (in some cases) the core assumptions we make in this paper, as we have shown (Propositions 3.1 and 3.2). However, observational studies (e.g., pre-post studies of model deployments) are often easier to run in practice. Our hope is that this work can serve as a springboard towards increasing the utility, and thus adoption, of RCTs for ML models deployed in high-risk settings.

Acknowledgements

The authors are grateful for the helpful feedback of anonymous reviewers as well as fruitful discussions with Jonathan Zhang, Erik Skalnes, and Trung Phung.

References

Roy Adams, Katharine E Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N Hager, Sara E Cosgrove, Andrew Markowski, Eili Y Klein, Edward S Chen, Mustapha O Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C Linton, Anushree R Ahluwalia, Albert W Wu, and Suchi Saria. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature medicine*, 28(7):1455–1460, July 2022.

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. *arXiv*, 2403:v3, 2024.

Eli Ben-Michael, D James Greiner, Kosuke Imai, and Zhichao Jiang. Safe policy learning through extrapolation: Application to pre-trial risk assessment. *Journal of the American Statistical Association*, pages 1–23, 2025.

Aaron Boussina, Supreeth P. Shashikumar, Atul Malhotra, Robert L. Owens, Robert El-Kareh, Christopher A. Longhurst, Kimberly Quintero, Allison Donahue, Theodore C. Chan, Shamim Nemati, and Gabriel Wardi. Impact of a deep learning sepsis prediction model on quality of care and survival. *npj Digital Medicine*, 7, 1 2024. doi: 10.1038/s41746-023-00986-6. URL http://dx.doi.org/10.1038/s41746-023-00986-6.

FDA. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. *U.S. Food and Drug Administration*, 2024. URL https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed January 14th, 2024.

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*, 331:2018, 6 2024a. doi: 10.1001/jama.2024.6259. URL http://dx.doi.org/10.1001/jama.2024.6259.

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 pneumonia. *JAMA*, 331:2007, 6 2024b. doi: 10.1001/jama.2024.6248. URL http://dx.doi.org/10.1001/jama.2024.6248.

Ryan Han, Julián N Acosta, Zahra Shakeri, John P A Ioannidis, Eric J Topol, and Pranav Rajpurkar. Randomised controlled trials evaluating artificial intelligence in clinical practice: A scoping review. *The Lancet Digital Health*, 6: e367–e373, 5 2024. doi: 10.1016/s2589-7500(24)00047-5. URL http://dx.doi.org/10.1016/s2589-7500(24)00047-5.

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*, 186:167–189, 5 2023. doi: 10.1093/jrsssa/qnad010. URL https://doi.org/10.1093/jrsssa/qnad010.

Shalmali Joshi, Iñigo Urteaga, Wouter AC van Amsterdam, George Hripcsak, Pierre Elias, Benjamin Recht, Noémie Elhadad, James Fackler, Mark P Sendak, Jenna Wiens, et al. AI as an intervention: improving clinical outcomes relies on a causal approach to AI development and validation. *Journal of the American Medical Informatics Association*, page ocae301, 2025.

Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088. PMLR, 2019.

David Ouyang and Joseph Hogan. We need more randomized clinical trials of AI, 2024.

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

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

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.

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.

Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, Kelly Kester, Sahil Sandhu, Kristin Corey, Nathan Brajer, Christelle Tan, Anthony Lin, Tres Brown, Susan Engelbosch, Kevin Anstrom, Madeleine Clare Elish, Katherine Heller, Rebecca Donohoe, Jason Theiling, Eric Poon, Suresh Balu, Armando Bedoya, and Cara O' Brien. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Medical Informatics*, 8:e15182, 7 2020. doi: 10.2196/15182. URL http://dx.doi.org/10.2196/15182.

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.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 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, 10 2024. doi: 10.1056/aioa2400865. URL http://dx.doi.org/10.1056/aioa2400865.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Supplementary Material

Jacob M. Chen[1] **Michael Oberst**[1]

[1]Department of Computer Science, Johns Hopkins University

A BRIEF CAUSAL INFERENCE OVERVIEW

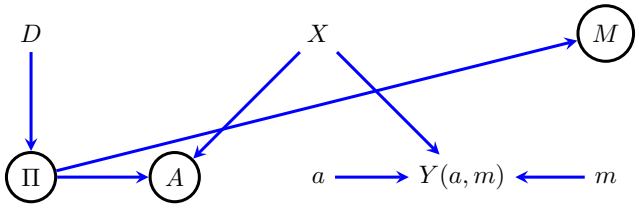

Figure 6: SWIG showing $\mathcal{G}(a, m)$.

In this section, we give a brief overview of additional topics in causal inference that we use in our proofs and provide references for further reading on these topics.

First, independencies implied in the full data distribution $P(\mathbf{V})$ can be read off from $\mathcal{G}$ using the d-separation criterion [Pearl, 2014]. Next, we follow [Richardson and Robins, 2013] to convert a DAG $\mathcal{G}$, given an intervened on variable W, to a single-world intervention graph (SWIG) using a node-splitting transformation as follows: (1) replace all children of W (here, V) with the potential outcome random variable $V(w)$, (2) add the intervened on variable w into $\mathcal{G}$ as a new vertex, and (3) change all outgoing edges from W to originate from w instead. This new SWIG is denoted by $\mathcal{G}(w)$. It is possible to intervene on multiple variables in $\mathcal{G}$ by repeatedly applying the node-splitting transformation described above. Fig. 6 shows the SWIG $\mathcal{G}(a, m)$, the SWIG containing the potential outcome random variable $Y(a, m)$ corresponding to the target estimand $\mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e))]$.

In order to connect potential outcome distributions to observed distributions, [Malinsky et al., 2019] propose a set of three rules known as the *potential outcome calculus* (po-calculus). Most salient to our proofs is Rule 2 of po-calculus, which states that $P(V(w) \mid \mathbf{X}) = P(V \mid \mathbf{X}, W = w)$ if $V(w) \perp\!\!\!\perp W \mid \mathbf{X}$ in $\mathcal{G}(w)$, where $\mathbf{X}$ is any set of random variables in $\mathcal{G}(w)$, including the empty set. Rule 2 is also referred to as the *conditional ignorability* assumption commonly used in causal inference, which states that a potential outcome of interest is independent of a treatment of interest conditional on a sufficiently rich set of covariates $\mathbf{X}$. The benefit of using Rule 2 of po-calculus is that it allows us to use d-separation in a SWIG to directly verify whether conditional ignorability holds given a graph. For instance, note that $Y(a, m) \perp\!\!\!\perp A, M \mid X$ in the SWIG shown in Fig. 6. Rule 2 of po-calculus thus allows us to conclude that $P(Y(a, m) \mid X) = P(Y \mid A = a, M = m, X)$.

We now formally define some concepts from our discussion above that we use in our proofs.

Corollary A.1 (Consistency)**.** *Under Assumption 2.1, if* $A = a, M = m$*, then* $Y = Y(A = a, M = m)$.

Corollary A.2 (Conditional Ignorability)**.** *Under Assumption 2.1, since* $Y(a, m) \perp\!\!\!\perp A, M \mid X$*, then* $P(Y(a, m) \mid X) = P(Y \mid A = a, M = m, X)$.

B ALGORITHMIC VIEWPOINT OF COMPUTING BOUNDS

Here, we give an algorithmic construction of the lower and upper bounds presented in Theorem 3.1. First, we define sub-algorithms in Algorithm 1 that compute the lower and upper bounds conditional on a value of $x \in \mathcal{X}$. Next, we define the algorithm that constructs the complete bounds in Algorithm 2 by iterating through all possible values of $x \in \mathcal{X}$ and computing the weighted average of the bounds for each x.

Algorithm 1 Sub-Algorithms for computing lower and upper bounds conditional on X.

Input: x
Output: Tight lower bound on Y conditional on $X = x$
function CONDITIONAL_LOWER(x)
    if $\pi_e(x) \neq a_0$ **then**
        if $\tilde{\mathbf{\Pi}}^e_{\leq}(x) \neq \varnothing$ **then return** $\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)]$
        else if $\tilde{\mathbf{\Pi}}^e_{\leq}(x) = \varnothing$ **then return** Y_{min}
        end if
    else if $\pi_e(x) = a_0$ **then**
        if $\mathbf{\Pi}^e(x) \neq \varnothing$ **then return** $\mathbb{E}[Y \mid X = x, \Pi \in \mathbf{\Pi}^e(x)]$
        else if $\mathbf{\Pi}^e(x) = \varnothing$ **then return** Y_{min}
        end if
    end if
end function

Input: x
Output: Tight upper bound on Y conditional on $X = x$
function CONDITIONAL_UPPER(x)
    if $\pi_e(x) \neq a_0$ **then**
        if $\tilde{\mathbf{\Pi}}^e_{\geq}(x) \neq \varnothing$ **then return** $\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)]$
        else if $\tilde{\mathbf{\Pi}}^e_{\geq}(x) = \varnothing$ **then return** Y_{max}
        end if
    else if $\pi_e(x) = a_0$ **then**
        if $\mathbf{\Pi}^e(x) \neq \varnothing$ **then return** $\mathbb{E}[Y \mid X = x, \Pi \in \mathbf{\Pi}^e(x)]$
        else if $\mathbf{\Pi}^e(x) = \varnothing$ **then return** Y_{max}
        end if
    end if
end function

Algorithm 2 Algorithm for computing lower and upper bounds for $\mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e)]$.

$lower\_bound \leftarrow 0$
$upper\_bound \leftarrow 0$
for $x \in \mathcal{X}$ **do**
    $lower\_bound \leftarrow lower\_bound +$ CONDITIONAL_LOWER$(x) \times P(X = x)$
    $upper\_bound \leftarrow upper\_bound +$ CONDITIONAL_UPPER$(x) \times P(X = x)$
end for
return $(lower\_bound, upper\_bound)$

C DATA GENERATING PROCESS OF SIMULATION STUDY

Here, we define in detail the data generating process for the simulation study in Section 5. The trialed models are defined as $\pi_0 = 0$, $\pi_1 = \mathbf{1}\{X = 1\}$, and $\pi_2 = \mathbf{1}\{X \in \{2, 3\}\}$. We use $\text{acc}(\pi_i)$ to denote the accuracy of policy π_i at predicting

disease O. One can verify that this data generating process obeys the structural causal model given in Assumption 2.1.

$$
\begin{aligned}
X &\sim \text{discrete uniform}[0,3] \\
D &\sim \text{discrete uniform}[0,2] \\
\Pi &= \pi_D \\
A &= \pi_D(X) \\
O &\sim \text{Bernoulli}(\mathbf{1}\{X=0\}0.9 + \mathbf{1}\{X=1\}0.7 + \mathbf{1}\{X=2\}0.6 + \mathbf{1}\{X=3\}0.5) \\
M &= \text{acc}(\Pi) \\
Y &\sim \text{Bernoulli}(\mathbf{1}\{X=0\}(0.46 + ((1+m)/2)\cdot 0.18A) + \mathbf{1}\{X=1\}(0.37 + ((1+m)/2)\cdot 0.56A) \\
&\quad + \mathbf{1}\{X=2\}(0.46 + ((1+m)/2)\cdot 0.48A) + \mathbf{1}\{X=3\}(0.55 + ((1+m)/2)\cdot 0.4A)
\end{aligned}
$$

We now explain in detail how the probabilities of the survival rate Y are computed. First, conditional on patients who have the disease ($O=1$), the survival rates without and with alerts are as follows:

Table 2: Probability of survival ($\mathbb{E}[Y(a,m)]$) under no alerts ($\mathbb{E}[Y(0,m)]$) and alerts ($\mathbb{E}[Y(1,m)]$) for the simulation study.

	$X=0$	$X=1$
$\mathbb{E}[Y(0,m) \mid O=1]$	0.4	0.1
$\mathbb{E}[Y(1,m) \mid O=1]$	$0.4 + ((1+m)/2)\cdot 0.2$	$0.1 + ((1+m)/2)\cdot 0.8$

	$X=2$	$X=3$
$\mathbb{E}[Y(0,m) \mid O=1]$	0.1	0.1
$\mathbb{E}[Y(1,m) \mid O=1]$	$0.1 + ((1+m)/2)\cdot 0.8$	$0.1 + ((1+m)/2)\cdot 0.8$

When patients do not have the disease ($O=0$), their probability of survival is 1, that is $\mathbb{E}[Y(a,m) \mid O=0] = 1$ for all values of A and M. The assumption that lower model performance has a weakening effect on the positive effect of alerting is implicit in Table 2. For patients with $X=0$ and $O=1$, the survival probability with alerts range from 0.5 when accuracy is 0 to 0.6 when accuracy is 1. For all other patients, the survival probability with alerts range from 0.5 when accuracy is 0 to 0.9 when accuracy is 1.

Now, to compute the probabilities of survival given in Table 1, we apply the law of total probability – $\mathbb{E}[Y(a,m)] = \mathbb{E}[Y(a,m) \mid O=0]p(O=0) + \mathbb{E}[Y(a,m) \mid O=1]p(O=1)$ – for each value of X. We give this explicit computation in the following table.

Table 3: Probability of survival ($\mathbb{E}[Y(a,m)]$) computed explicitly when summed over the likelihood of developing the disease ($O=1$).

	$X=0$	$X=1$
$\mathbb{E}[Y(a,m)]$	$1\cdot 0.1 + (0.4 + ((1+m)/2)0.2a)\cdot 0.9$	$1\cdot 0.3 + (0.1 + ((1+m)/2)0.8a)\cdot 0.7$

	$X=2$	$X=3$
$\mathbb{E}[Y(a,m)]$	$1\cdot 0.4 + (0.1 + ((1+m)/2)0.8a)\cdot 0.6$	$1\cdot 0.5 + (0.1 + ((1+m)/2)0.8a)\cdot 0.5$

After simplifying the expressions above, we get the probabilities of survival shown in Table 1. Note that the dependence of Y on O does not violate Assumption 2.1 and does not change Fig. 2 because O is an unobserved variable on the directed path from X to Y.

D PROOFS

Proposition 3.1 (Falsification of Assumption 3.1)**.** *Let $\mathcal{X}$ denote the full space of possible covariate values. Under Assumption 2.1, given data from an RCT that includes at least two trialed models π_1 and π_2 with different levels of performance $f_M(\pi_1) < f_M(\pi_2)$, and whose actions agree on a non-empty set of patients $\mathcal{X}_{agree} := \{x \in \mathcal{X} \mid \pi_1(x) = \pi_2(x)\}$ such that $P(X \in \mathcal{X}_{agree}) > 0$, the observation that*

$$\mathbb{E}[Y \mid X \in \mathcal{X}_{agree}, \Pi = \pi_2] < \mathbb{E}[Y \mid X \in \mathcal{X}_{agree}, \Pi = \pi_1],$$

implies that Assumption 3.1 is false.

Proof. We will show that if Assumption 3.1 holds, then the stated observation will yield a contradiction. We will first show that the Assumption 3.1 implies a point-wise inequality over x, and then show that this inequality holds when aggregating over $X \in \mathcal{X}_{\text{agree}}$. First, choose any $x \in \mathcal{X}_{\text{agree}}$. Then we can write

$$\begin{aligned}
\mathbb{E}[Y \mid X = x, \Pi = \pi_i] &= \mathbb{E}[Y \mid X = x, A = \pi_i(x), M = f_M(\pi_i), \Pi = \pi_i] && (12)\\
&= \mathbb{E}[Y(A = \pi_i(x), M = f_M(\pi_i)) \mid X = x, A = \pi_i(x), M = f_M(\pi_i), \Pi = \pi_i] && (13)\\
&= \mathbb{E}[Y(A = \pi_i(x), M = f_M(\pi_i)) \mid X = x] && (14)
\end{aligned}$$

where Eq. (12) follows from the implication $\{\Pi = \pi_i, X = x\} \implies \{A = \pi_i(x), M = f_M(\pi_i)\}$, Eq. (13) follows from consistency (Corollary A.1), and Eq. (14) follows from conditional ignorability (Corollary A.2). Since Eq. (14) holds for both policies π_1, π_2, we can then write that

$$\begin{aligned}
&\mathbb{E}[Y \mid X = x, \Pi = \pi_2] - \mathbb{E}[Y \mid X = x, \Pi = \pi_1]\\
&\quad = \mathbb{E}[Y(A = \pi_2(x), M = f_M(\pi_2)) - Y(A = \pi_1(x), M = f_M(\pi_1)) \mid X = x] && (15)\\
&\quad \geq 0 && (16)
\end{aligned}$$

where Eq. (15) follows from linearity of expectation, and Eq. (16) follows from Assumption 3.1, since $\pi_1(x) = \pi_2(x)$ by construction, and $f_M(\pi_2) > f_M(\pi_1)$. To aggregate, we first observe that $X \perp\!\!\!\perp \Pi$ in our data generating process (as the policies are assigned randomly and independently of X). Hence $P(X \mid \Pi = \pi_2, X \in \mathcal{X}_{\text{agree}}) = P(X \mid \Pi = \pi_1, X \in \mathcal{X}_{\text{agree}}) = P(X \mid X \in \mathcal{X}_{\text{agree}})$. Accordingly, we can write that

$$\begin{aligned}
&\mathbb{E}[Y \mid X \in \mathcal{X}_{\text{agree}}, \Pi = \pi_2] - \mathbb{E}[Y \mid X \in \mathcal{X}_{\text{agree}}, \Pi = \pi_1]\\
&\quad = \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_2] dP(x \mid \Pi = \pi_2, X \in \mathcal{X}_{\text{agree}}) - \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_1] dP(x \mid \Pi = \pi_1, X \in \mathcal{X}_{\text{agree}})\\
&\quad = \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_2] dP(x \mid X \in \mathcal{X}_{\text{agree}}) - \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_1] dP(x \mid X \in \mathcal{X}_{\text{agree}})\\
&\quad = \int_x \left(\mathbb{E}[Y \mid X = x, \Pi = \pi_2] - \mathbb{E}[Y \mid X = x, \Pi = \pi_1]\right) dP(x \mid X \in \mathcal{X}_{\text{agree}})\\
&\quad \geq 0
\end{aligned}$$

where the final inequality follows from the point-wise inequality in Eq. (16), and which directly gives the implication

$$\mathbb{E}[Y \mid X \in \mathcal{X}_{\text{agree}}, \Pi = \pi_2] \geq \mathbb{E}[Y \mid X \in \mathcal{X}_{\text{agree}}, \Pi = \pi_1],$$

which is contradicted in the case where the stated observation (the relationship $<$ instead of $\geq$) holds. □

Proposition 3.2 (Falsification of Assumption 3.2)**.** *Under Assumption 2.1, given data from an RCT that includes at least two trialed models π_1 and π_2 with different levels of performance $f_M(\pi_1) < f_M(\pi_2)$, and which both models take the neutral action a_0 on a non-empty set of patients $\mathcal{X}_{a_0} := \{x \in \mathcal{X} \mid \pi_1(x) = \pi_2(x) = a_0\}$ such that $P(X \in \mathcal{X}_{a_0}) > 0$, the observation that*

$$\mathbb{E}[Y \mid X \in \mathcal{X}_{a_0}, \Pi = \pi_2] \neq \mathbb{E}[Y \mid X \in \mathcal{X}_{a_0}, \Pi = \pi_1],$$

implies that Assumption 3.2 is false.

Proof. Our proof follows a similar structure to that of Proposition 3.1. We will show that if Assumption 3.2 holds, then the stated observation would yield a contradiction. We will first show that the Assumption 3.2 implies a point-wise equality over x, and then show that this inequality holds when aggregating over $X \in \mathcal{X}_{a_0}$. First, choose any $x \in \mathcal{X}_{a_0}$. Then we can write

$$\mathbb{E}[Y \mid X = x, \Pi = \pi_i] = \mathbb{E}[Y(A = \pi_i(x), M = f_M(\pi_i)) \mid X = x] \quad (17)$$

using the same argument as in Proposition 3.1 (namely, the implication that $\{\Pi = \pi_i, X = x\} \implies \{A = \pi_i(x), M = f_M(\pi_i)\}$, consistency (Corollary A.1), and conditional ignorability (Corollary A.2). Since Eq. (17) holds for both policies

π_1, π_2, we can then write that

$$\begin{aligned}
&\mathbb{E}[Y \mid X = x, \Pi = \pi_2] - \mathbb{E}[Y \mid X = x, \Pi = \pi_1] \\
&\quad = \mathbb{E}[Y(A = \pi_2(x), M = f_M(\pi_2)) - Y(A = \pi_1(x), M = f_M(\pi_1)) \mid X = x] && (18) \\
&\quad = \mathbb{E}[Y(A = a_0, M = f_M(\pi_2)) - Y(A = a_0, M = f_M(\pi_1)) \mid X = x] && (19) \\
&\quad = 0 && (20)
\end{aligned}$$

where Eq. (18) follows from linearity of expectation, Eq. (19) follows from the fact that $x \in \mathcal{X}_{a_0}$, and Eq. (20) follows from Assumption 3.2, which states that $Y(A = a_0, M = m) = Y(A = a_0, M = m')$ for any m, m'.

To aggregate, we first observe (similar to the proof of Proposition 3.1) that $X \perp\!\!\!\perp \Pi$ in our data generating process (as the policies are assigned randomly and independently of X). As a result, we can write that

$$\begin{aligned}
&\mathbb{E}[Y \mid X \in \mathcal{X}_{a_0}, \Pi = \pi_2] - \mathbb{E}[Y \mid X \in \mathcal{X}_{a_0}, \Pi = \pi_1] \\
&\quad = \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_2] dP(x \mid \Pi = \pi_2, X \in \mathcal{X}_{a_0}) - \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_1] dP(x \mid \Pi = \pi_1, X \in \mathcal{X}_{a_0}) \\
&\quad = \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_2] dP(x \mid X \in \mathcal{X}_{a_0}) - \int_x \mathbb{E}[Y \mid X = x, \Pi = \pi_1] dP(x \mid X \in \mathcal{X}_{a_0}) \\
&\quad = \int_x \left(\mathbb{E}[Y \mid X = x, \Pi = \pi_2] - \mathbb{E}[Y \mid X = x, \Pi = \pi_1]\right) dP(x \mid X \in \mathcal{X}_{a_0}) \\
&\quad = 0
\end{aligned}$$

where the final equality follows from the point-wise equality in Eq. (20), and which directly gives the implication

$$\mathbb{E}[Y \mid X \in \mathcal{X}_{a_0}, \Pi = \pi_2] = \mathbb{E}[Y \mid X \in \mathcal{X}_{a_0}, \Pi = \pi_1],$$

which is contradicted in the case where the stated observation (an inequality instead of equality) holds. □

D.1 PROOF OF THEOREM 3.1

Before we give the proof of Theorem 3.1, we restate Definition 3.1 from the main text, for ease of reference when reviewing the proof.

Definition 3.1 (Policy/Model Sets)**.** For each value of $x \in \mathcal{X}$, we define the sets of trialed policies/models (possibly none) that agree with $\pi_e(x)$ and subsets of this set based on the performance characteristics of those trialed models[1].

$$\begin{aligned}
\mathbf{\Pi}^e(x) &\coloneqq \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x)\} \\
\mathbf{\Pi}^e_{\leq}(x) &\coloneqq \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x), f_M(\pi) \leq f_M(\pi_e)\} \\
\mathbf{\Pi}^e_{\geq}(x) &\coloneqq \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x), f_M(\pi) \geq f_M(\pi_e)\}
\end{aligned}$$

We also further define subsets of $\mathbf{\Pi}^e_{\leq}$ and $\mathbf{\Pi}^e_{\geq}$ that contain only the next-worst or next-best performing model[2].

$$\begin{aligned}
\tilde{\mathbf{\Pi}}^e_{\leq}(x) &\coloneqq \arg\max_{\pi \in \mathbf{\Pi}^e_{\leq}(x)} f_M(\pi), \\
\tilde{\mathbf{\Pi}}^e_{\geq}(x) &\coloneqq \arg\min_{\pi \in \mathbf{\Pi}^e_{\geq}(x)} f_M(\pi)
\end{aligned}$$

Armed with Definition 3.1, we first state some useful inequalities that will form the core of the proof for Theorem 3.1

Lemma D.1 (Independence under neutral actions)**.** *Under the assumed data generating process (Assumption 2.1) and Assumption 3.2,*

$$Y \perp\!\!\!\perp M \mid X, A = a_0 \tag{21}$$

[1] All these sets are defined with respect to the model π_e and could be written more precisely with π_e as an argument instead of in the superscript (e.g., $\mathbf{\Pi}(x, \pi_e)$) but we use the superscript notation for conciseness.

[2] Where relevant, we use the convention that $\arg\min_{\pi \in \varnothing}(f_M(\pi)) = \varnothing$.

Proof. This claim follows from Assumption 3.2 in a straightforward fashion. Let m and m' be any two distinct values of M, then

$$\begin{aligned}
P(Y \mid X=x, M=m, A=a_0) &= P(Y(A=a_0, M=m) \mid X=x, M=m, A=a_0) && \text{Consistency} \\
&= P(Y(A=a_0, M=m') \mid X=x, M=m, A=a_0) && \text{By Assumption 3.2} \\
&= P(Y(A=a_0, M=m') \mid X=x, M=m', A=a_0) && Y(a,m) \perp\!\!\!\perp M, A \mid X \\
&= P(Y \mid X=x, M=m', A=a_0) && \text{Consistency}
\end{aligned}$$

The claim follows from the fact that we have shown equality $P(Y \mid X=x, M=m, A=a_0) = P(Y \mid X=x, M=m', A=a_0)$ for arbitrary m, m'. □

Lemma D.2 (Outcome Equalities / Inequalities)**.** *Under Assumptions 3.1 and 3.2, the following inequalities hold, where we use π_e as shorthand for $\pi_e(x)$,*

$$\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(x) \neq \varnothing\}\mathbb{E}[Y \mid X=x, A=\pi_e, M=f_M(\pi_e)] \geq \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(x) \neq \varnothing\}\mathbb{E}[Y \mid X=x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)] \tag{22}$$

$$\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(x) \neq \varnothing\}\mathbb{E}[Y \mid X=x, A=\pi_e, M=f_M(\pi_e)] \leq \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(x) \neq \varnothing\}\mathbb{E}[Y \mid X=x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] \tag{23}$$

$$\begin{aligned}
&\mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}[Y \mid X=x, A=\pi_e, M=f_M(\pi_e)] = \\
&\qquad\qquad \mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}[Y \mid X=x, \Pi \in \mathbf{\Pi}^e(x)]
\end{aligned} \tag{24}$$

Proof. The proof for each of these relations follows a similar structure. For each, we only need to consider the relation when the stated indicator (identical on either side) is equal to 1, since all are trivially true when the indicator is equal to zero.

For Eq. (22), the event $\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)\}$ implies $\{A = \pi_e(x), M \leq f_M(\pi_e)\}$ from the definition of $\tilde{\mathbf{\Pi}}^e_{\leq}(x)$ (see Definition 3.1). By the independence $Y \perp\!\!\!\perp \Pi \mid X, A, M$, we have it that $\mathbb{E}[Y \mid X=x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)] = E[Y \mid X=x, A=\pi_e, M \leq f_M(\pi_e)]$, and this conditional expectation is well-defined when $\tilde{\mathbf{\Pi}}^e_{\leq}(x) \neq \varnothing$. Finally, we make use of Assumption 3.1, which implies that $\mathbb{E}[Y \mid X=x, A=\pi_e, M=f_M(\pi_e)] \geq \mathbb{E}[Y \mid \bar{X}=x, A=\pi_e, M \leq f_M(\pi_e)]$. The stated inequality follows.

For Eq. (23), the event $\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)\}$ implies $\{A = \pi_e(x), M \geq f_M(\pi_e)\}$ from the definition of $\tilde{\mathbf{\Pi}}^e_{\geq}(x)$ (see Definition 3.1). By the independence $Y \perp\!\!\!\perp \Pi \mid X, A, M$, we have it that $\mathbb{E}[Y \mid X=x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] = E[Y \mid X=x, A=\pi_e, M \geq f_M(\pi_e)]$, and this conditional expectation is well-defined when $\tilde{\mathbf{\Pi}}^e_{\geq}(x) \neq \varnothing$. Finally, we make use of Assumption 3.1, which implies that $\mathbb{E}[Y \mid X=x, A=\pi_e, M=f_M(\pi_e)] \leq \mathbb{E}[Y \mid \bar{X}=x, A=\pi_e, M \geq f_M(\pi_e)]$. The stated inequality follows.

Finally, for Eq. (24), we can observe that

$$\begin{aligned}
&\mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}[Y \mid X=x, \Pi \in \mathbf{\Pi}^e(x)] \\
&= \mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}[Y \mid X=x, A=\pi_e(x), \Pi \in \mathbf{\Pi}^e(x)] && (25) \\
&= \mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}_M[\mathbb{E}[Y \mid X=x, A=\pi_e(x), \Pi \in \mathbf{\Pi}^e(x), M] \mid X=x, A=\pi_e(x), \Pi \in \mathbf{\Pi}^e(x)] && (26) \\
&= \mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}_M[\mathbb{E}[Y \mid X=x, A=\pi_e(x), M] \mid X=x, A=\pi_e(x), \Pi \in \mathbf{\Pi}^e(x)] && (27) \\
&= \mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}_M[\mathbb{E}[Y \mid X=x, A=a_0, M] \mid X=x, A=\pi_e(x), \Pi \in \mathbf{\Pi}^e(x)] && (28) \\
&= \mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}_M[\mathbb{E}[Y \mid X=x, A=a_0] \mid X=x, A=\pi_e(x), \Pi \in \mathbf{\Pi}^e(x)] && (29) \\
&= \mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}[Y \mid X=x, A=a_0] && (30) \\
&= \mathbf{1}\{\mathbf{\Pi}^e(x) \neq \varnothing, \pi_e = a_0\}\mathbb{E}[Y \mid X=x, A=\pi_e, M=f_M(\pi_e)] && (31)
\end{aligned}$$

where Eq. (25) follows from the fact that the event $\{\Pi \in \mathbf{\Pi}^e(x)\} \implies \{A = \pi_e(x)\}$ by the definition of $\mathbf{\Pi}^e(x)$ (see Definition 3.1). Equation (26) simply applies the law of total probability, including M in the inner expectation (and where we use $\mathbb{E}_M$ as helpful shorthand to remind the reader that the outer expectation is taken over M). Equation (27) uses the fact that $Y \perp\!\!\!\perp \Pi \mid A, M, X$ in our data-generating process to remove Π from the inner expectation. Equation (28) replaces $\pi_e(x)$ with a_0 due to the outside indicator that restricts to x where $\pi_e(x) = a_0$, so whenever this expression is non-zero, then $\pi_e(x) = a_0$. Finally, Eq. (29) uses Lemma D.1, which implies that $\mathbb{E}[Y \mid X=x, A=a_0, M=m] = \mathbb{E}[Y \mid X=x, A=a_0]$, and Eq. (30) uses the fact that the inner expectation is a constant value, to remove the outer expectation over M. From Eq. (30), we can simply add back the conditioning on $M = f_M(\pi_e)$, again using Lemma D.1, and recall that $\pi_e = a_0$ under the indicator, to arrive at Eq. (31), which completes the proof. □

We will also make use of the following inequalities that follow from boundedness of Y under Assumption 3.3.

Lemma D.3 (Boundedness)**.** *Under Assumption 3.3, the following inequalities hold by the fact that $Y \in [Y_{min}, Y_{max}]$.*

$$\begin{aligned}
\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}\mathbb{E}[Y \mid X = x, A = \pi_e, M = f_M(\pi_e)] &\geq \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}Y_{min} \\
\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}\mathbb{E}[Y \mid X = x, A = \pi_e, M = f_M(\pi_e)] &\leq \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}Y_{max} \\
\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(x) = a_0\}\mathbb{E}[Y \mid X = x, A = a_0, M = f_M(\pi_e)] &\geq \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(x) = a_0\}Y_{min} \\
\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(x) = a_0\}\mathbb{E}[Y \mid X = x, A = a_0, M = f_M(\pi_e)] &\leq \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(x) = a_0\}Y_{max}
\end{aligned}$$

Proof. Each claim is immediate from the fact that Y is bounded between Y_{min} and Y_{max}, with the additional observation that for each inequality, the indicators are identical on either side. □

We are now prepared to prove our main result.

Theorem 3.1. *Given the data generating process in Assumption 2.1, and under Assumptions 3.1 to 3.3, the policy value of a model / policy π_e is bounded as*

$$L(\pi_e) \leq \mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e))] \leq U(\pi_e), \tag{3}$$

where

$$\begin{aligned}
L(\pi_e) = \mathbb{E}\big[&\mathbf{1}\{\pi_e \neq a_0\}\big(\\
&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)] && (4) \\
&+\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}Y_{min}\big) && (5) \\
&+\mathbf{1}\{\pi_e = a_0\}\big(\\
&\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)] && (6) \\
&+\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}Y_{min}\big)\big] && (7) \\
U(\pi_e) = \mathbb{E}\big[&\mathbf{1}\{\pi_e \neq a_0\}\big(\\
&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)] \\
&+\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}Y_{max}\big) \\
&+\mathbf{1}\{\pi_e = a_0\}\big(\\
&\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)] \\
&+\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}Y_{max}\big)\big]
\end{aligned}$$

These bounds are still valid if we replace $\tilde{\mathbf{\Pi}}^e_{\leq}(X)$ with $\mathbf{\Pi}^e_{\leq}(X)$ and $\tilde{\mathbf{\Pi}}^e_{\geq}(X)$ with $\mathbf{\Pi}^e_{\geq}(X)$.

Proof. We begin with the lower bound, and note that the upper bound follows similarly.

Lower Bound First, we observe that the given set indicators form a partition over $\mathcal{X}$ (a set of disjoint subsets of $\mathcal{X}$ whose union is equal to $\mathcal{X}$), such that

$$\begin{aligned}
1 = &\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\} + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0\} \\
&+ \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\} + \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\}
\end{aligned} \tag{32}$$

Here, the subsets of $X \in \mathcal{X}$ satisfying each of $\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}$ and $\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}$ form a partition over $\mathcal{X}$ by definition as does $\{\mathbf{\Pi}^e(X) \neq \varnothing\}$ and $\{\mathbf{\Pi}^e(X) = \varnothing\}$. Hence, we can write that

$$
\begin{aligned}
&\mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e))] \\
&\quad = \mathbb{E}[\mathbb{E}[Y(A = \pi_e, M = f_M(\pi_e)) \mid X]] && (33)\\
&\quad = \mathbb{E}[\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e))]] && (34)\\
&\quad = \mathbb{E}[(\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\} + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0\} \\
&\qquad\quad + \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\} + \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\})\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)]] && (35)\\
&\quad = \mathbb{E}\big[\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \\
&\qquad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \\
&\qquad + \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \\
&\qquad + \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)]\big] && (36)
\end{aligned}
$$

where Eq. (33) follows from the law of iterated expectation, Eq. (34) follows from consistency (Corollary A.1) and the fact that $Y(a, m) \perp\!\!\!\perp A, M \mid X$ (Corollary A.2), and Eq. (35) follows from Eq. (32). After distributing terms in Eq. (36), the lower bound follows from the application of Lemmas D.2 and D.3 to each term.

First, $\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \geq \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)]$ follows from Lemma D.2.

Next, $\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \geq \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0\}Y_{min}$ follows from Lemma D.3.

Next, $\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] = \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(x)]$ follows from Lemma D.2.

Finally, $\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \geq \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\}Y_{min}$ follows from Lemma D.3.

As each term above is either equal to or less than or equal to their respective corresponding terms, the sum of all the components above will be less than or equal to the target estimand.

Upper Bound For the upper bound, we use the partition given by

$$
\begin{aligned}
1 = &\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \pi_e(X) \neq a_0\} + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \pi_e(X) \neq a_0\} \\
&+ \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\} + \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\} && (37)
\end{aligned}
$$

and the argument follows similarly, such that the upper bound follows from the application of Lemmas D.2 and D.3 to each term.

First, $\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \leq \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)]$ follows from Lemma D.2.

Next, $\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \leq \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \pi_e(X) \neq a_0\}Y_{max}$ follows from Lemma D.3.

Next, $\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] = \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(x)]$ follows from Lemma D.2.

Finally, $\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, A = \pi_e, M = f_M(\pi_e)] \leq \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\}Y_{max}$ follows from Lemma D.3. □

Proposition 3.3 (Bound Decomposition)**.** *The gap between the bounds in Theorem 3.1 can be written as*

$$U(\pi_e) - L(\pi_e) = \mathbb{E}[\delta(X, Y, \Pi)]$$

where

$$
\begin{aligned}
&\delta(X, Y, \Pi) = \\
&\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}(Y_{max} - Y_{min}) && (8)\\
+&\mathbf{1}\{\pi_e \neq a_0\}\big[\\
&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\} \cdot && (9)\\
&(\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)] - \mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)]) \\
+&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\} \cdot && (10)\\
&(Y_{max} - \mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)]) \\
+&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\} \cdot && (11)\\
&(\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)] - Y_{min})\big]
\end{aligned}
$$

Moreover, $\delta(X, Y, \Pi) \geq 0$ *almost surely under the assumptions of Theorem 3.1.*

Proof. We will start by observing that certain terms cancel out in the difference of the bound $U(\pi_e) - L(\pi_e)$. We begin by recalling the definition of these bounds from Theorem 3.1, rearranged slightly.

$$
\begin{aligned}
L(\pi_e) = \mathbb{E}[&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)] \\
&+ \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{min} \\
&+ \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbf{1}\{\pi_e = a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \mathbf{\Pi}^e(x)] && (38)\\
&+ \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e = a_0\}Y_{min}] && (39)\\
U(\pi_e) = \mathbb{E}[&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] \\
&+ \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{max} \\
&+ \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbf{1}\{\pi_e = a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \mathbf{\Pi}^e(x)] && (40)\\
&+ \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e = a_0\}Y_{max}] && (41)
\end{aligned}
$$

By linearity of expectation, we can remove identical terms, i.e., we can observe that in the difference $U(\pi_e) - L(\pi_e)$, the terms in Eqs. (38) and (40) cancel, leaving us with the following after collecting similar terms Eqs. (39) and (41).

$$
\begin{aligned}
&U(\pi_e) - L(\pi_e) \\
&= \mathbb{E}[\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] && (42)\\
&\quad - \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)] && (43)\\
&\quad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{max} && (44)\\
&\quad - \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{min} && (45)\\
&\quad + (\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e = a_0\})(Y_{\text{max}} - Y_{\text{min}})] && (46)
\end{aligned}
$$

Now we will conduct two splits of indicators, to reflect finer-grained subgroups.

- First, we note that we can partition the subset of $\mathcal{X}$ satisfying $\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}$ into two further subsets: The set of x where the *only* agreeing policies are those with worse performance, and the set where there also exists agreeing policies with equal or greater performance. Note that, if there exists trial policies with exactly equal performance to the new policy, both $\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}$ and $\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}$ must be true. We can argue similarly for $\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}$, and write

$$
\begin{aligned}
\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\} &= \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\} + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\} \\
\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\} &= \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\} + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}
\end{aligned}
$$

- Second, we note that we can partition the subset of $\mathcal{X}$ satisfying $\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}$ into two further subsets: The set of x where the *only* agreeing trial policies have greater performance, and the set where there are no agreeing trial policies at

all. We can argue similarly for $\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}$, and write

$$\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\} = \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\} + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}$$
$$\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\} = \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\} + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}$$

With these equalities in mind, we can rewrite the difference as follows by expanding terms

$$\begin{aligned}
&U(\pi_e) - L(\pi_e) \\
&= \mathbb{E}[\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] && \text{From Eq. (42)} && (47)\\
&\quad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] && \text{From Eq. (42)} && (48)\\
&\quad - \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)] && \text{From Eq. (43)} && (49)\\
&\quad - \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)] && \text{From Eq. (43)} && (50)\\
&\quad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{max} && \text{From Eq. (44)} && (51)\\
&\quad + \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{max} && \text{From Eq. (44)} && (52)\\
&\quad - \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{min} && \text{From Eq. (45)} && (53)\\
&\quad - \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{min} && \text{From Eq. (45)} && (54)\\
&\quad + (\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e = a_0\})(Y_{\max} - Y_{\min})] && \text{From Eq. (46)} && (55)
\end{aligned}$$

And by rearranging terms, we arrive at

$$\begin{aligned}
&U(\pi_e) - L(\pi_e) \\
&= \mathbb{E}[\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}(Y_{\max} - Y_{\min}) && \text{Eqs. (52), (54) and (55)}\\
&\quad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\} \\
&\qquad (\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] - \mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)]) && \text{Eqs. (47) and (49)}\\
&\quad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}(Y_{\max} - \mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)]) && \text{Eqs. (50) and (51)}\\
&\quad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}(\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] - Y_{\min})] && \text{Eqs. (48) and (53)}
\end{aligned}$$

After rearranging terms slightly, this gives us the desired result of the form of $\delta(X, Y, \Pi)$ stated in the theorem. The fact that $\delta(X, Y, \Pi) \geq 0$ almost surely follows from the fact that $Y_{\max} \geq \mathbb{E}[Y \mid C] \geq Y_{\min}$ for any conditioning set C according to Assumption 3.3, and the fact that $(\mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(x)] - \mathbb{E}[Y \mid X = x, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(x)])$ is nonnegative by Assumption 3.1. □

Theorem 3.2 (Tightness of bounds in Theorem 3.1)**.** *For any observational distribution $P(X, Y, A, M, \Pi, D)$ consistent with the assumptions of Theorem 3.1, there exist two structural causal models $\mathcal{M}_L, \mathcal{M}_U$ such that both are consistent with Assumptions 2.1 and 3.1 to 3.3, both give rise to that same observational distribution P, and where the policy value of any policy π_e under $\mathcal{M}_L, \mathcal{M}_U$ is given by $L(\pi_e), U(\pi_e)$ from Theorem 3.1, respectively. Hence, these bounds cannot be improved without further assumptions.*

Proof. Recall that we make use of the more stringent sets of comparison arms

$$\tilde{\mathbf{\Pi}}^e_{\leq}(x) = \underset{\pi \in \mathbf{\Pi}^e_{\leq}(x)}{\arg\max}\, f_M(\pi) \qquad \tilde{\mathbf{\Pi}}^e_{\geq}(x) = \underset{\pi \in \mathbf{\Pi}^e_{\geq}(x)}{\arg\min}\, f_M(\pi).$$

From here, we will construct a pair of SCMs $\mathcal{M}_L, \mathcal{M}_U$ that satisfy our criteria, which are defined as follows, using f^L to denote the structural equations under $\mathcal{M}_L$ and f^U to denote the structural equations under $\mathcal{M}_U$, and f (without a superscript) is used to denote structural equations that are shared between the two.

Shared Structural Equations for $\mathcal{M}_L, \mathcal{M}_U$: Both SCMs share the following equations that respect Assumptions 3.1 and 3.2, which give rise to a shared distribution over $P(X, A, M, \Pi, D)$, and these can be chosen to match any such observed distribution.

$$D = f_D(\epsilon_D), \qquad X = f_X(\epsilon_X) \qquad \Pi = \pi_D$$
$$M = f_M(\Pi), \qquad A = \Pi(X)$$

Differences between $\mathcal{M}_L, \mathcal{M}_U$: These SCMs will differ in terms of how Y is generated. Let $f_Y^L(A, X, M, \epsilon_Y)$, $f_Y^U(A, X, M, \epsilon_Y)$ denote the structural equations for $\mathcal{M}_L, \mathcal{M}_U$ respectively. We will define these functions for every possible set of inputs, using knowledge of the true conditional distribution $P(Y \mid X = x, A = a, M = m)$ wherever this combination of inputs (x, a, m) has positive density under the observed distribution $P(x, a, m) > 0$.

We will define these functions constructively, by defining their behavior for every value of (x, a, m).

1. First, fix any value of $x \in \mathcal{X}$. For this value of x, we need to define the value of f_Y^L, f_Y^U for all values of $a \in \mathcal{A}, m \in \mathbb{R}$. To do so, let

$$\mathbf{\Pi}^a(x) \coloneqq \{\pi \in \Pi \mid \pi(x) = a\}$$

be the set of all policies (possibly an empty set) that have output a on the input x, and let $\mathbf{M}_a(x) = (m_1, \ldots, m_K)$ be the (ordered) set of performance values for these policies, where m_1 denotes the worst performance $f_M(\pi_i)$ (breaking ties arbitrarily) of all policies π_i where $\pi_i \in \mathbf{\Pi}^a(x)$, and m_K denotes the best performance, where K is the size of the set $\mathbf{\Pi}^a(x)$.

2. Now we consider any arbitrary $a \in \mathcal{A}$, in addition to our fixed $x \in \mathcal{X}$. Here, there are two cases to consider:

   - If $\mathbf{\Pi}^a(x)$ is empty for this a, then for all $m \in \mathbb{R}$, we let $f_Y^L(x, a, m, \epsilon_Y) = Y_{\text{min}}$, $f_Y^U(x, a, m, \epsilon_Y) = Y_{\text{max}}$. In other words, at this point in x, if no trialed policy takes action a, we assume the worst (for f_Y^L) and the best (for f_Y^U) possible outcomes. We can easily verify that both f_Y^L, f_Y^U satisfy Assumptions 3.1 and 3.2 since the output is a constant for each function, and satisfy Assumption 3.3 since the output remains bounded.
   - If $\mathbf{\Pi}^a(x)$ (and consequently $\mathbf{M}_a(x)$) is non-empty, then we first define the behavior of f_Y^L, f_Y^U at all the (observable) performance values in $\mathbf{M}_a(x)$ to match the conditional distribution $P(Y \mid x, a, m)$.

$$f_Y^U(x, a, m, \epsilon_Y) = f_Y^L(x, a, m, \epsilon_Y) \sim P(Y \mid x, a, m), \forall m \in \mathbf{M}_a(x),$$

   with the additional constraint that f_Y^L, f_Y^U are constant with respect to m when $a = a_0$, which itself must be achievable since we assume that the true SCM generating P adheres to this constraint by Assumption 3.2. Note that we can always achieve the equivalence of distribution shown above by taking ϵ_Y to be a uniform random variable in $[0, 1]$, and defining our function as sampling from $P(Y \mid x, a, m)$ using the inverse CDF $f_Y(x, a, m, \epsilon_Y) = F^{-1}_{Y|x,a,m}(\epsilon_Y)$. Because we assume that $P(Y \mid x, a, m)$ does not violate our assumptions, it should be clear that $f_Y^L(x, a, m, \epsilon_Y), f_Y^U(x, a, m, \epsilon_Y)$ do not violate our assumptions for values of $m \in \mathbf{M}_a(x)$. In addition, we have that for any $m \notin \mathcal{M}(x)$, our construction above does not violate Assumption 3.2 (since in this case, $f_Y^L(x, a_0, m, \epsilon_Y)$ is constant for all values of m).
   We have now defined the behavior of f_Y^L, f_Y^U when $a = a_0$, and when $a \neq a_0, m \in \mathbf{M}_a(x)$. Now it remains to define the behavior of f_Y^L, f_Y^U for other values of m when $a \neq a_0$. For any value $m' \notin \mathbf{M}_a(x)$, there are three possible scenarios: It is smaller than the smallest value (m_1), larger than the largest value m_K, or in-between two values, which we denote $h_{\text{prev}}(m') < m' < h_{\text{next}}(m')$ without loss of generality, where $h_{\text{prev}}(m') = \max(m \in \mathbf{M}_a(x) \mid m < m')$ and $h_{\text{next}}(m') = \min(m \in \mathbf{M}_a \mid m > m')$. Here, $h_{\text{prev}}(m')$ corresponds to the performance of the "next-worst" policy among those deployed, and $h_{\text{next}}(m')$ corresponds to the performance of the "next-best" policy. We define behavior on these sets as follows

$$f_Y^L(x, a, m', \epsilon_Y) = \begin{cases} Y_{\text{min}}, & \text{if } m' < \min(\mathbf{M}_a(x)) \\ f_Y^L(x, a, h_{\text{prev}}(m'), \epsilon_Y) & \text{if } m' > \min(\mathbf{M}_a(x)), m \notin \mathbf{M}_a(x) \end{cases}$$

$$f_Y^U(x, a, m', \epsilon_Y) = \begin{cases} Y_{\text{max}}, & \text{if } m' > \max(\mathbf{M}_a(x)) \\ f_Y^U(x, a, h_{\text{next}}(m'), \epsilon_Y) & \text{if } m' < \max(\mathbf{M}_a(x)), m \notin \mathbf{M}_a(x) \end{cases}$$

   In words, we have "filled in" the missing gaps in f_Y^L, f_Y^U for all values of m using piecewise constant functions: For any $m' \notin \mathbf{M}_a(x)$, if m' is worse than any observed performance, we assume the worst-case for the lower bound, and if m' is better than any observed performance, we assume the best-case for the upper bound. Otherwise, we have $h_{\text{prev}}(m') < m'$ and/or $m' < h_{\text{next}}(m')$, and we assume for the lower bound that the outcomes at m' match those at $h_{\text{prev}}(m')$, and for the upper bound we assume the outcomes at m' match that at $h_{\text{next}}(m')$. Because we have maintained monotonicity with respect to m, our construction continues to satisfy our core assumptions.

3. We have now fully defined f_Y^L, f_Y^U, having defined these functions for any input (x, a, m), and shown that they satisfy

our core assumptions Assumptions 3.1 to 3.3. Putting it together, we have it that

$$f_Y^L(x,a,m,\epsilon_Y) = \begin{cases} Y_{\min}, & \text{if } \mathbf{\Pi}^a(x) = \varnothing, \\ Y_{\min}, & \text{if } \mathbf{\Pi}^a(x) \neq \varnothing, m < \min(\mathbf{M}_a(x)) \\ F^{-1}_{Y|x,a,h_{\text{prev}}(m)}(\epsilon_Y), & \text{if } \mathbf{\Pi}^a(x) \neq \varnothing, m > \min(\mathbf{M}_a(x)), m \notin \mathbf{M}_a(x), \\ F^{-1}_{Y|x,a,m}(\epsilon_Y), & \text{if } \mathbf{\Pi}^a(x) \neq \varnothing, m \in \mathbf{M}_a(x), \end{cases} \tag{56}$$

$$f_Y^U(x,a,m,\epsilon_Y) = \begin{cases} Y_{\max}, & \text{if } \mathbf{\Pi}^a(x) = \varnothing, \\ Y_{\max}, & \text{if } \mathbf{\Pi}^a(x) \neq \varnothing, m > \max(\mathbf{M}_a(x)) \\ F^{-1}_{Y|x,a,h_{\text{next}}(m)}(\epsilon_Y), & \text{if } \mathbf{\Pi}^a(x) \neq \varnothing, m < \max(\mathbf{M}_a(x)), m \notin \mathbf{M}_a(x), \\ F^{-1}_{Y|x,a,m}(\epsilon_Y), & \text{if } \mathbf{\Pi}^a(x) \neq \varnothing, m \in \mathbf{M}_a(x), \end{cases} \tag{57}$$

where $F^{-1}_{Y|x,a,m}$ is the inverse conditional CDF of Y given X, A, M, derived from P, and where $\mathbf{\Pi}^a(x) \coloneqq \{\pi \in \Pi \mid \pi(x) = a\}$ and $\mathbf{M}_a(x) \coloneqq \{f_M(\pi) : \pi \in \mathbf{\Pi}^a(x)\}$, as defined previously above.

Verifying conditions We have now defined the SCMs $\mathcal{M}_L, \mathcal{M}_U$, and shown that these SCMs are both consistent with our assumptions. We will now briefly verify that both SCMs give rise to the same observed distribution $P(X, Y, A, M, \Pi, D)$, and then demonstrate that these SCMs achieve the upper and lower bounds that are given in Theorem 3.1.

First, we have it by construction that both SCMs yield the observed distribution $P(X, Y, A, M, \Pi, D)$, so it remains to demonstrate that they agree with the observed distribution $P(Y \mid X, A, M, D, \Pi)$, which we can write equivalently as $P(Y \mid X, A, M)$, since $D, \Pi \perp\!\!\!\perp Y \mid X, A, M$ under our assumed data-generating process. Note that $P(Y \mid X, A, M, \Pi, D)$ is only well-defined for x, a, m with positive density (if X is continuous) or probability mass (if X is discrete). Assuming that $P(x) > 0$ for all $x \in \mathcal{X}$, we have constructed f_Y^L, f_Y^U to agree with $P(Y \mid x, a, m)$ for all a, m where there exists a trialed policy π that outputs $a = \pi(x)$ with performance $m = f_M(\pi)$. We note that for any other value of a', m', we have it that $P(a', m' \mid x) = 0$, and hence the entire set (x, a', m') has zero density, and it is precisely on these never-observed subsets of inputs where $\mathcal{M}_L, \mathcal{M}_U$ disagree.

Second, we can verify that the policy values under f_Y^L and f_Y^U evaluate to $L(\pi_e)$ and $U(\pi_e)$, respectively. Recalling that $Y(A = \pi_e, M = f_M(\pi_e)) = f_Y(X, \pi_e(X), f_M(\pi_e), \epsilon_Y)$, and recalling Eq. (32), we can write that under $\mathcal{M}_L$,

$$\begin{aligned} \mathbb{E}_{\mathcal{M}_L}[Y(A = \pi_e, M = f_M(\pi_e))] = \mathbb{E}\big[&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\} f_Y^L(X, \pi_e(X), f_M(\pi_e), \epsilon_Y) && (58) \\ &+ \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0\} f_Y^L(X, \pi_e(X), f_M(\pi_e), \epsilon_Y) && (59) \\ &+ \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\} f_Y^L(X, \pi_e(X), f_M(\pi_e), \epsilon_Y) && (60) \\ &+ \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\} f_Y^L(X, \pi_e(X), f_M(\pi_e), \epsilon_Y)] && (61) \end{aligned}$$

where we can consider each component in the sum individually by linearity of expectation, and the fact that $\mathbb{E}[\mathbf{1}\{x \in \Omega\} f(x, \epsilon_Y)] = \mathbb{E}[\mathbf{1}\{x \in \Omega\} \mathbb{E}[f(x, \epsilon_Y) \mid x \in \Omega]]$ for any set Ω. We consider each term under the definition of f_Y^L in Eq. (56).

- For Eq. (58), we can observe that for all $x \in \mathcal{X}$ satisfying $\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, a = \pi_e(x)$, the set $\mathbf{\Pi}^a(x)$ is non-empty by definition of $\tilde{\mathbf{\Pi}}^e_{\leq}(X)$, and moreover that $m_e \geq \min(\mathbf{M}_a(x))$. As a result, we have it that $f_Y^L(x, a, m_e, \epsilon_Y) = F^{-1}_{Y|x,a,h_{\text{prev}}(m_e)}(\epsilon_Y) \sim P(Y \mid x, a, h_{\text{prev}}(m_e))$, and thus this term can be re-written as

$$\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\} \mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)] \tag{62}$$

  since for any $\pi_i \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)$, we have it that $\pi_i(X) = \pi_e(X)$ and $f_M(\pi_i) = h_{\text{prev}}(m_e)$ by definition of $h_{\text{prev}}(m_e)$ and $\tilde{\mathbf{\Pi}}^e_{\leq}(X)$.

- For Eq. (59), we can observe that for all $x \in \mathcal{X}$ satisfying $\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, a = \pi_e(x)$, the set $\mathbf{\Pi}^a(x)$ is either empty, or non-empty where $f_M(\pi_e) < \min(\mathbf{M}_a(x))$, by definition of $\tilde{\mathbf{\Pi}}^e_{\leq}(X)$. In either case, we have it that $f_Y^L = Y_{\min}$, and so this term is equal to

$$\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0\} Y_{\min} \tag{63}$$

- For Eq. (60), we can observe that for all $x \in \mathcal{X}$ satisfying $\mathbf{\Pi}^e(X) \neq \varnothing, a = \pi_e(x) = a_0$, the set $\mathbf{\Pi}^a(x)$ is non-empty by definition of $\mathbf{\Pi}^e(X)$, and moreover that f_Y^L is invariant to the choice of m, and so this term is equal to $\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, A = \pi_e(X), M = f_M(\pi_e)]$, which is equal to
$$\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)] \tag{64}$$
from Eq. (24) of Lemma D.2
- For Eq. (61), we can observe that for all $x \in \mathcal{X}$ satisfying $\mathbf{\Pi}^e(X) = \varnothing$, the set $\mathbf{\Pi}^a(x)$ is empty by definition of $\mathbf{\Pi}^e(X)$, and so $f_Y^L = Y_{\min}$. Thus, this term is equal to
$$\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\}Y_{\min} \tag{65}$$

Collecting terms Eqs. (62) to (65) gives us that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{M}_L}[Y(A = \pi_e, M = f_M(\pi_e))] \\
&\quad = \mathbb{E}\big[\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)] \\
&\qquad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0\}Y_{\min} \\
&\qquad + \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)] \\
&\qquad + \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0\}Y_{\min}\big]
\end{aligned}$$

which is equivalent to $L(\pi_e)$ and completes the proof for the lower bound. For the upper bound, the argument is similar, and roughly symmetric, but uses the partition given by Eq. (37).

Conclusion Because it is possible to construct structural causal models that are consistent with our assumptions and that have counterfactual policy values that are exactly $L(\pi_e)$ and $U(\pi_e)$, the bounds in Theorem 3.1 cannot be improved without further assumptions. □

Proposition 3.4. *The bounds in Theorem 3.1 can be written* $L(\pi_e) = \mathbb{E}[\psi_L(Y, X, \Pi)]$ *and* $U(\pi_e) = \mathbb{E}[\psi_U(Y, X, \Pi)]$, *where* ψ_L *and* ψ_U *are defined as follows*

$$\begin{aligned}
&\psi_L(Y, X, \Pi) \\
&\coloneqq \begin{cases}
Y \cdot \frac{\mathbf{1}\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)\}}{P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X))}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0 \\
Y_{min}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0 \\
Y \cdot \frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}^e(X)\}}{P(\Pi \in \mathbf{\Pi}^e(X))}, & \text{if } \mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0 \\
Y_{min}, & \text{if } \mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0
\end{cases} \\
&\psi_U(Y, X, \Pi) \\
&\coloneqq \begin{cases}
Y \cdot \frac{\mathbf{1}\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)\}}{P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X))}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \pi_e(X) \neq a_0 \\
Y_{max}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \pi_e(X) \neq a_0 \\
Y \cdot \frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}^e(X)\}}{P(\Pi \in \mathbf{\Pi}^e(X))}, & \text{if } \mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0 \\
Y_{max}, & \text{if } \mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0
\end{cases}
\end{aligned}$$

Moreover, since ψ_L, ψ_U *are known functions of the data, these bounds can be estimated as*

$$\hat{L}(\pi_e) \coloneqq n^{-1} \sum_i \psi_L(Y_i, X_i, \Pi_i)$$

$$\hat{U}(\pi_e) \coloneqq n^{-1} \sum_i \psi_U(Y_i, X_i, \Pi_i)$$

where $\sqrt{n}(L - \hat{L}) \xrightarrow{d} N(0, \sigma^2(\psi_L))$ *where* $\sigma^2(\psi_L)$ *is the variance of* ψ_L *and* $\xrightarrow{d}$ *denotes convergence in distribution, with similar convergence of* $\hat{U}$, *and hence*

$$\left[\hat{L}(\pi_e) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\hat{\sigma}(\psi_L)}{\sqrt{n}}, \right.$$
$$\left. \hat{U}(\pi_e) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\hat{\sigma}(\psi_U)}{\sqrt{n}}\right]$$

is an asymptotically valid $(1-\alpha)$-confidence interval, where $\hat{\sigma}(\psi)$ is the empirical standard deviation of ψ and Φ^{-1} is the inverse of the standard normal CDF.

Proof. First, note that the conditional expectation of Y is always finite due to Assumption 3.3.

Lemma D.4. *Let $\mathbf{\Pi}'(X)$ be a function that maps from $\mathcal{X}$ to any subset (including the empty set) of Π, and let $\mathcal{X}'$ be a subset of $\mathcal{X}$. If $P(\Pi \in \mathbf{\Pi}'(x)) > 0, \forall x \in \mathcal{X}'$, then*

$$\mathbb{E}\left[Y\frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}'(X)\}}{P(\Pi \in \mathbf{\Pi}'(X))}\mathbf{1}\{X \in \mathcal{X}'\}\right] = \mathbb{E}[\mathbb{E}[Y \mid \Pi \in \mathbf{\Pi}'(X), X]\mathbf{1}\{X \in \mathcal{X}'\}] \tag{66}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[Y\frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}'(X)\}}{P(\Pi \in \mathbf{\Pi}'(X))}\mathbf{1}\{X \in \mathcal{X}'\}\right] &= \mathbb{E}\left[\mathbb{E}[Y\mathbf{1}\{\Pi \in \mathbf{\Pi}'(X)\} \mid X]\frac{\mathbf{1}\{X \in \mathcal{X}'\}}{P(\Pi \in \mathbf{\Pi}'(X))}\right] && (67)\\
&= \mathbb{E}\left[\mathbb{E}[Y \mid \Pi \in \mathbf{\Pi}'(X), X]P(\Pi \in \mathbf{\Pi}'(X) \mid X)\frac{\mathbf{1}\{X \in \mathcal{X}'\}}{P(\Pi \in \mathbf{\Pi}'(X))}\right] && (68)\\
&= \mathbb{E}[\mathbb{E}[Y \mid \Pi \in \mathbf{\Pi}'(X), X]\mathbf{1}\{X \in \mathcal{X}'\}] && (69)
\end{aligned}$$

where the first equality is well-defined on both sides by the assumption that for any $X \in \mathcal{X}'$, $P(\Pi \in \mathbf{\Pi}'(X)) > 0$. For the second-to-last line, note that this follows from the basic fact that A, B, C

$$\begin{aligned}
\mathbb{E}[A \cdot \mathbf{1}\{B \in \mathcal{B}\} \mid C] &= \mathbb{E}[A \cdot \mathbf{1}\{B \in \mathcal{B}\} \mid B \in \mathcal{B}, C]P(B \in \mathcal{B} \mid C)\\
&\quad + \mathbb{E}[A \cdot \mathbf{1}\{B \in \mathcal{B}\} \mid B \notin \mathcal{B}, C]P(B \notin \mathcal{B} \mid C)\\
&= \mathbb{E}[A \mid B \in \mathcal{B}, C]P(B \in \mathcal{B} \mid C)
\end{aligned}$$

and the last line follows from the fact that $\Pi \perp\!\!\!\perp X$ under Assumption 2.1, so that $P(\Pi \in \mathbf{\Pi}'(X) \mid X) = P(\Pi \in \mathbf{\Pi}'(X))$. □

Note that Lemma D.4 applies to all of the pairs (e.g., $\mathbf{\Pi}^e(X) \neq \varnothing$ and $\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0\}$) used in Proposition 3.4. Thus, we can directly write the following through linearity of expectations and two applications of Lemma D.4.

$$\begin{aligned}
L(\pi_e) &= \mathbb{E}[\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)]\\
&\quad + \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{min}\\
&\quad + \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbf{1}\{\pi_e = a_0\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)]\\
&\quad + \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e = a_0\}Y_{min}]\\
&= \mathbb{E}[Y\frac{\mathbf{1}\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)\}}{P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X))}\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\\
&\quad + Y_{min}\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\\
&\quad + \frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}^e(X)\}}{P(\Pi \in \mathbf{\Pi}^e(X))}\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbf{1}\{\pi_e = a_0\}\\
&\quad + Y_{min}\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e = a_0\}]
\end{aligned}$$

We apply Lemma D.4 in two instances: one where the indicator function $\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}$ is turned on and one where the indicator function $\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}$ is turned on. In the case where $\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}$ is true, $P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)) > 0$ is also true because there is at least one trial policy in the set $\tilde{\mathbf{\Pi}}^e_{\leq}(X)$. Similarly, when $\{\mathbf{\Pi}^e(X) \neq \varnothing\}$ is satisfied, $P(\Pi \in \mathbf{\Pi}^e(X)) > 0$ will also be satisfied. Thus, when applying Lemma D.4, the assumption required in the lemma that $P(\Pi \in \mathbf{\Pi}'(x)) > 0$ is satisfied by the decomposition of the lower bound, and we do not require any additional assumptions regarding the probability of deploying a particular set of trial models.

Because the sets $(\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e \neq a_0)$, $(\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e \neq a_0)$, $(\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e = a_0)$, and $(\mathbf{\Pi}^e(X) = \varnothing, \pi_e = a_0)$ are disjoint, only one product of the indicator functions inside the expectation above ever evaluates to 1. Thus, we can equivalently express the expression inside the expectation above as the piecewise function

$$\psi_L(Y, X, \Pi) \coloneqq \begin{cases} Y \cdot \frac{\mathbf{1}\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)\}}{P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X))}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing, \pi_e(X) \neq a_0 \\ Y_{\text{min}}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing, \pi_e(X) \neq a_0 \\ Y \cdot \frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}^e(X)\}}{P(\Pi \in \mathbf{\Pi}^e(X))}, & \text{if } \mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0 \\ Y_{\text{min}}, & \text{if } \mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0 \end{cases}$$

Thus, $L(\pi_e) = \mathbb{E}[\psi_L(Y, X, \Pi)]$.

The proof for $U(\pi_e)$ follows similarly. We directly write the following through linearity of expectations and two applications of Lemma D.4.

$$\begin{aligned} U(\pi_e) = \mathbb{E}[&\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)] \\ &+ \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\}Y_{max} \\ &+ \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbf{1}\{\pi_e = a_0\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)] \\ &+ \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e = a_0\}Y_{max}] \\ = \mathbb{E}[&Y \frac{\mathbf{1}\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)\}}{P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X))}\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing\}\mathbf{1}\{\pi_e \neq a_0\} \\ &+ Y_{max}\mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing\}\mathbf{1}\{\pi_e \neq a_0\} \\ &+ \frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}^e(X)\}}{P(\Pi \in \mathbf{\Pi}^e(X))}\mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbf{1}\{\pi_e = a_0\} \\ &+ Y_{max}\mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}\mathbf{1}\{\pi_e = a_0\}] \end{aligned}$$

Because this different enumeration of the sets $(\tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \pi_e \neq a_0)$, $(\tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \pi_e \neq a_0)$, $(\mathbf{\Pi}^e(X) \neq \varnothing, \pi_e = a_0)$, and $(\mathbf{\Pi}^e(X) = \varnothing, \pi_e = a_0)$ is also disjoint, only one product of the indicator functions inside the expectation above ever evaluates to 1. Thus, we can equivalently express the expression inside the expectation above as the piecewise function

$$\psi_U(Y, X, \Pi) \coloneqq \begin{cases} Y \cdot \frac{\mathbf{1}\{\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X)\}}{P(\Pi \in \tilde{\mathbf{\Pi}}^e_{\geq}(X))}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\geq}(X) \neq \varnothing, \pi_e(X) \neq a_0 \\ Y_{\text{max}}, & \text{if } \tilde{\mathbf{\Pi}}^e_{\geq}(X) = \varnothing, \pi_e(X) \neq a_0 \\ Y \cdot \frac{\mathbf{1}\{\Pi \in \mathbf{\Pi}^e(X)\}}{P(\Pi \in \mathbf{\Pi}^e(X))}, & \text{if } \mathbf{\Pi}^e(X) \neq \varnothing, \pi_e(X) = a_0 \\ Y_{\text{max}}, & \text{if } \mathbf{\Pi}^e(X) = \varnothing, \pi_e(X) = a_0 \end{cases}$$

Thus, $U(\pi_e) = \mathbb{E}[\psi_U(Y, X, \Pi)]$.

To show asymptotic normality, it suffices to observe that ψ_U, ψ_L are known functions of the data, such that the problem reduces to mean estimation using samples. The asymptotic behavior is then just a consequence of the central limit theorem [Vaart, 1998], and the validity of the confidence intervals follows from the fact that we use the $1 - \alpha/2$ lower bound for L, such that the probability of failing to cover L is asymptotically $1 - \alpha/2$, and similarly the $1 - \alpha/2$ upper bound for U. The validity of the given interval follows from application of the union bound. □