# Just Trial Once: Ongoing Causal Validation of Machine Learning Models

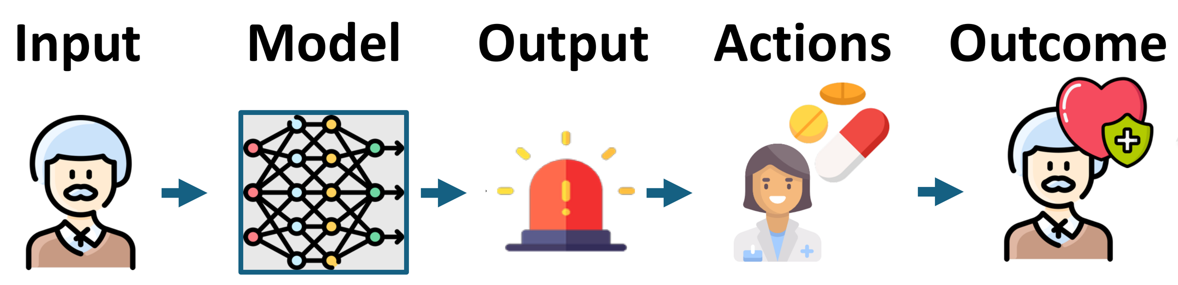
Jacob M. Chen†, Michael Oberst†
jchen459@jhu.edu, moberst@jhu.edu
†Dept. of Computer Science, Johns Hopkins University

Link to full paper!

JOHNS HOPKINS UNIVERSITY

## Motivation

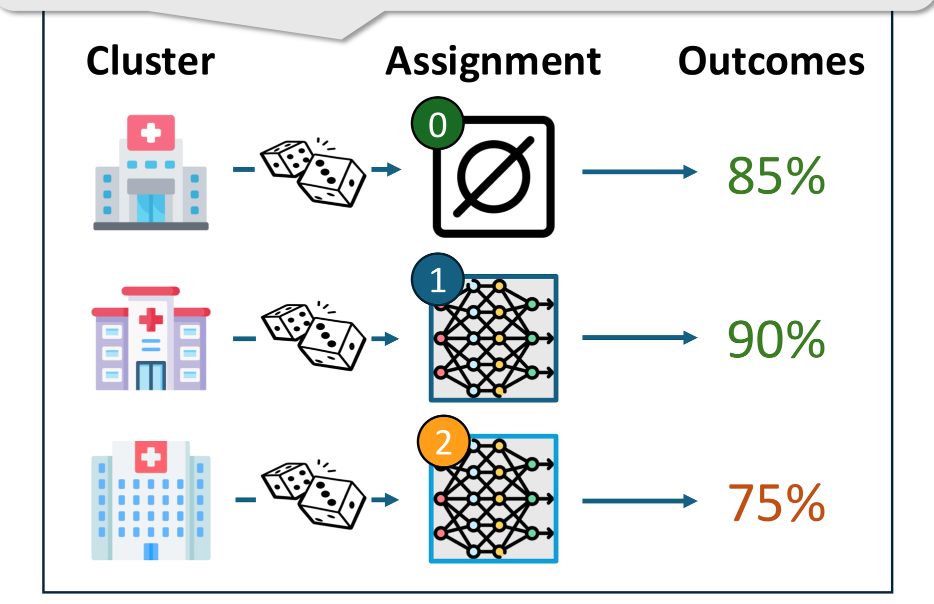
1. Machine learning (ML) models are increasingly deployed in high-risk domains like healthcare and criminal justice as tools to support human decision-makers [1, 2].
2. The randomized controlled trial (RCT) framework is not designed for ML-enabled systems, which (unlike drugs) are often updated frequently to handle performance degradation [3].
3. We give 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.



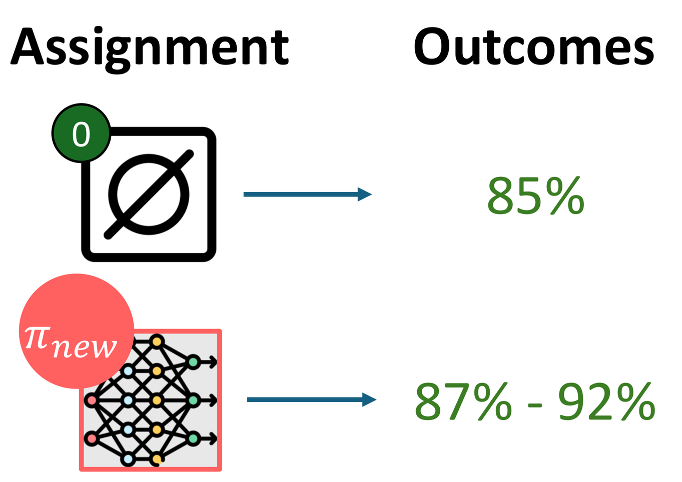

What is the **causal effect** of deploying this model on **patient outcomes?**

## Our Goal: Just Trial Once

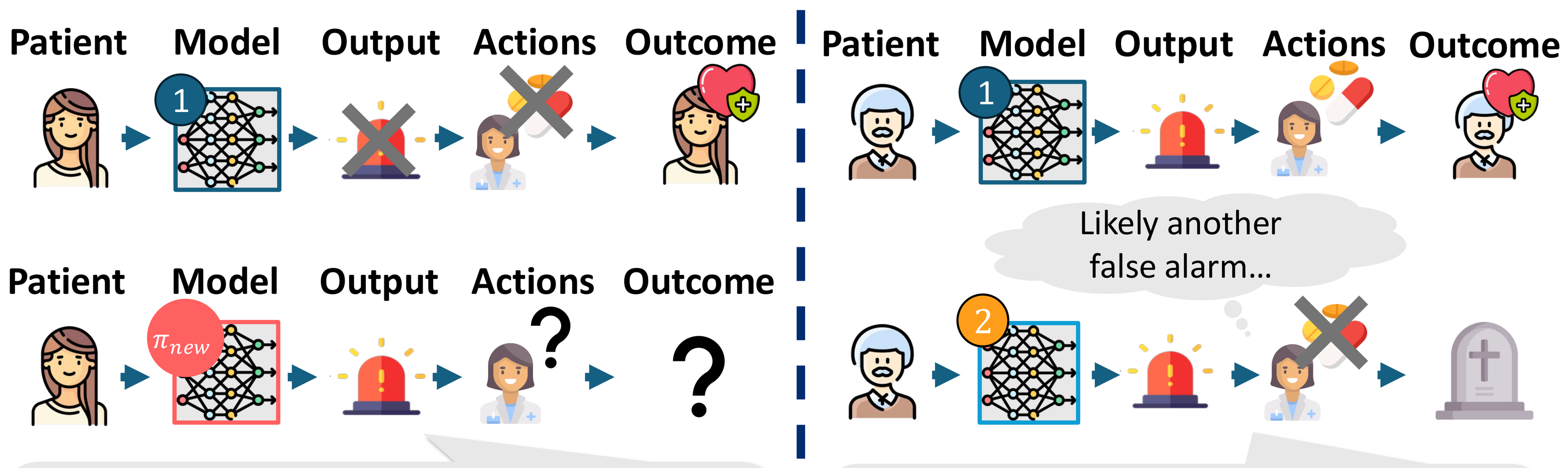
Given **data from a cluster RCT** where multiple models were trialed with varying outcomes (e.g. survival rate)…




…**bound the causal effect** of deploying a new, never-deployed model between a lower bound (*L*) and upper bound (*U*).




## Two Challenges: Coverage and Trust



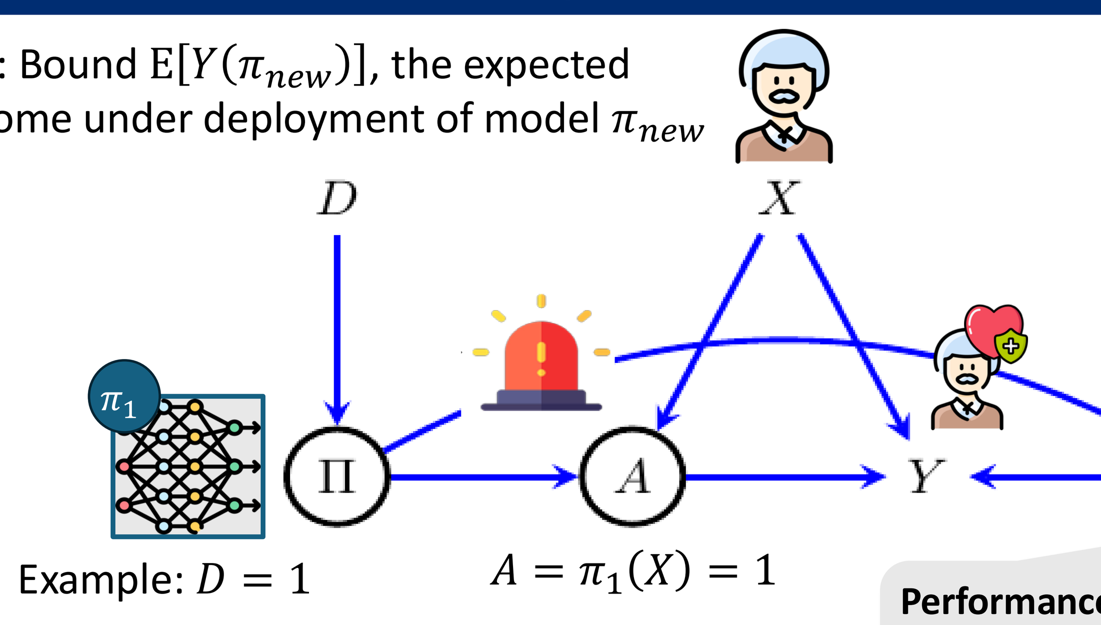

**Our data is unlikely to cover all possible outputs**: If our new model raises alerts on patients who never received alerts in the trial, we lack data on what would occur to these types of patients!

**Causal impact depends on clinician trust**: Even for the same patient with the same model output, outcomes could differ depending on perceived model reliability and performance.

## Model and Problem Setup

**Goal**: Bound $\mathbb{E}[Y(\pi_{new})]$, the expected outcome under deployment of model $\pi_{new}$

$D$: Indicator of trial arm
$\Pi$: Deployed model/policy
$A$: Model Output
$X$: Model Inputs
$Y$: Clinical Outcome
$M$: Model performance metric




**Performance Assumption:** There exists a computable metric that captures overall model performance / trust (e.g., false alarm rate).

**Assumed causal data-generating process**

## Evaluating a New Model: Falsifiable Assumptions

**Assumption 1: Performance Monotonicity** — *Has a falsification test: See Proposition 3.1*

Potential outcomes are non-decreasing in model performance metric, 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 2: Neutral Actions** — *Has a falsification test: See Proposition 3.2*

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 metric $M$. That is, for any two values $m_i \neq m_j$,

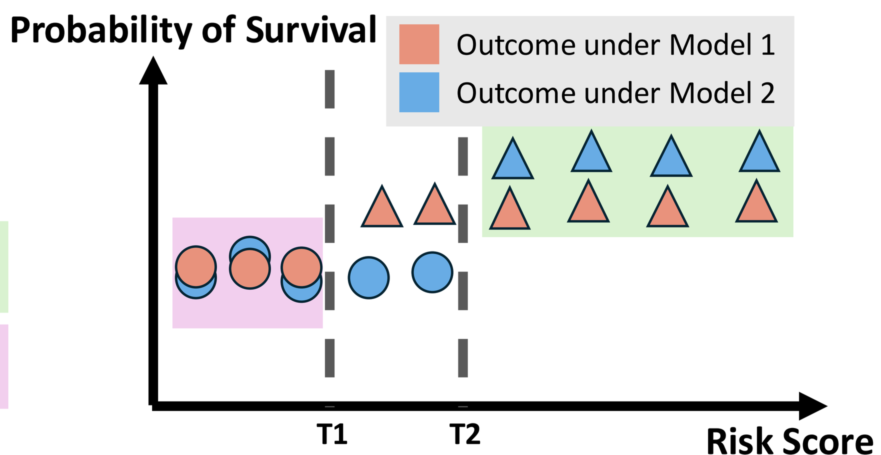
$$Y(A = a_0, M = m_i) = Y(A = a_0\ M = m_j)$$

**Assumption 3: Bounded Outcomes**

There exists constants $Y_{min}$ and $Y_{max}$ such that $Y_{min} \leq \text{Y} \leq Y_{max}$.

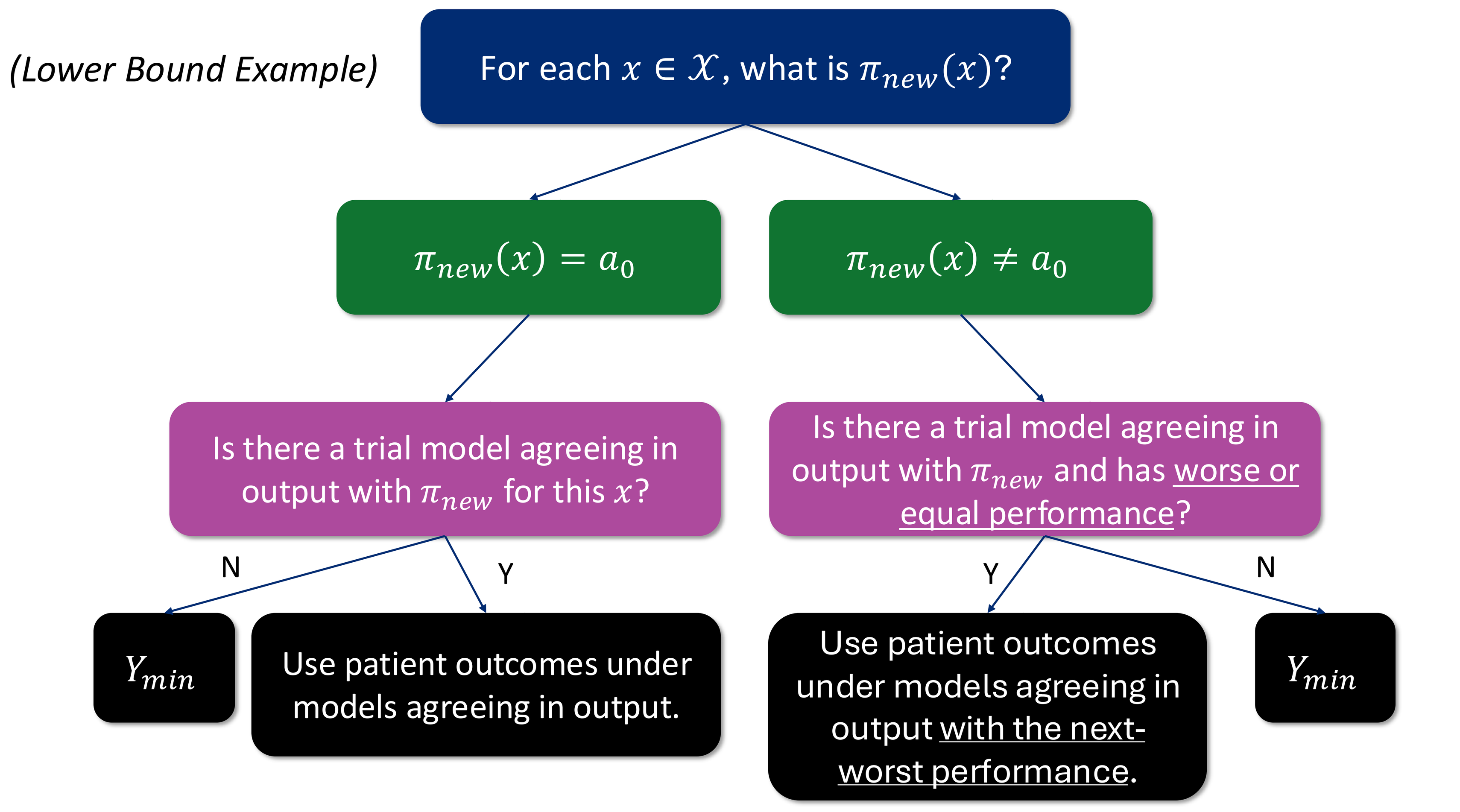
**Falsification in Practice**

- Model 2 has a greater performance metric than Model 1. Model 1 alerts above T1, and Model 2 alerts above T2. Triangles are alerts (A=1), and circles are no alerts (A=0).
- To falsify Assumption 1, compare the triangles in the green box.
- To falsify Assumption 2, compare the circles in the purple box.




## Intuition: Lower / Upper Bounds on Causal Impact

*(Lower Bound Example)*




The construction of the upper bound follows similarly. We show that **these bounds are tight** (*Theorem 3.2*) and give **inverse probability weighted-style estimators** with their corresponding confidence intervals (*Proposition 3.4*).

*More Details on Bound Construction (if desired).*

**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].

$$\mathbf{\Pi}^e(x) := \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x)\} \quad (1)$$
$$\mathbf{\Pi}^e_{\leq}(x) := \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x), f_M(\pi) \leq f_M(\pi_e)\} \quad (2)$$
$$\mathbf{\Pi}^e_{\geq}(x) := \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x), f_M(\pi) \geq f_M(\pi_e)\} \quad (3)$$

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].

$$\tilde{\mathbf{\Pi}}^e_{\leq}(x) := \arg\max_{\pi \in \mathbf{\Pi}^e_{\leq}(x)} f_M(\pi),$$
$$\tilde{\mathbf{\Pi}}^e_{\geq}(x) := \arg\min_{\pi \in \mathbf{\Pi}^e_{\geq}(x)} f_M(\pi)$$

1. Subset of models that agree in output with the new model for a patient $x$.
2. Subset of models that agree in output with the new model for a patient $x$ AND has equal or worse performance.
3. Subset of models that agree in output with the new model for a patient $x$ AND has equal or better performance.

**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 $\pi_e$ 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*

$$\begin{aligned} L(\pi_e) = \mathbb{E}\Big[&\mathbf{1}\{\pi_e \neq a_0\}\big( \\ &\quad \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \tilde{\mathbf{\Pi}}^e_{\leq}(X)] \quad (1) \\ &+ \mathbf{1}\{\tilde{\mathbf{\Pi}}^e_{\leq}(X) = \varnothing\}Y_{min}\big) \quad (2) \\ &+ \mathbf{1}\{\pi_e = a_0\}\big( \\ &\quad \mathbf{1}\{\mathbf{\Pi}^e(X) \neq \varnothing\}\mathbb{E}[Y \mid X, \Pi \in \mathbf{\Pi}^e(X)] \quad (3) \\ &+ \mathbf{1}\{\mathbf{\Pi}^e(X) = \varnothing\}Y_{min}\big)\Big] \quad (4) \end{aligned}$$

1. When the output is not a neutral action and there exists at least one agreeing model with worse or equal performance, use outcomes under the next-worst deployed model as the lower bound.

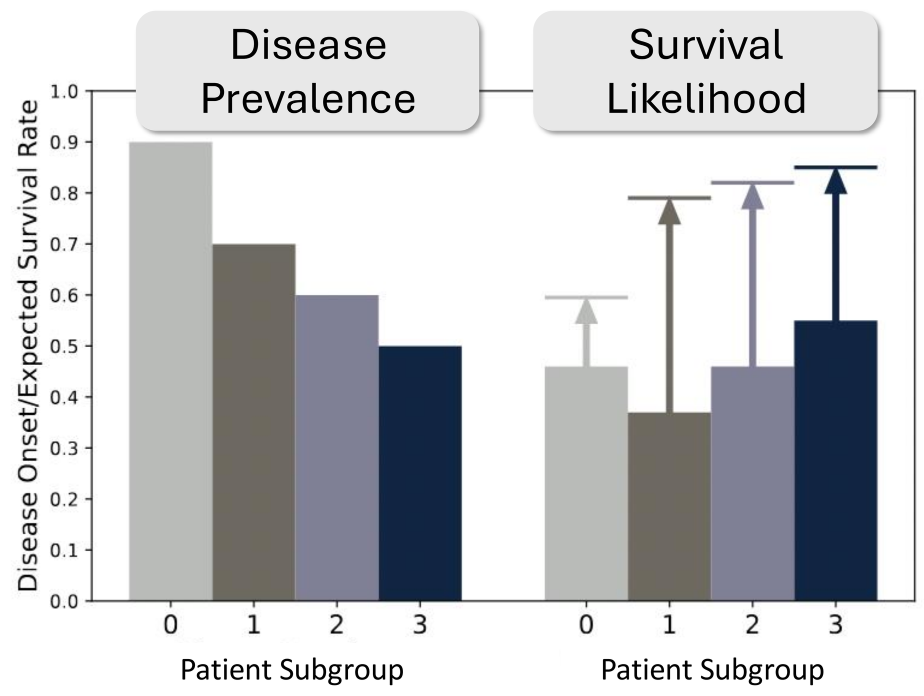
3. When the output is a neutral action and there exists at least one agreeing model, use outcomes under agreeing models as the lower bound.

2. 4. Otherwise, lower bound by the lowest possible value of the outcome.

## Simulation Study

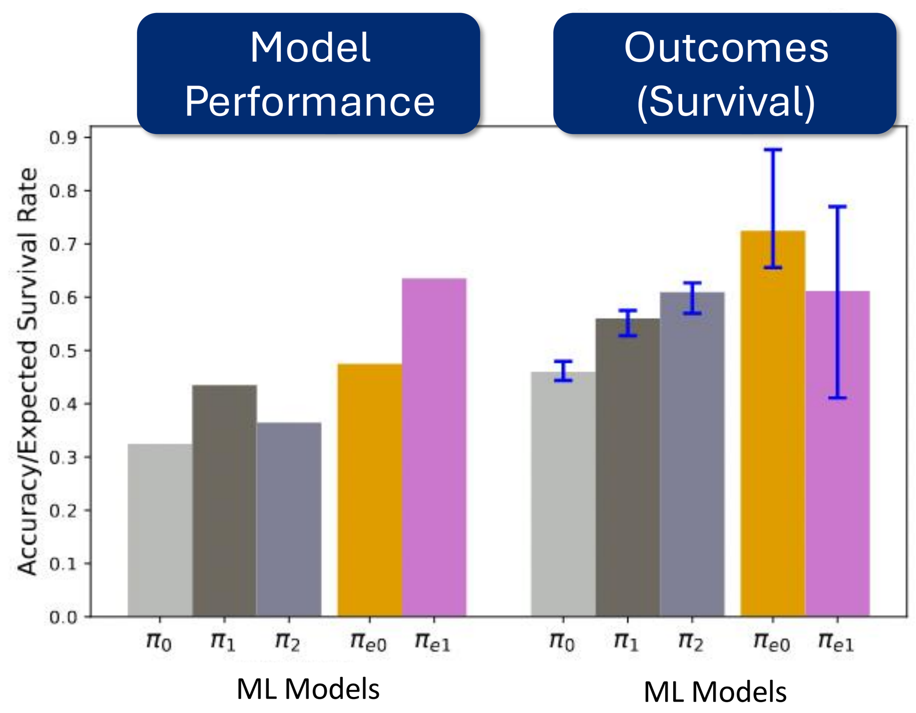
**Setup**

- Define four types of patients with varying likelihoods of developing disease and varying survival rates.
- Raising alerts on the highest-risk ("most obvious", X=0) patients is less helpful than raising alerts on other patients.




**Results**

- Model performance is the raw accuracy of the model in predicting disease onset.
- Bars indicate ground truth, and intervals indicate our bounds (incl. statistical uncertainty).
- Note: Model accuracy is not indicative of causal impact.




## References

[1] R. Adams, K. E. Henry, A. Sridharan, H. Soleimani, A. Zhan, N. Rawat, L. Johnson, D. N. Hager, S. E. Cosgrove, A. Markowski, et al. 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, 2022.
[2] S. K. Gohil, E. Septimus, K. Kleinman, N. Varma, T. R. Avery, L. Heim, R. Rahm, W. S. Cooper, M. Cooper, L. E. McLean, N. G. Nickolay, R. A. Weinstein, L. H. Burgess, et al. Stewardship prompts to improve antibiotic selection for pneumonia. JAMA, 331:2007, 6 2024.
[3] D. Ouyang and J. Hogan. We need more randomized clinical trials of AI, 2024.