Just Trial Once: Ongoing Causal Validation of Machine Learning Models

Jacob M Chen* and Michael Oberst*

*Department of Computer Science, Johns Hopkins University

Overview

1. Introduce the problem and its key challenges.
2. Introduce a formal setup for approaching the problem.
3. State assumptions that address the key challenges.
4. Bound the causal effect of interest.

Overview

1. Introduce the problem and its key challenges.

2. Introduce a formal setup for approaching the problem.

3. State assumptions that address the key challenges.

4. Bound the causal effect of interest.

Overview

1. Introduce the problem and its key challenges.
2. Introduce a formal setup for approaching the problem.
3. State assumptions that address the key challenges.
4. Bound the causal effect of interest.

Overview

1. Introduce the problem and its key challenges.

2. Introduce a formal setup for approaching the problem.

3. State assumptions that address the key challenges.

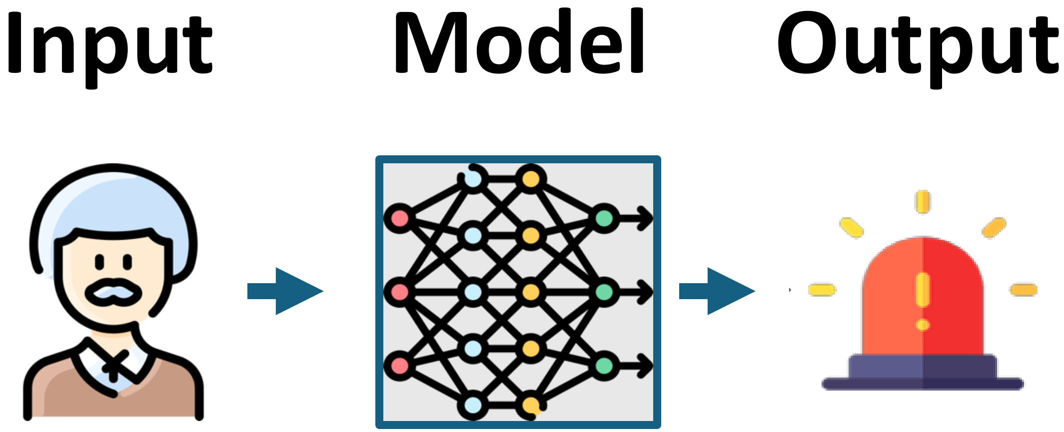
4. Bound the causal effect of interest.

Overview

1. Introduce the problem and its key challenges.
2. Introduce a formal setup for approaching the problem.
3. State assumptions that address the key challenges.
4. Bound the causal effect of interest.

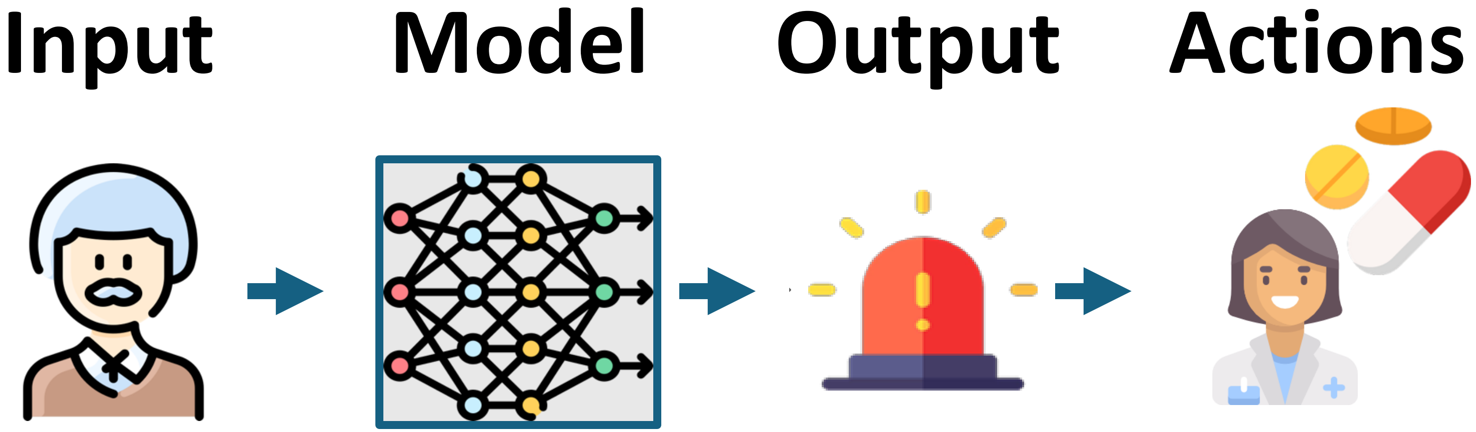
What is causal validation of an ML model?

What is causal validation of an ML model?



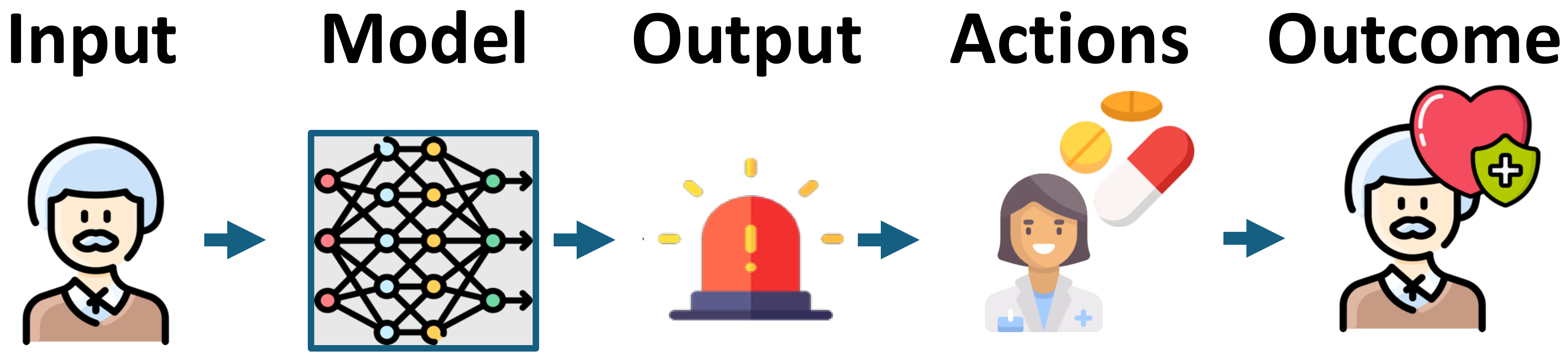
1. ML model produces an output based on patient information.

What is causal validation of an ML model?



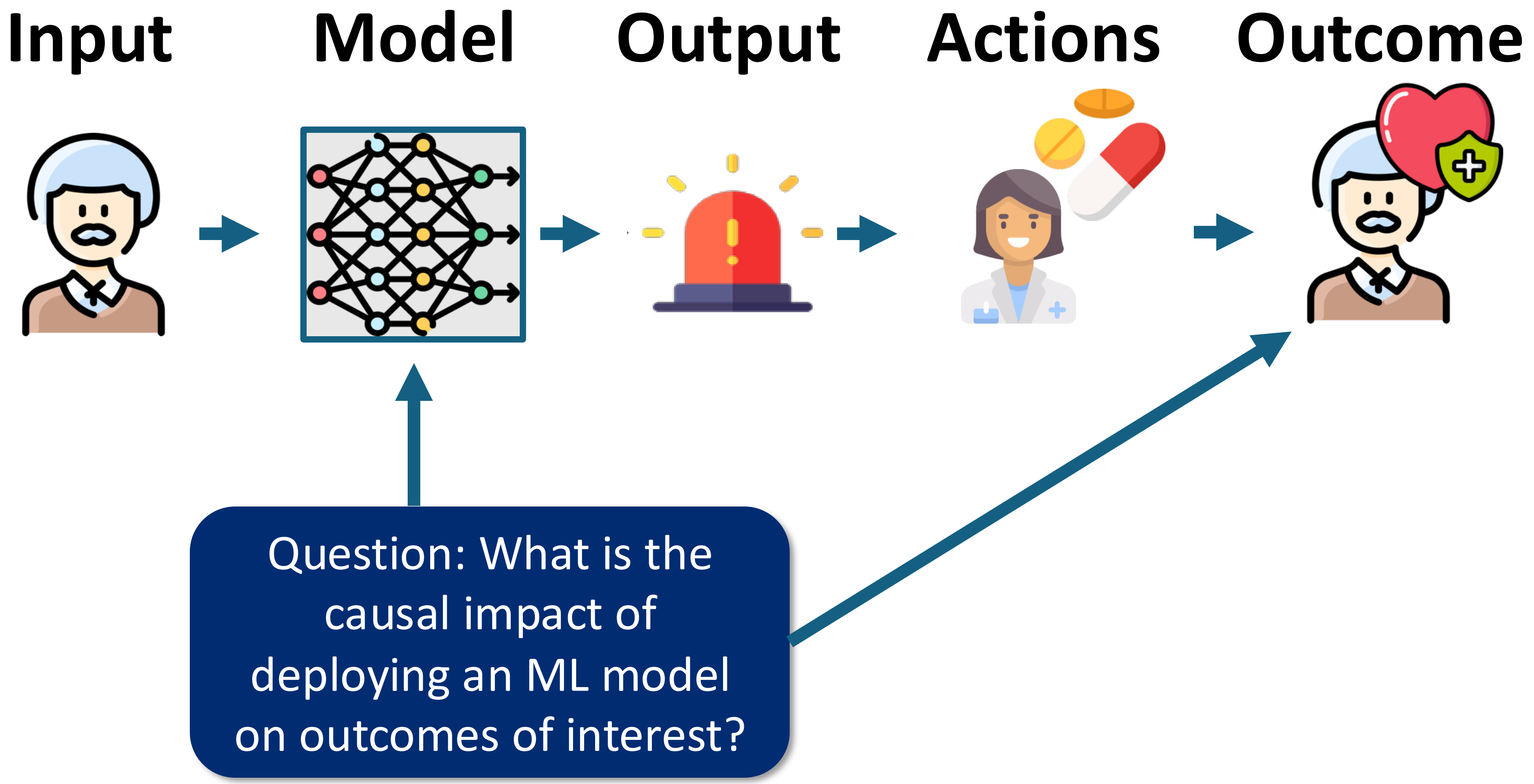
1. ML model produces an output based on patient information.
2. A decision-maker sees the model output and takes an action.

What is causal validation of an ML model?

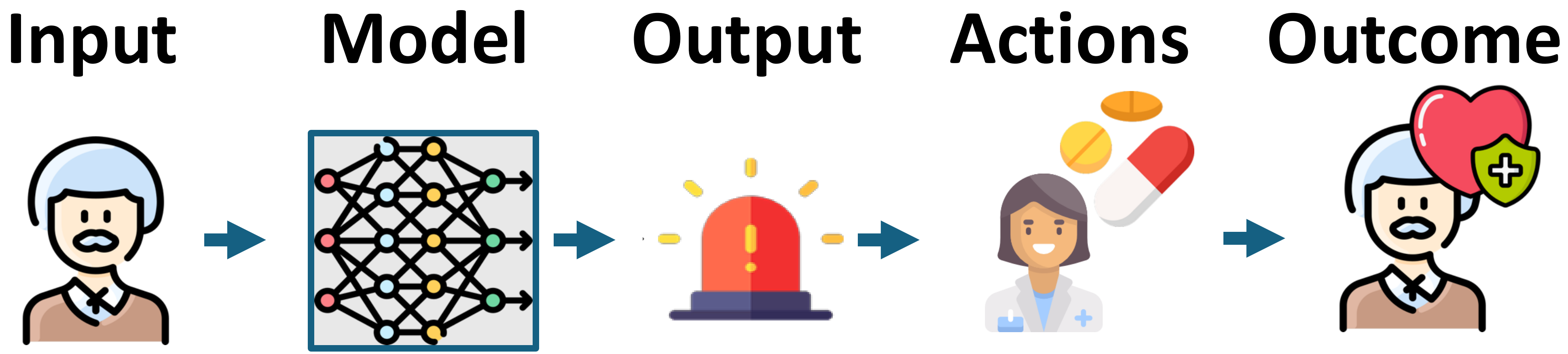


1. ML model produces an output based on patient information.
2. A decision-maker sees the model output and takes an action.
3. We observe outcomes, such as survival.

What is causal validation of an ML model?



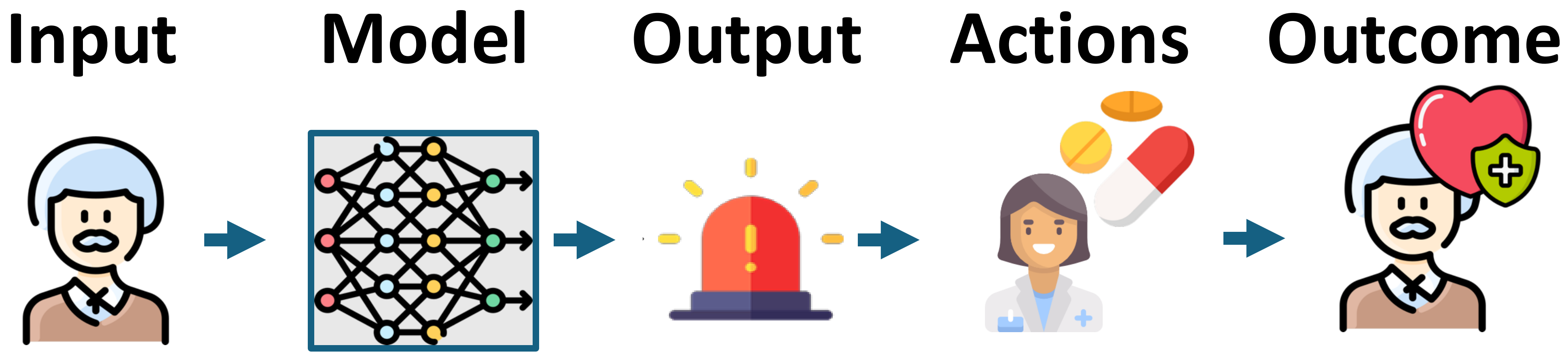
What is causal validation of an ML model?



This is a general problem:

- Does deploying ML models in hospitals improve patient survival?

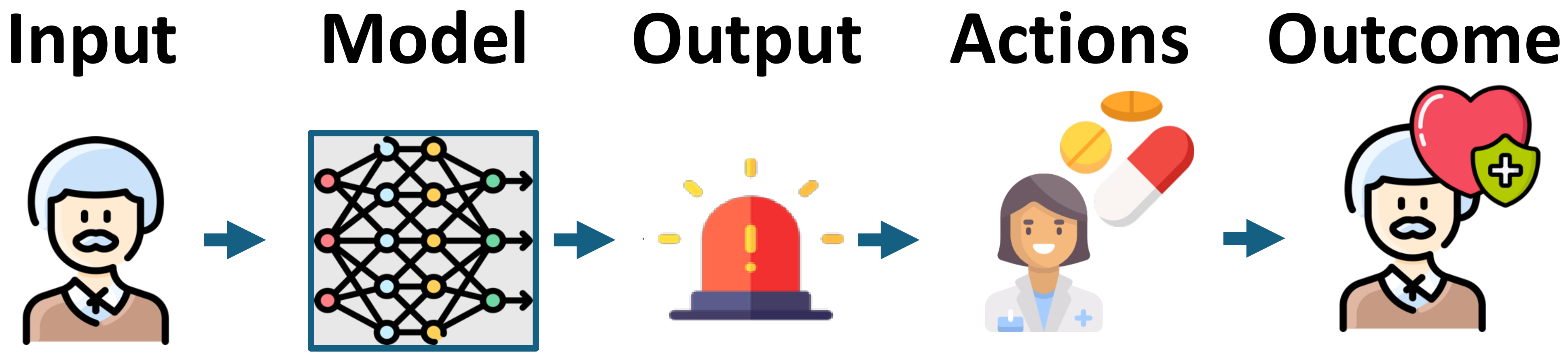
What is causal validation of an ML model?



This is a general problem:

- Does deploying ML models in hospitals improve patient survival?
- Does using AI coding assistants increase the speed and quality of code development for software developers?

What is causal validation of an ML model?

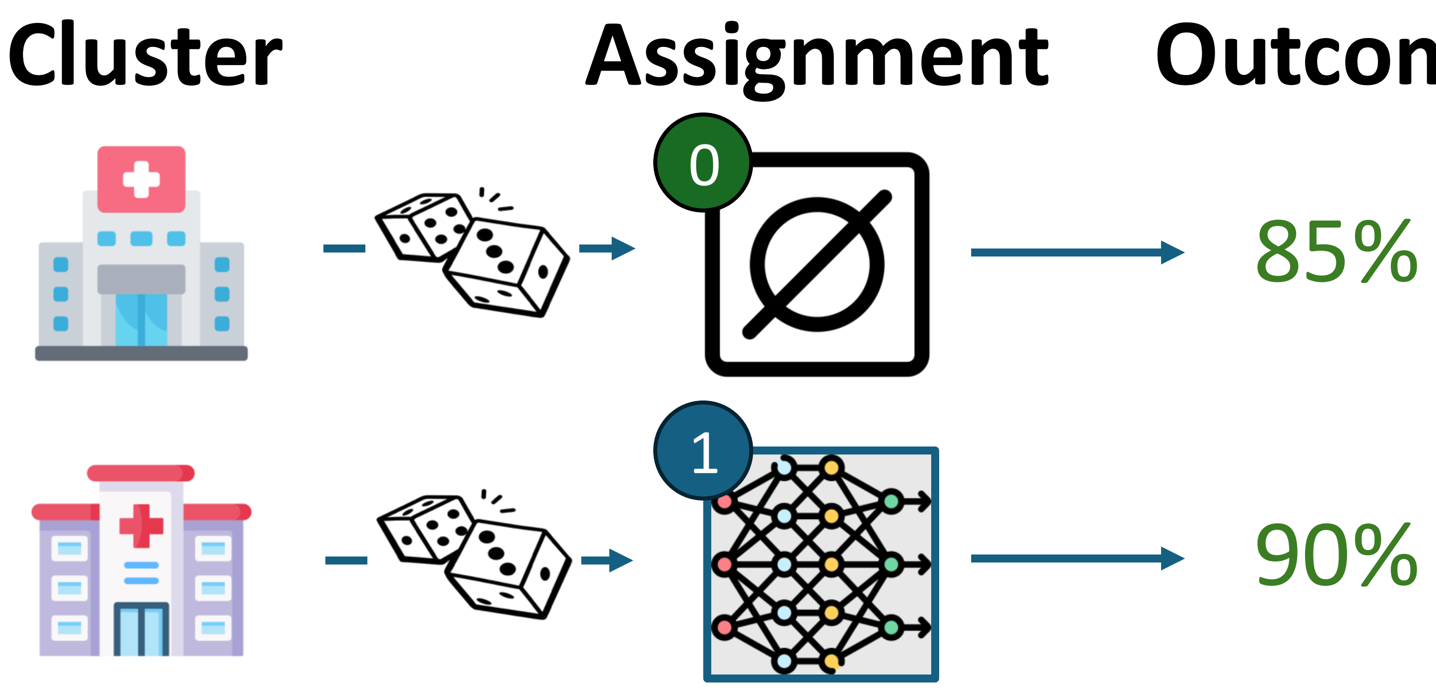


This is a general problem:

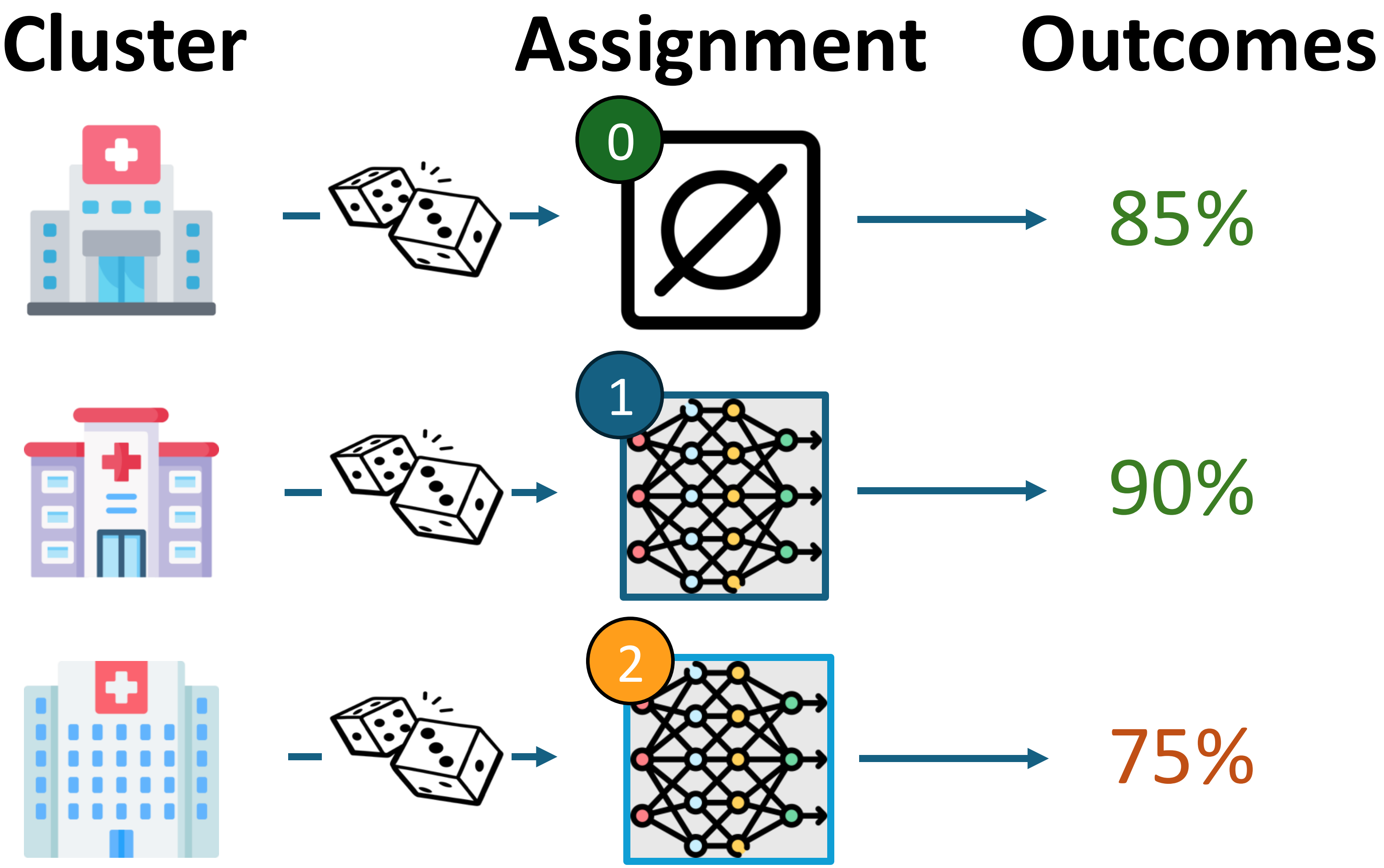
- Does deploying ML models in hospitals improve patient survival?
- Does using AI coding assistants increase the speed and quality of code development for software developers?
- Does using bail recommendation systems improve defendant return rates to the courtroom?

Cluster Randomized Controlled Trials (RCTs)

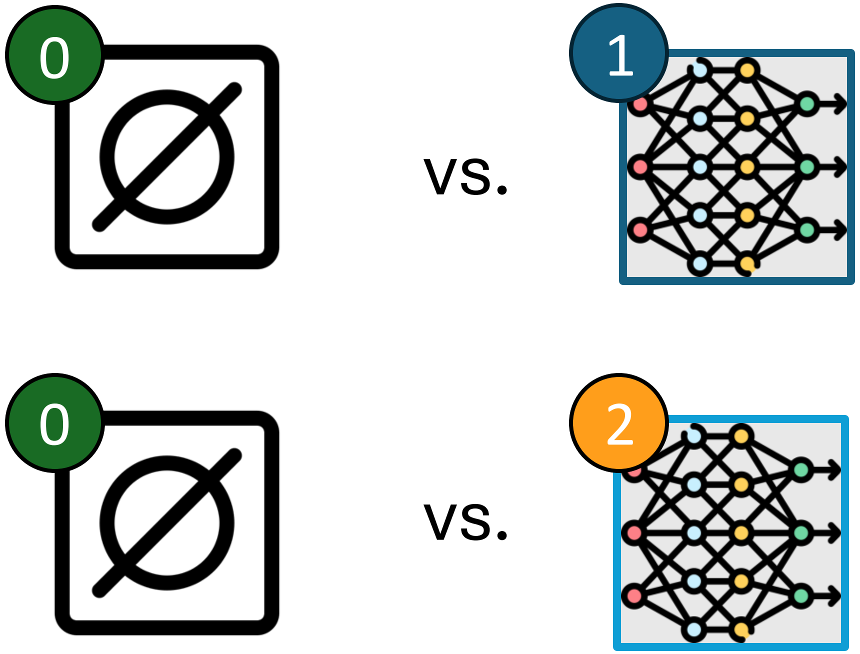
Cluster Randomized Controlled Trials (RCTs)



Cluster RCT Design with Multiple Models

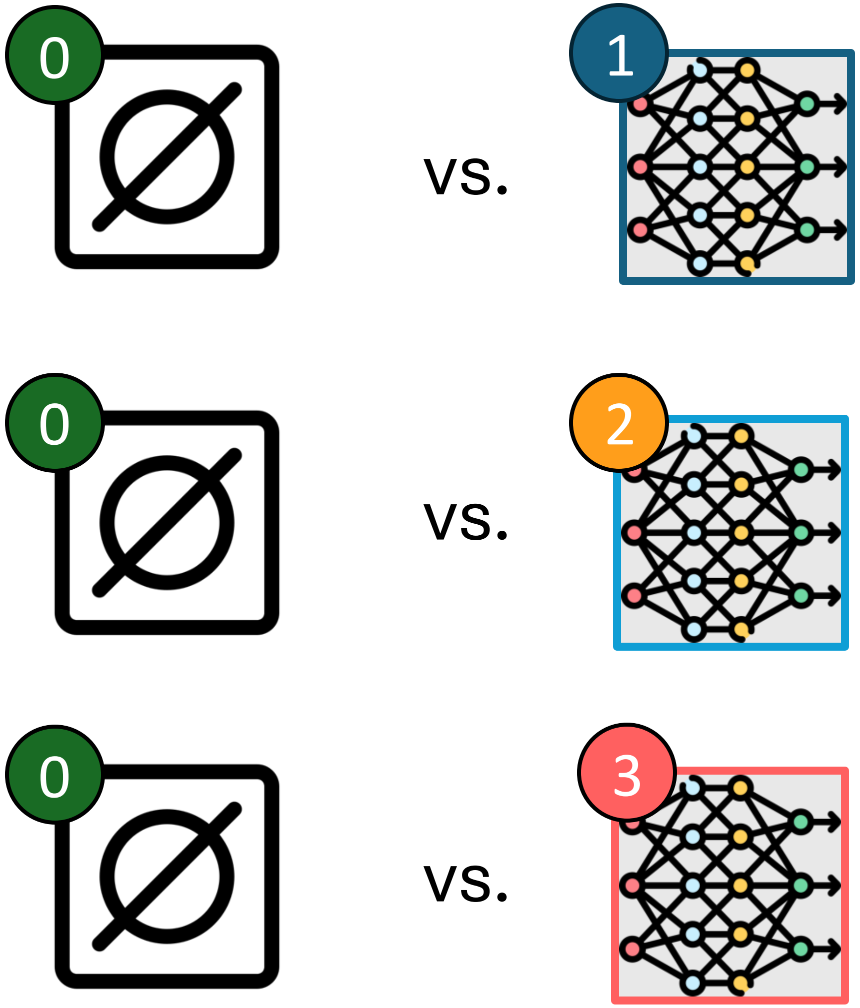


Limitations of Cluster RCTs



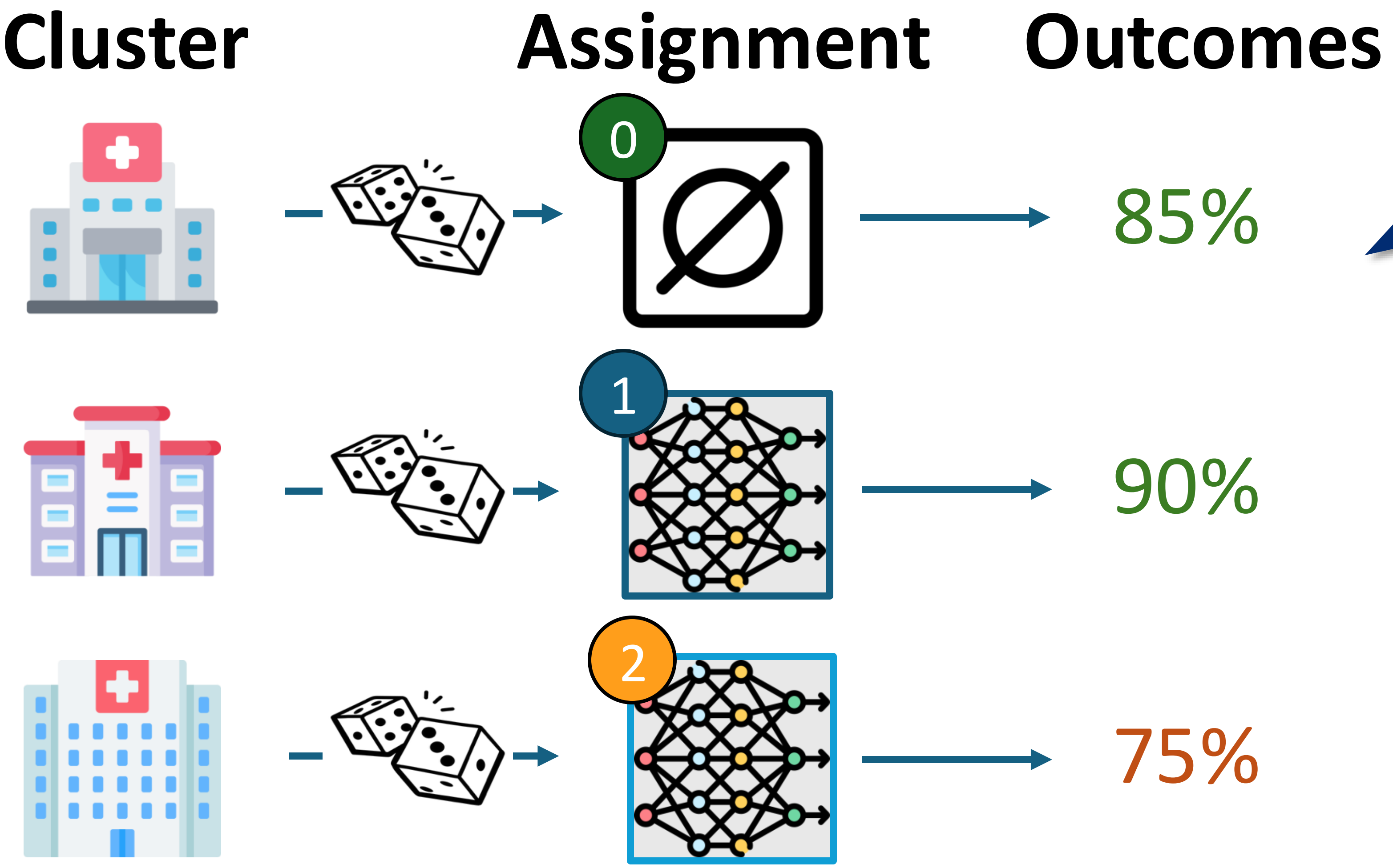
Allows for evaluation of models trialed in the cluster RCT.

Limitations of Cluster RCTs



Does not allow for evaluation of a new, never-trialed models.

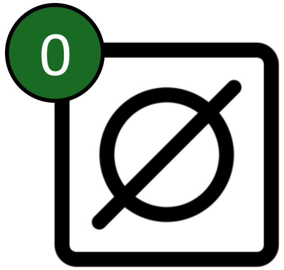
Our Goal: Just Trial Once



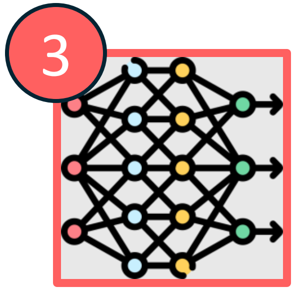
Given data from a cluster RCT...

Our Goal: Just Trial Once

Assignment | **Outcomes**



85%

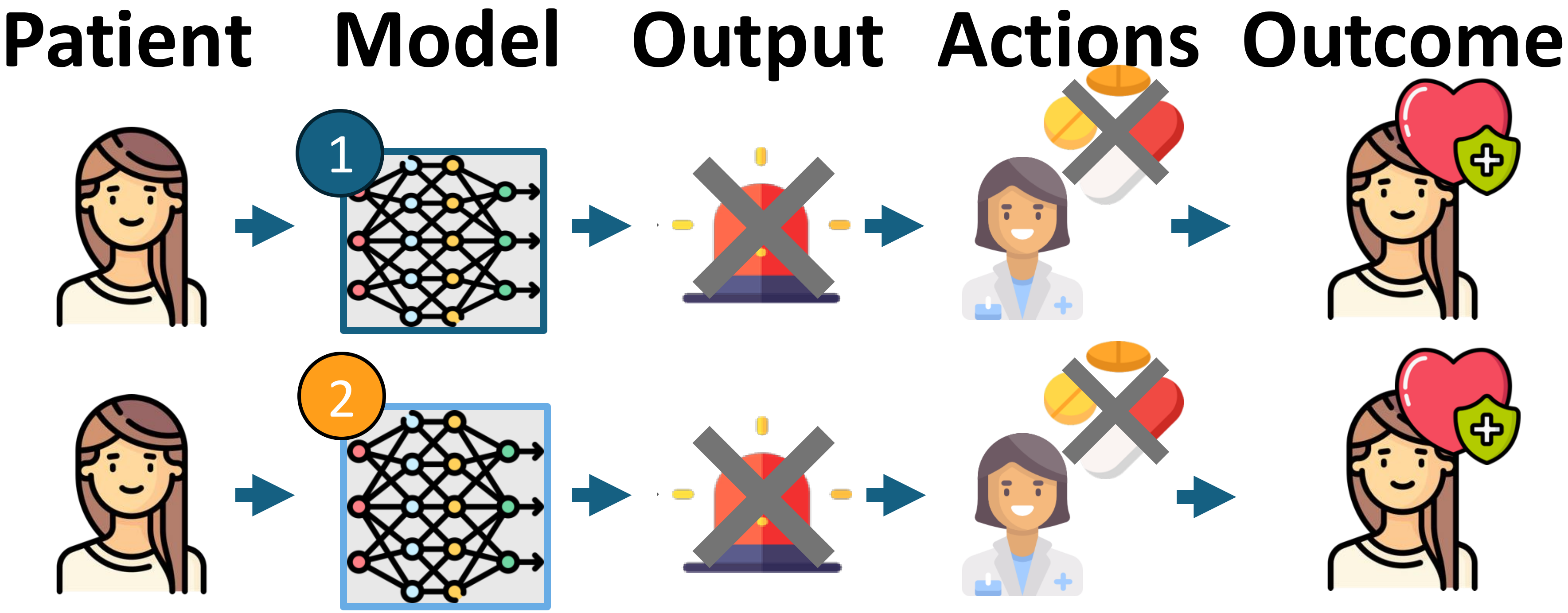


87% - 92%

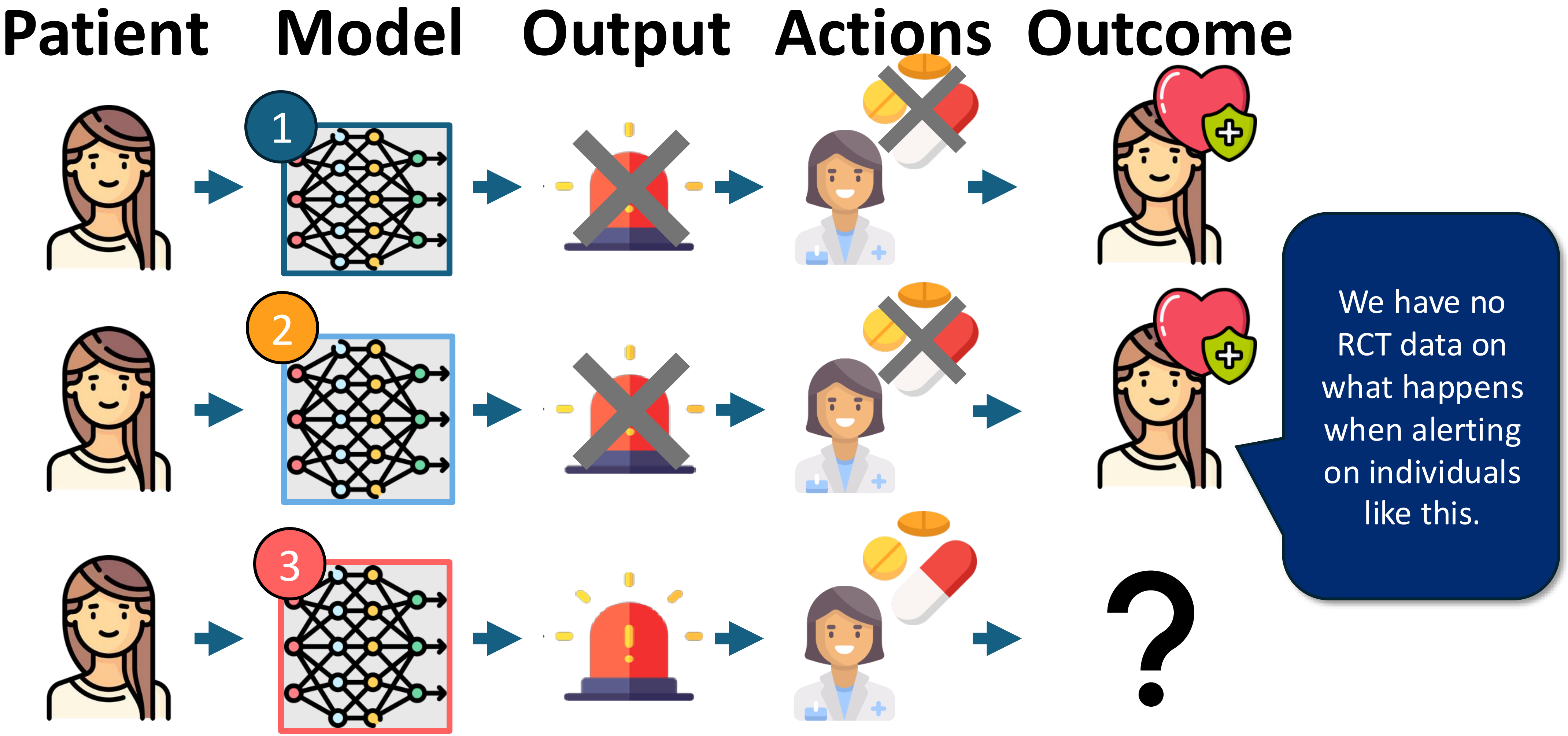
… bound the outcome of interest under a never-deployed model between *L* and *U*.

Two Challenges: **Coverage** and Trust

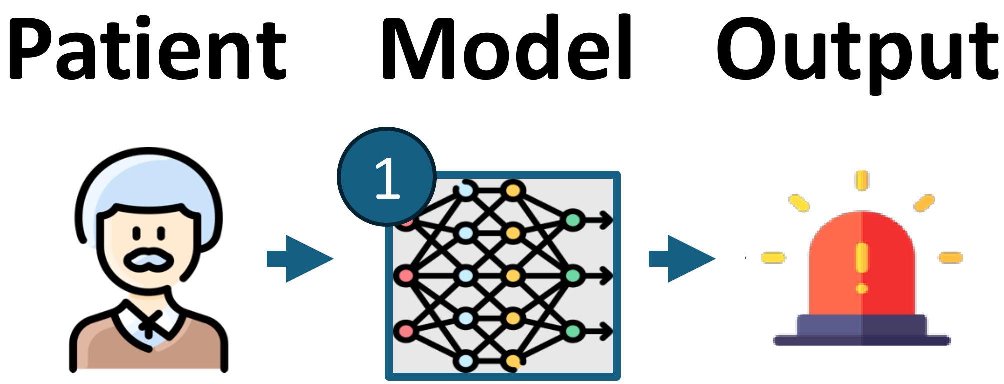
Two Challenges: **Coverage** and Trust



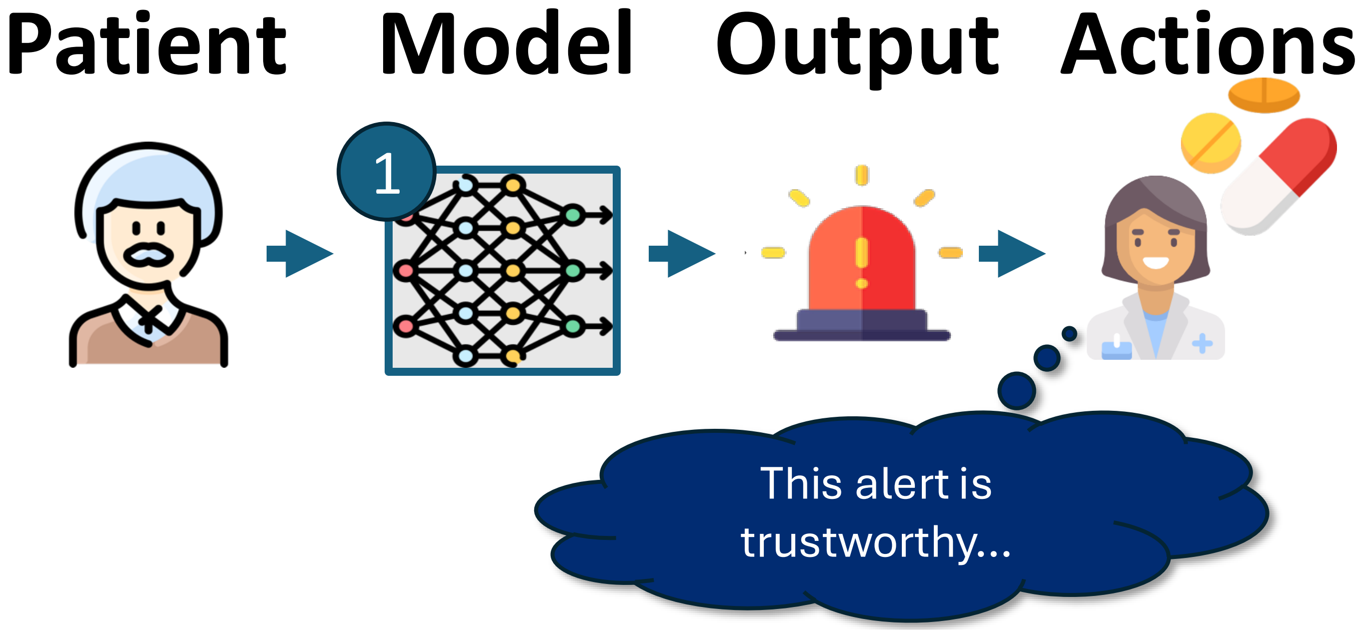
Two Challenges: **Coverage** and Trust



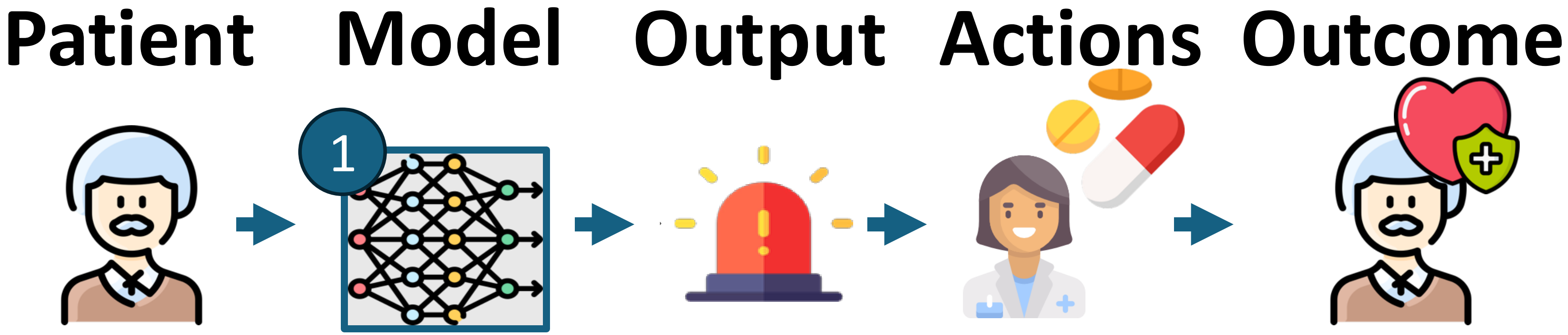
Two Challenges: Coverage and **Trust**



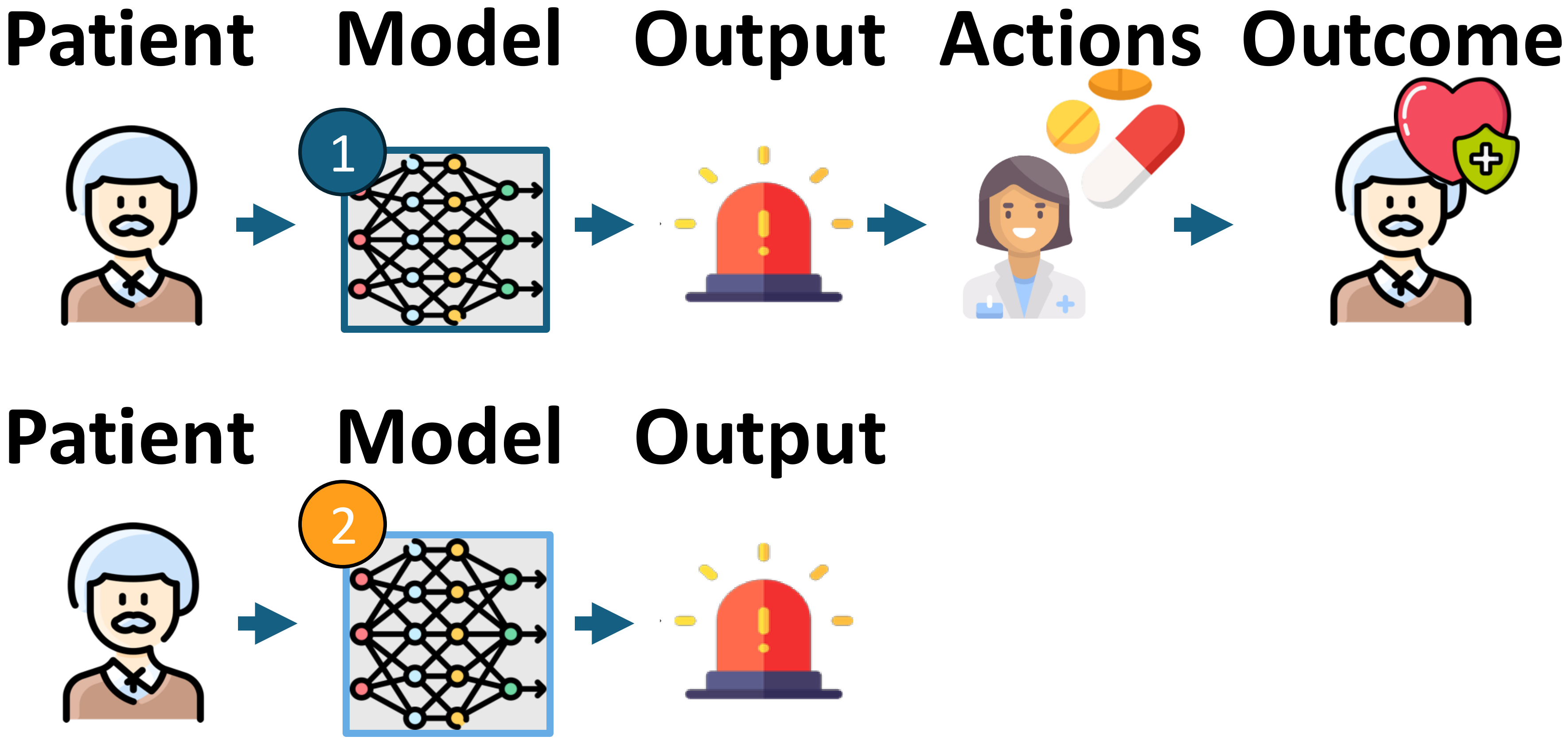
Two Challenges: Coverage and **Trust**



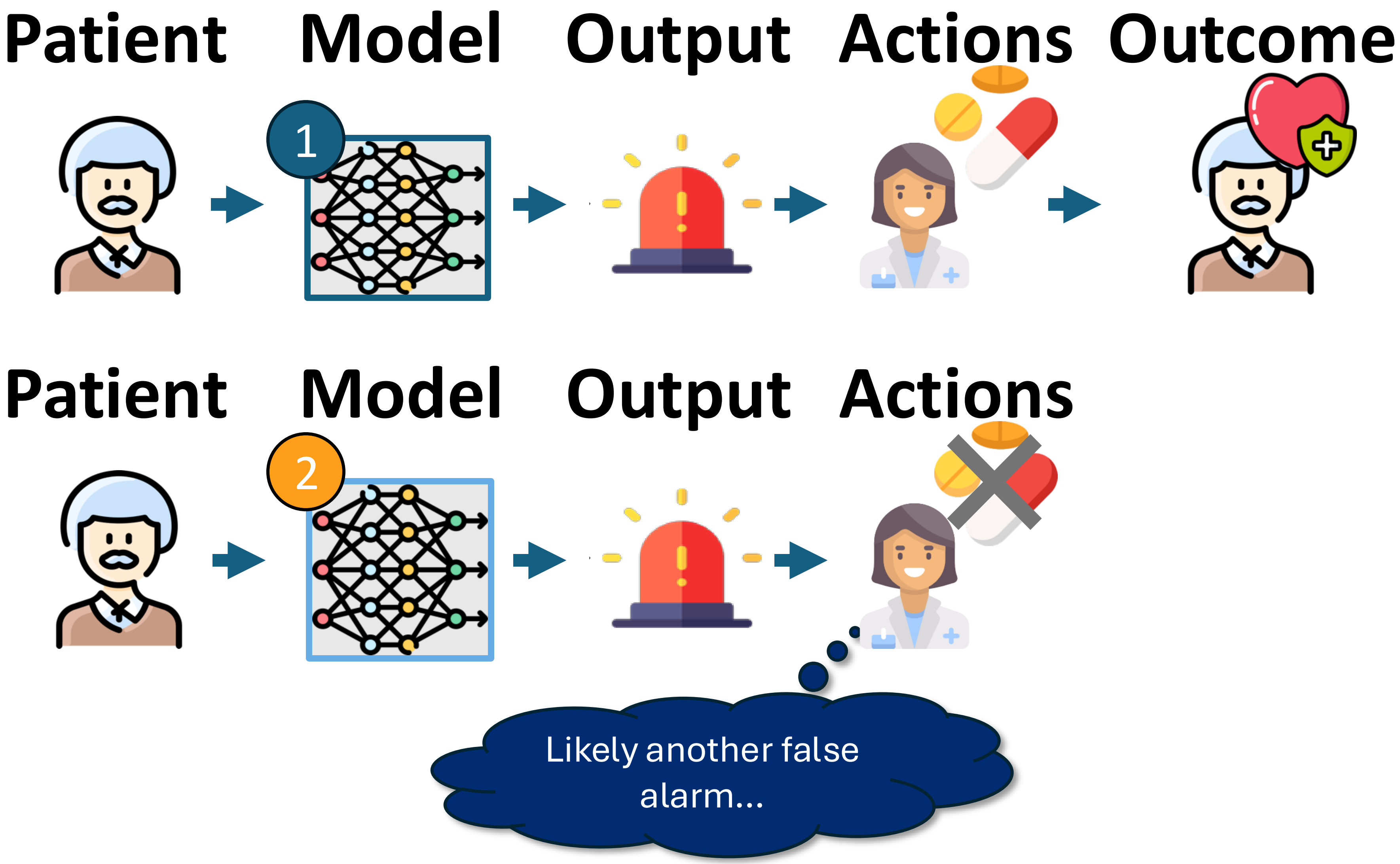
Two Challenges: Coverage and **Trust**



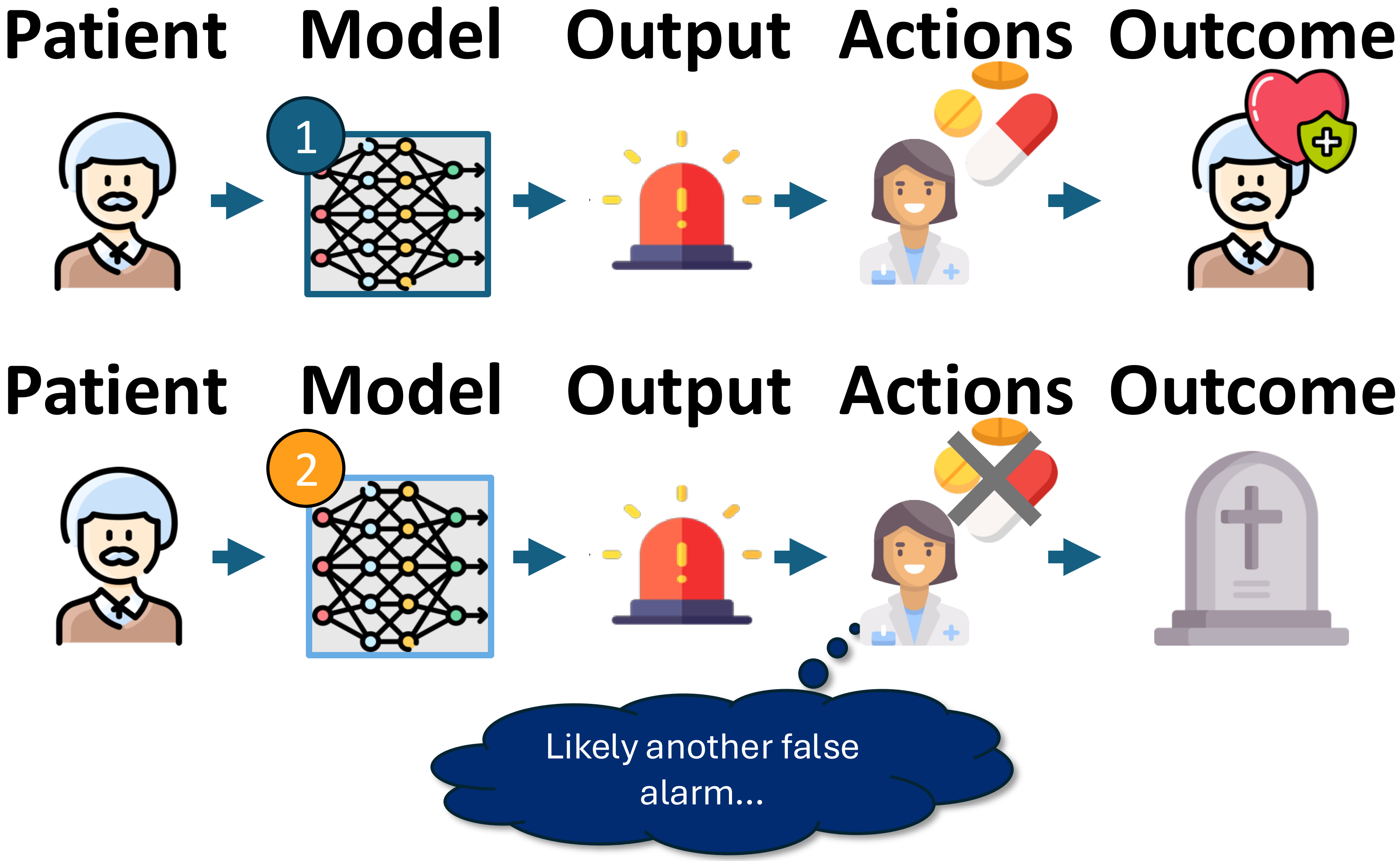
Two Challenges: Coverage and **Trust**



Two Challenges: Coverage and **Trust**



Two Challenges: Coverage and **Trust**

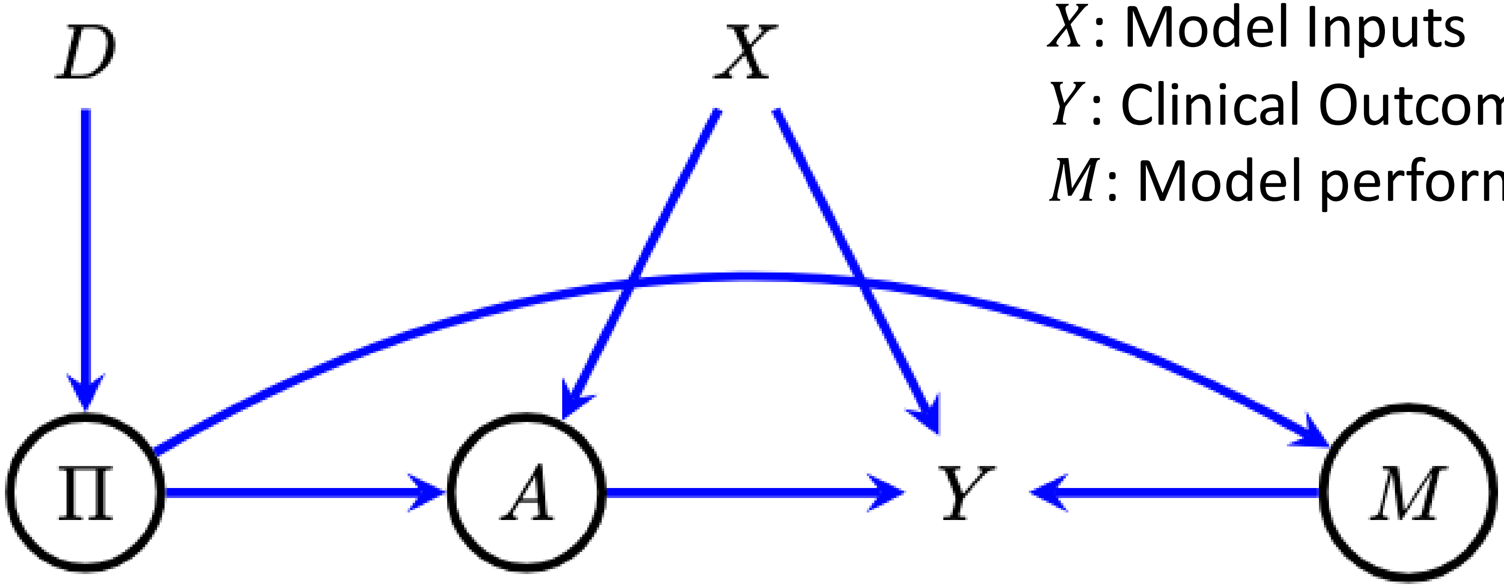


Contributions Overview

1. Introduce the problem and its key challenges.
2. Introduce a formal setup for approaching the problem.
3. State assumptions that address the key challenges.
4. Bound the causal effect of interest.

Problem Setup

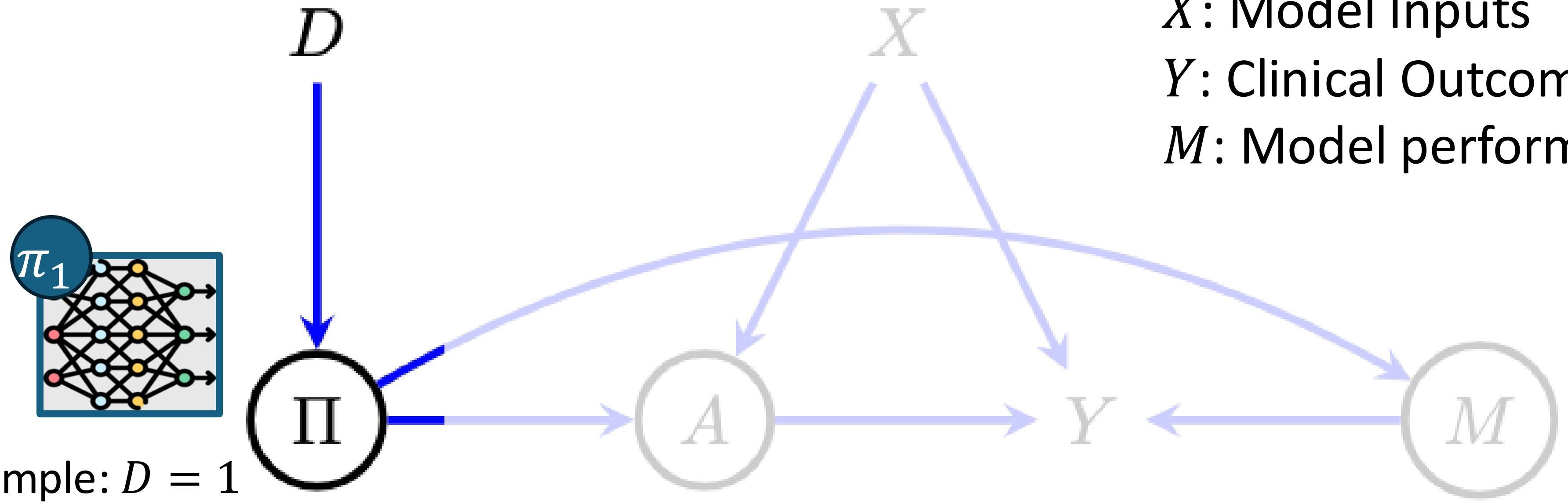
D: Indicator of trial arm
Π: Deployed model/policy
A: Model Output
X: Model Inputs
Y: Clinical Outcome
M: Model performance metric



Assumed causal data-generating process

Problem Setup

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

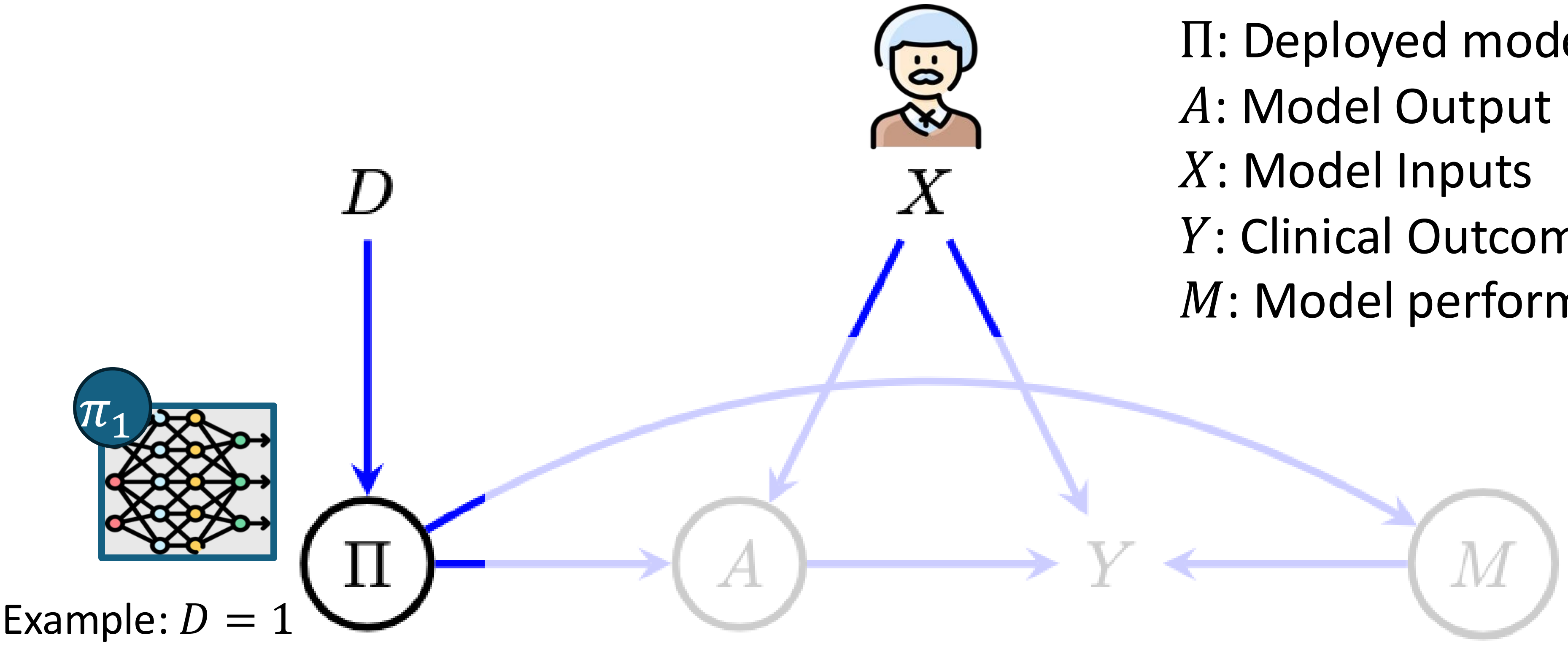


Example: $D = 1$

Assumed causal data-generating process

Problem Setup

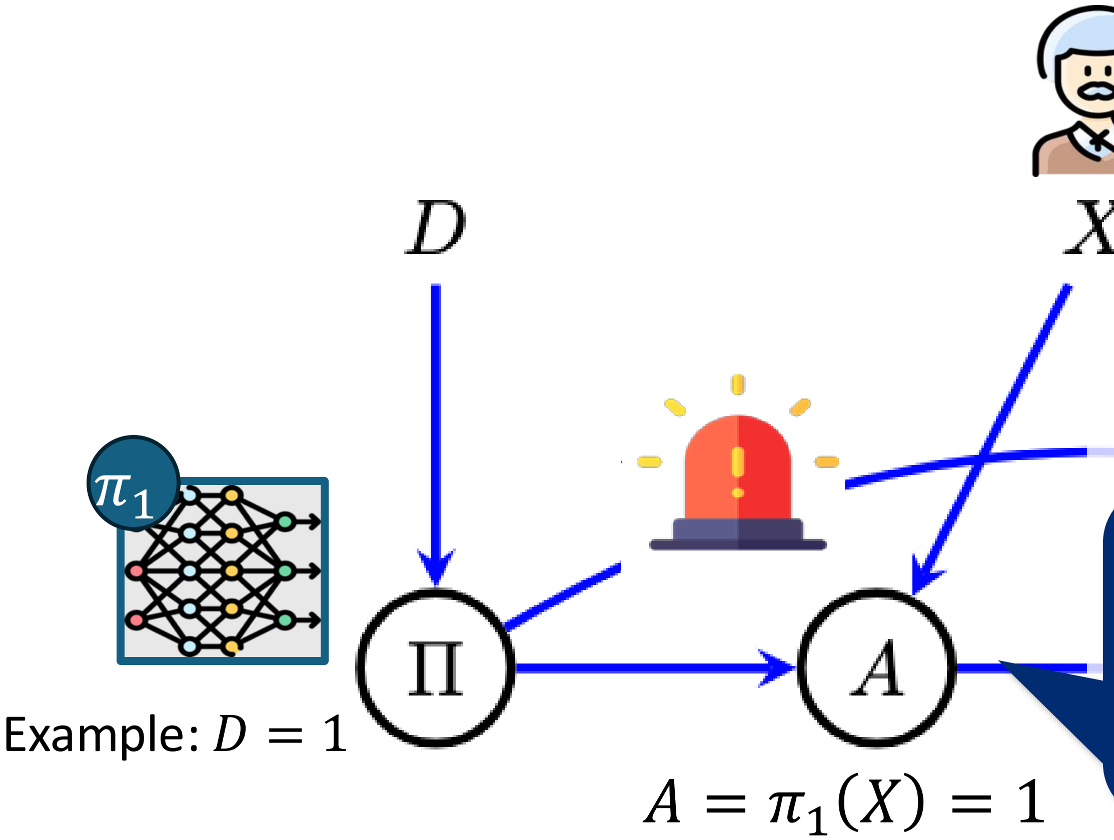
D: Indicator of trial arm
Π: Deployed model/policy
A: Model Output
X: Model Inputs
Y: Clinical Outcome
M: Model performance metric



Assumed causal data-generating process

Problem Setup

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

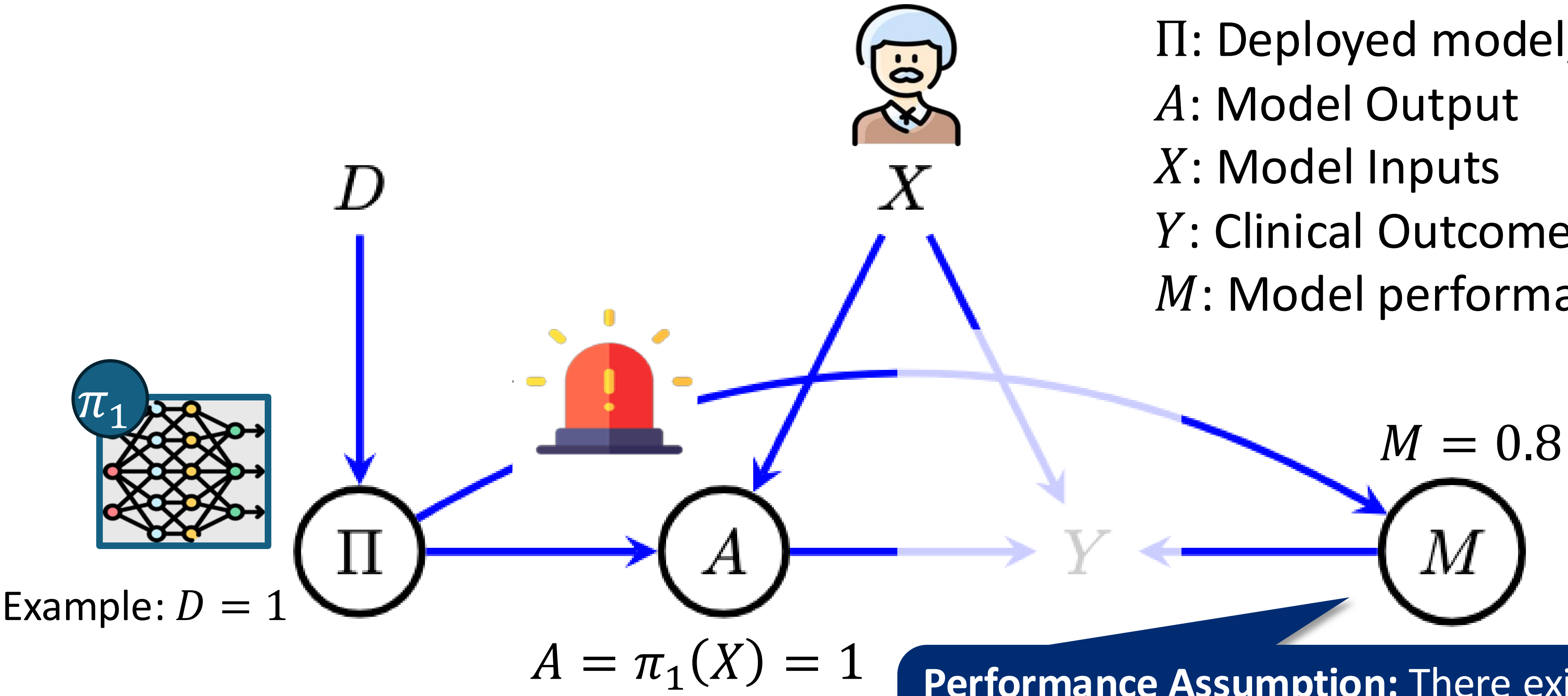


Model alerts are deterministic: we are unlikely to see all possible model outputs for all types of patients.

Assumed causal data-generating process

Problem Setup

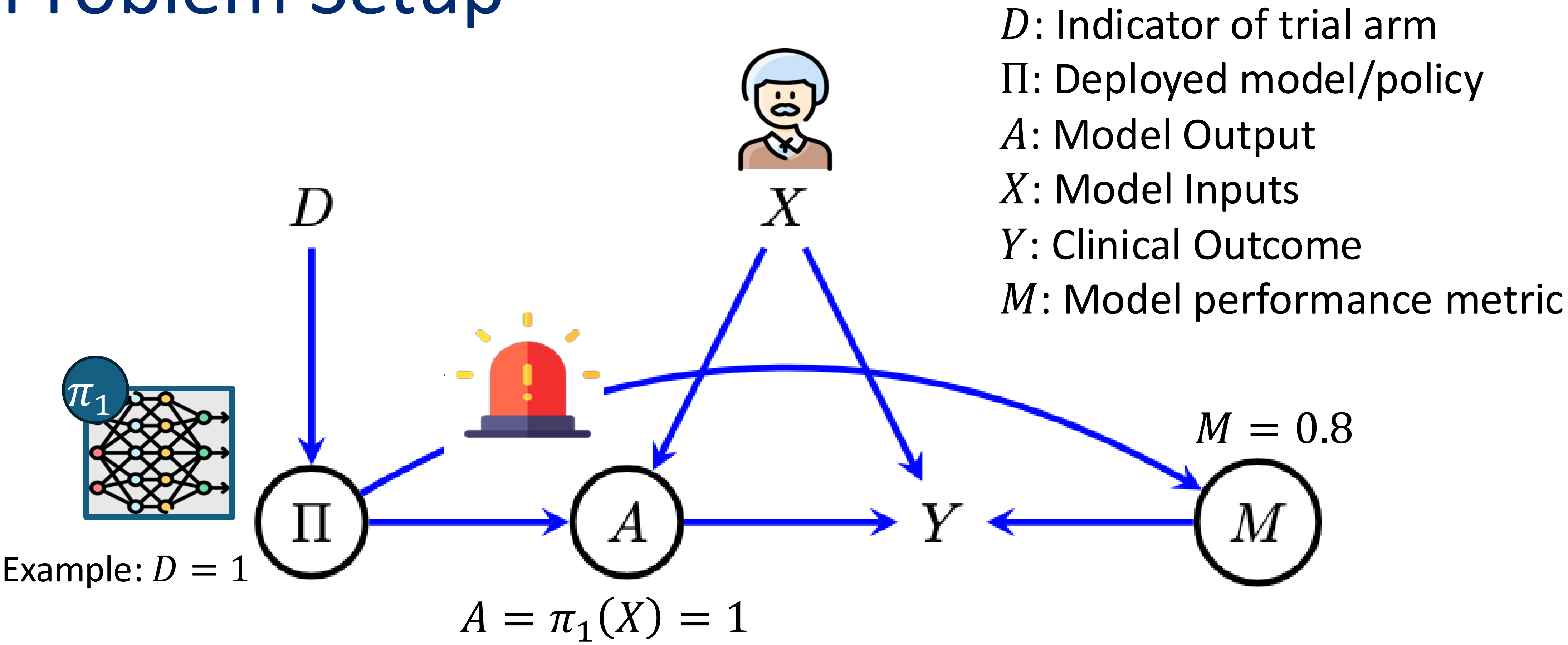
D: Indicator of trial arm
Π: Deployed model/policy
A: Model Output
X: Model Inputs
Y: Clinical Outcome
M: Model performance metric



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

Assumed causal data-generating process

Problem Setup

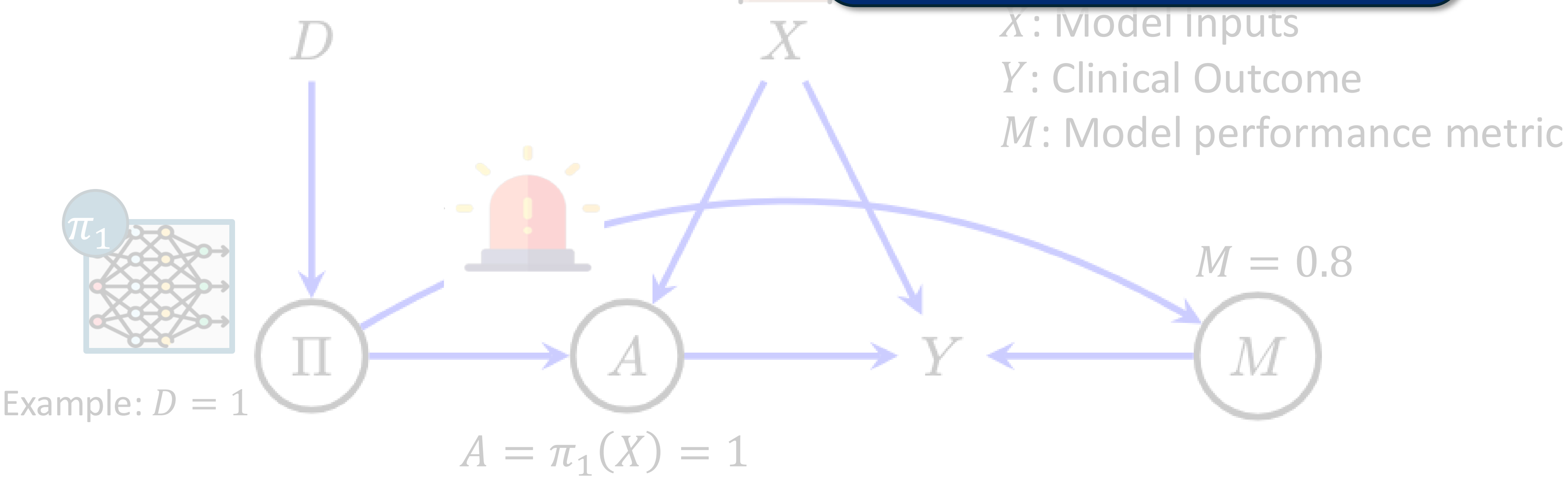


Assumed causal data-generating process

Problem Setup

Goal: Bound $\mathrm{E}[Y(\pi_{new})]$, the expected outcome under model π_{new}.

Potential outcomes notation: the outcome that would have occurred had we counterfactually deployed the new model.

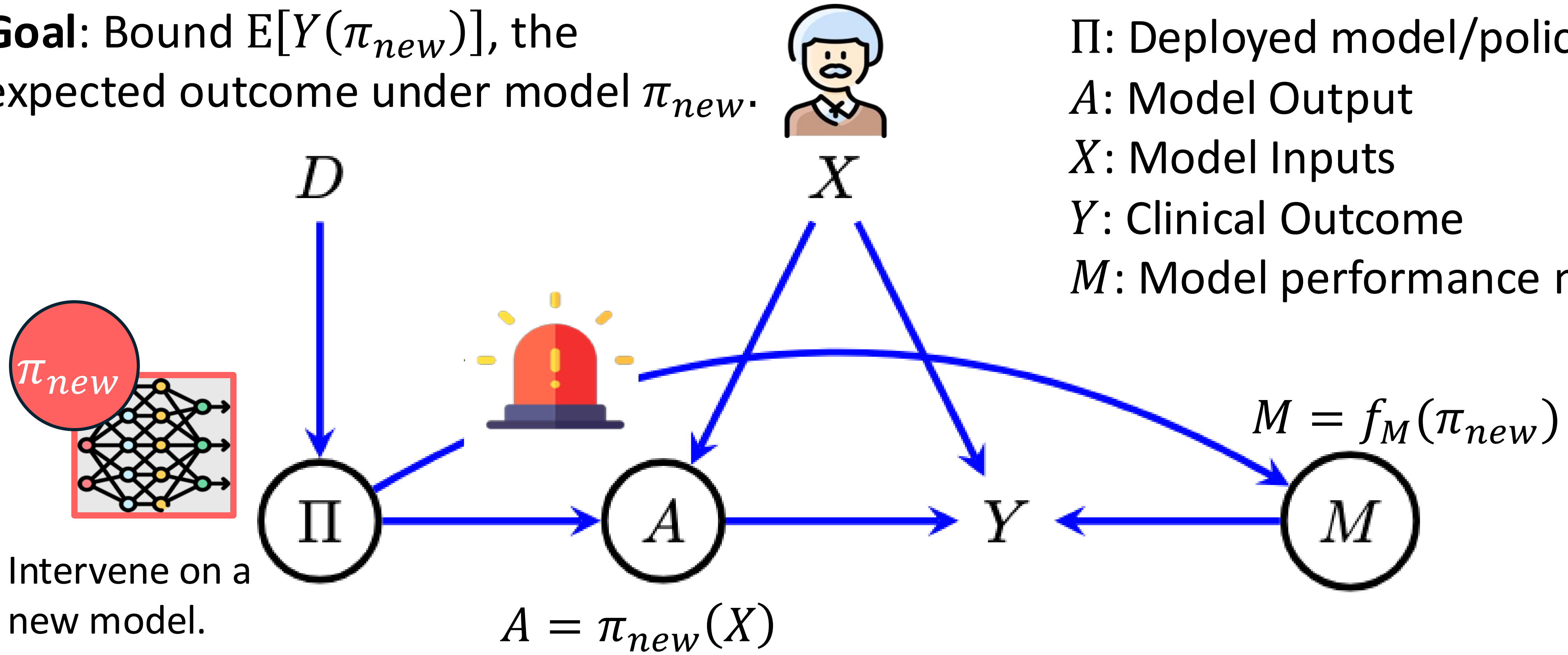


Assumed causal data-generating process

Problem Setup

Goal: Bound $\mathrm{E}[Y(\pi_{new})]$, the expected outcome under model π_{new}.

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



Assumed causal data-generating process

Contributions Overview

1. Introduce the problem and its key challenges.
2. Introduce a formal setup for approaching the problem.
3. State assumptions that address the key challenges.
4. Bound the causal effect of interest.

Assumptions to Evaluate a New Model

Assumption 1: Performance Monotonicity

Potential outcomes are non-decreasing in model performance metric, i.e., if $m_i < m_j$ then for all $a \in \mathcal{A}$,

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

Assumptions to Evaluate a New Model

Assumption 1: Performance Monotonicity

Potential outcomes are non-decreasing in model performance metric, i.e., if $m_i < m_j$ then for all $a \in \mathcal{A}$,

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

Given a fixed model output, a model with better performance metric will not make outcomes worse.

Assumptions to Evaluate a New Model

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 $a \in \mathcal{A}$,

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

Given a fixed model output, a model with better performance metric will not make outcomes worse.

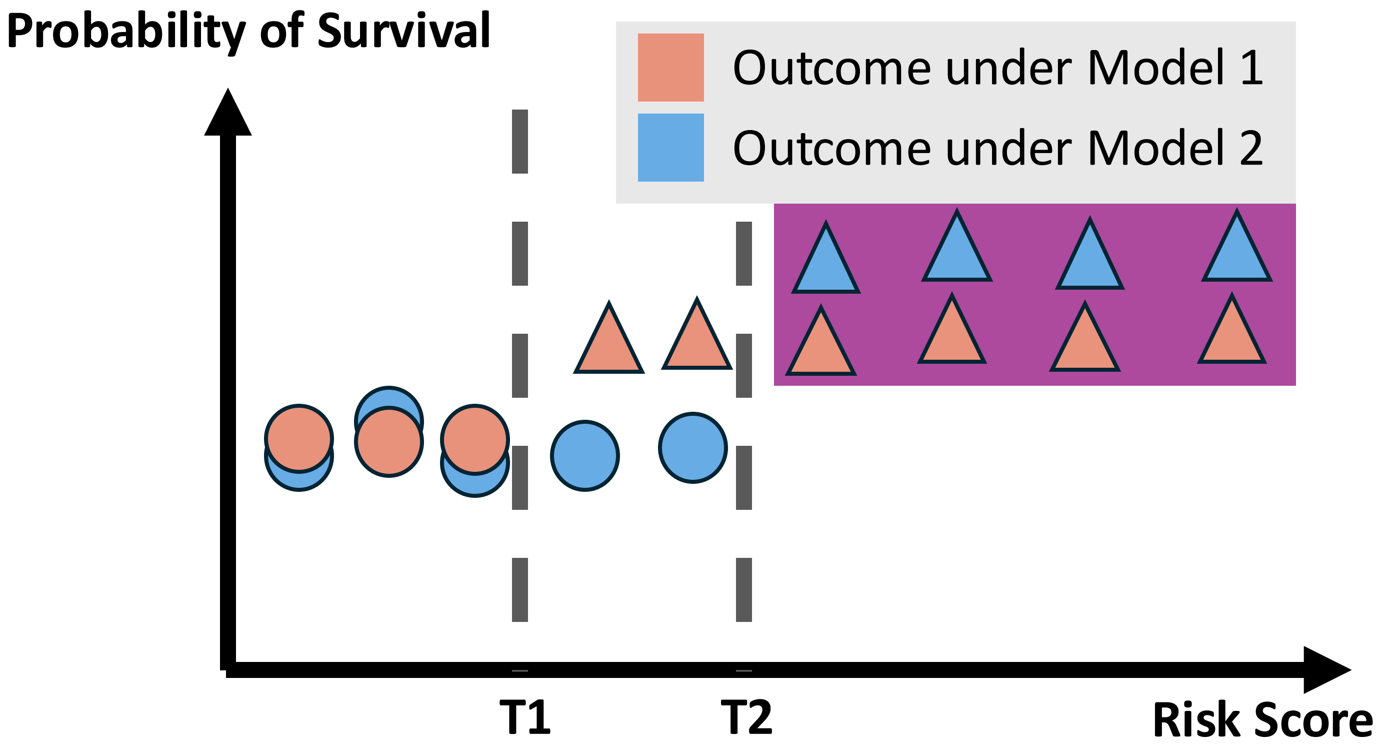
This assumption has observable implications in the RCT if multiple models are trialed. Thus, it can be falsified by comparing two empirical means.

Assumptions to Evaluate a New Model

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 $a \in \mathcal{A}$,

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



Given a fixed model output, a model with better performance metric will not make outcomes worse.

Model 2 has a better performance metric than Model 1.

Is Assumption 1 consistent with data? ✓

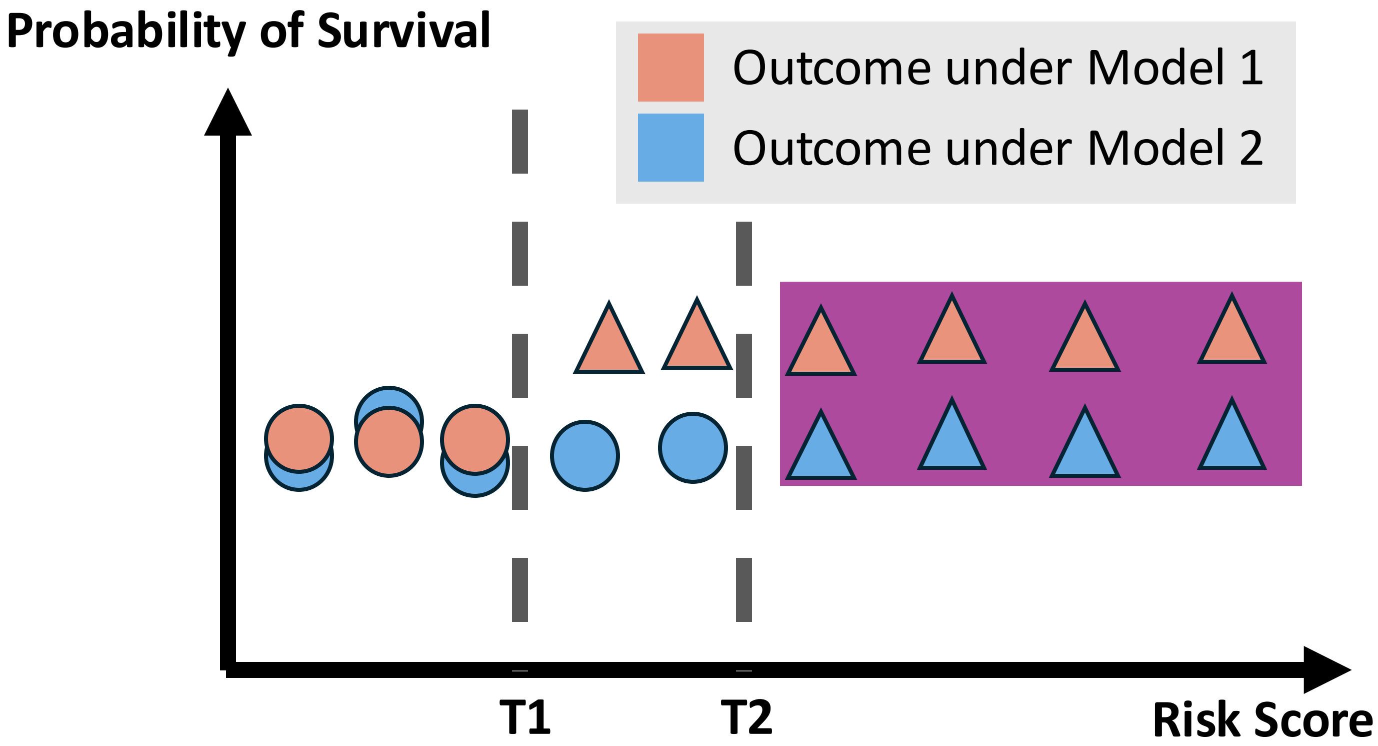
Assumptions to Evaluate a New Model

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 $a \in \mathcal{A}$,

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

Given a fixed model output, a model with better performance metric will not make outcomes worse.



Model 2 has a better performance metric than Model 1.

Is Assumption 1 consistent with data? ✕

Assumptions to Evaluate a New Model

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

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

Assumptions to Evaluate a New Model

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

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

Given that the model outputs the neutral action, the performance metric does not affect the outcome.

Assumptions to Evaluate a New Model

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

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

Given that the model outputs the neutral action, the performance metric does not affect the outcome.

The "control model" always outputs the neutral action (deferral, no alert, etc.). This assumption allows us to make use of control arm data.

Assumptions to Evaluate a New Model

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 $m_i \neq m_j$,

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

Given that the model outputs the neutral action, the performance metric does not affect the outcome.

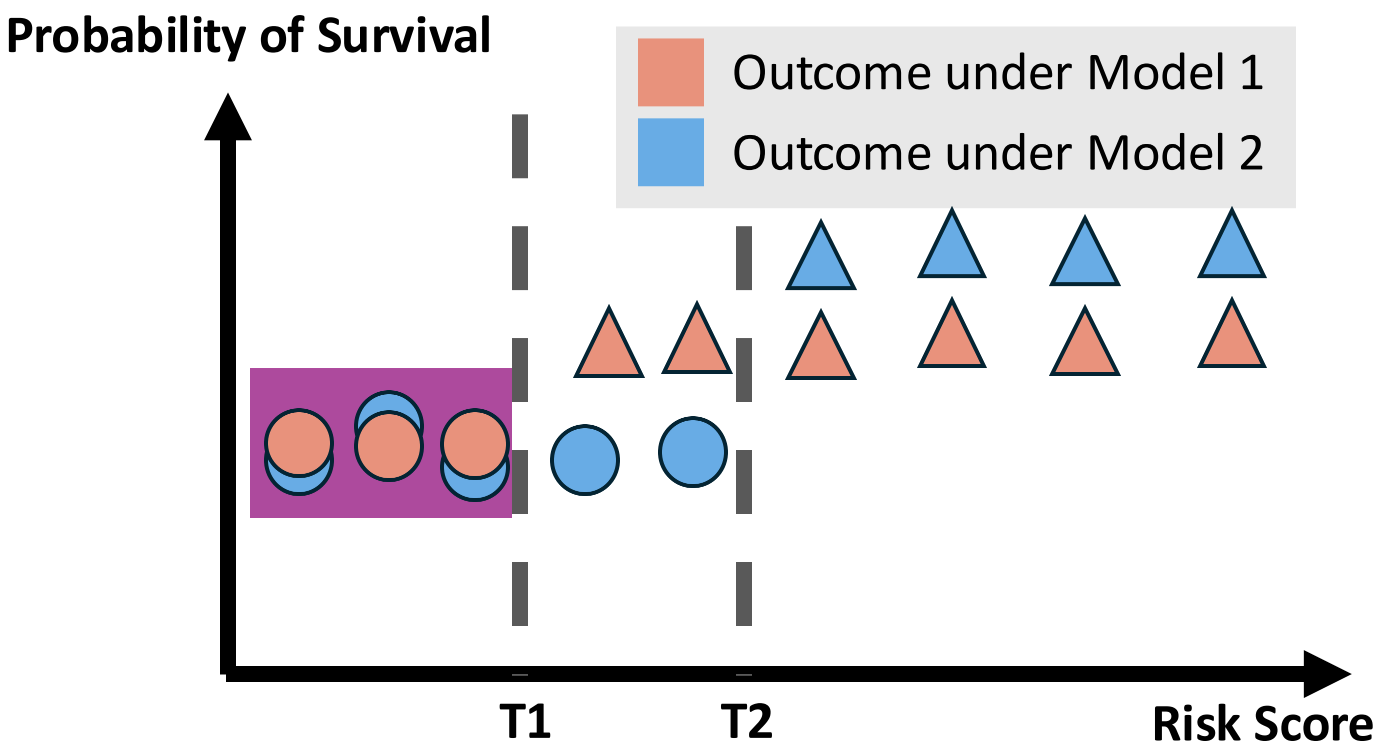
This assumption also has <u>observable implications in the RCT</u> that <u>can be falsified</u> by comparing two empirical means.

Assumptions to Evaluate a New Model

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 $m_i \neq m_j$,

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



Given that the model outputs the neutral action, the performance metric does not affect the outcome.

Model 2 has a better performance metric than Model 1.

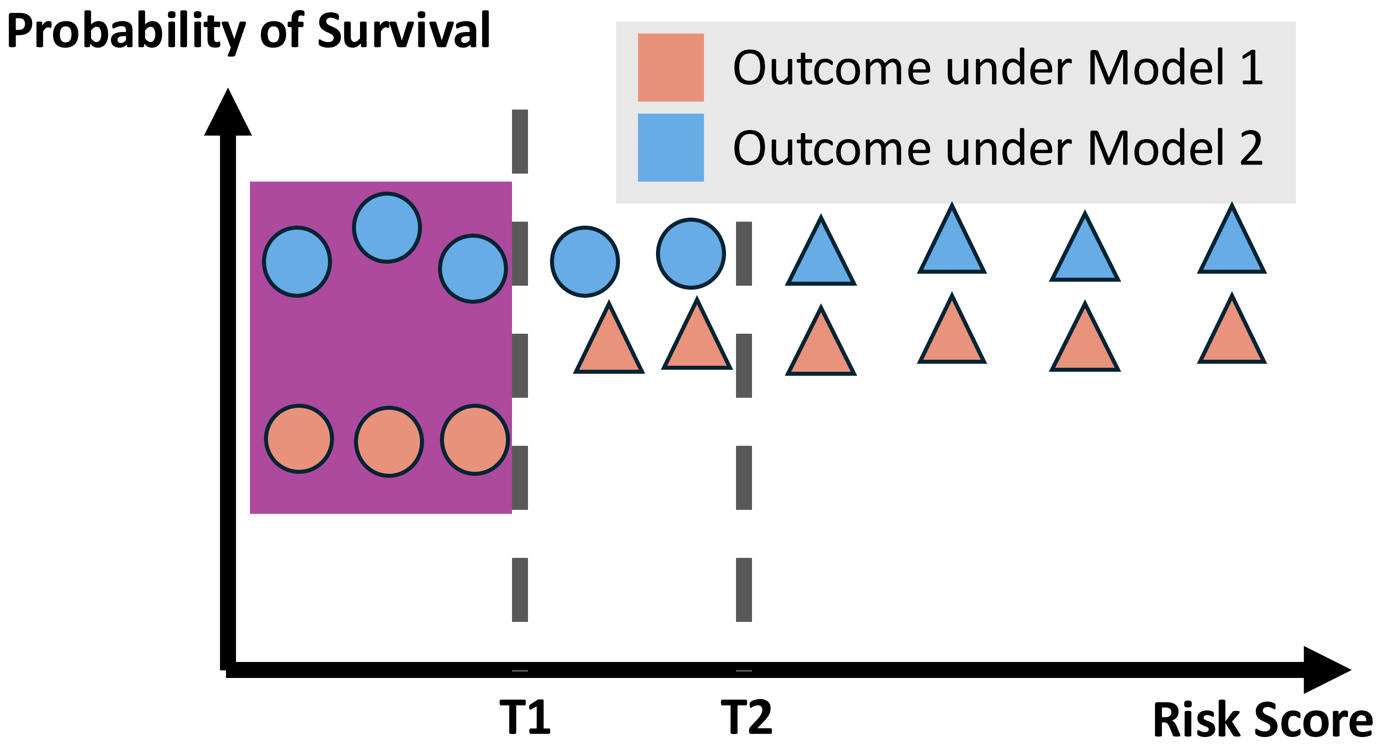
Is Assumption 2 consistent with data?

Assumptions to Evaluate a New Model

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 $m_i \neq m_j$,

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



Given that the model outputs the neutral action, the performance metric does not affect the outcome.

Model 2 has a better performance metric than Model 1.

Is Assumption 2 consistent with data? ✗

Assumptions to Evaluate a New Model

Assumption 3: Bounded Outcomes

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

Assumptions to Evaluate a New Model

Assumption 3: Bounded Outcomes
There exists constants Y_{min} and Y_{max} such that $Y_{min} \leq \mathrm{Y} \leq Y_{max}$.

Allows us to know what the best / worst-case scenarios are.

Assumptions to Evaluate a New Model

Assumption 3: Bounded Outcomes
There exists constants Y_{min} and Y_{max} such that $Y_{min} \leq Y \leq Y_{max}$.

Allows us to know what the best / worst-case scenarios are.

Boundedness is satisfied in practice with, for example, binary outcomes.

Assumptions to Evaluate a New Model

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

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 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 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 $Y_{min} \leq \mathrm{Y} \leq Y_{max}$.

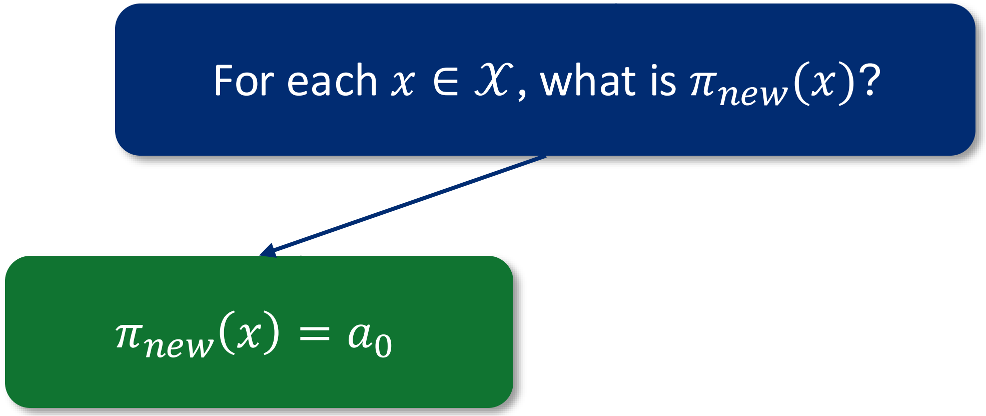
Contributions Overview

1. Introduce the problem and its key challenges.
2. Introduce a formal setup for approaching the problem.
3. State assumptions that address the key challenges.
4. Bound the causal effect of interest.

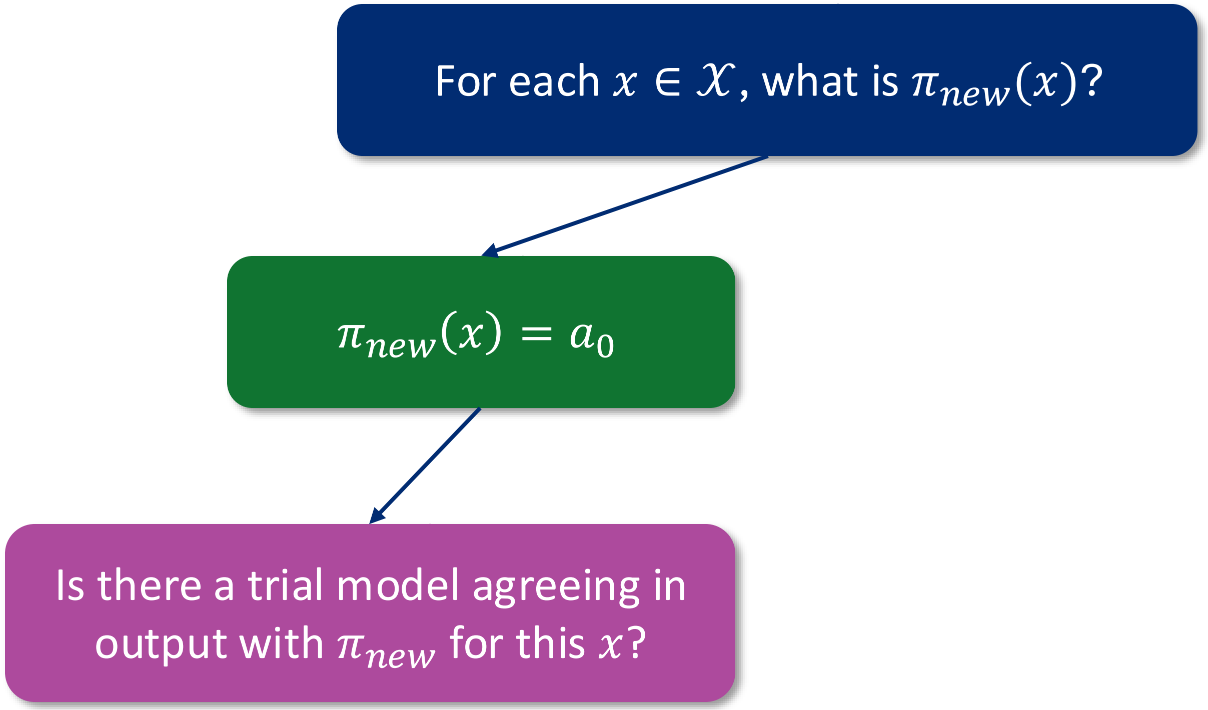
Lower Bound on Causal Impact

For each $x \in \mathcal{X}$, what is $\pi_{new}(x)$?

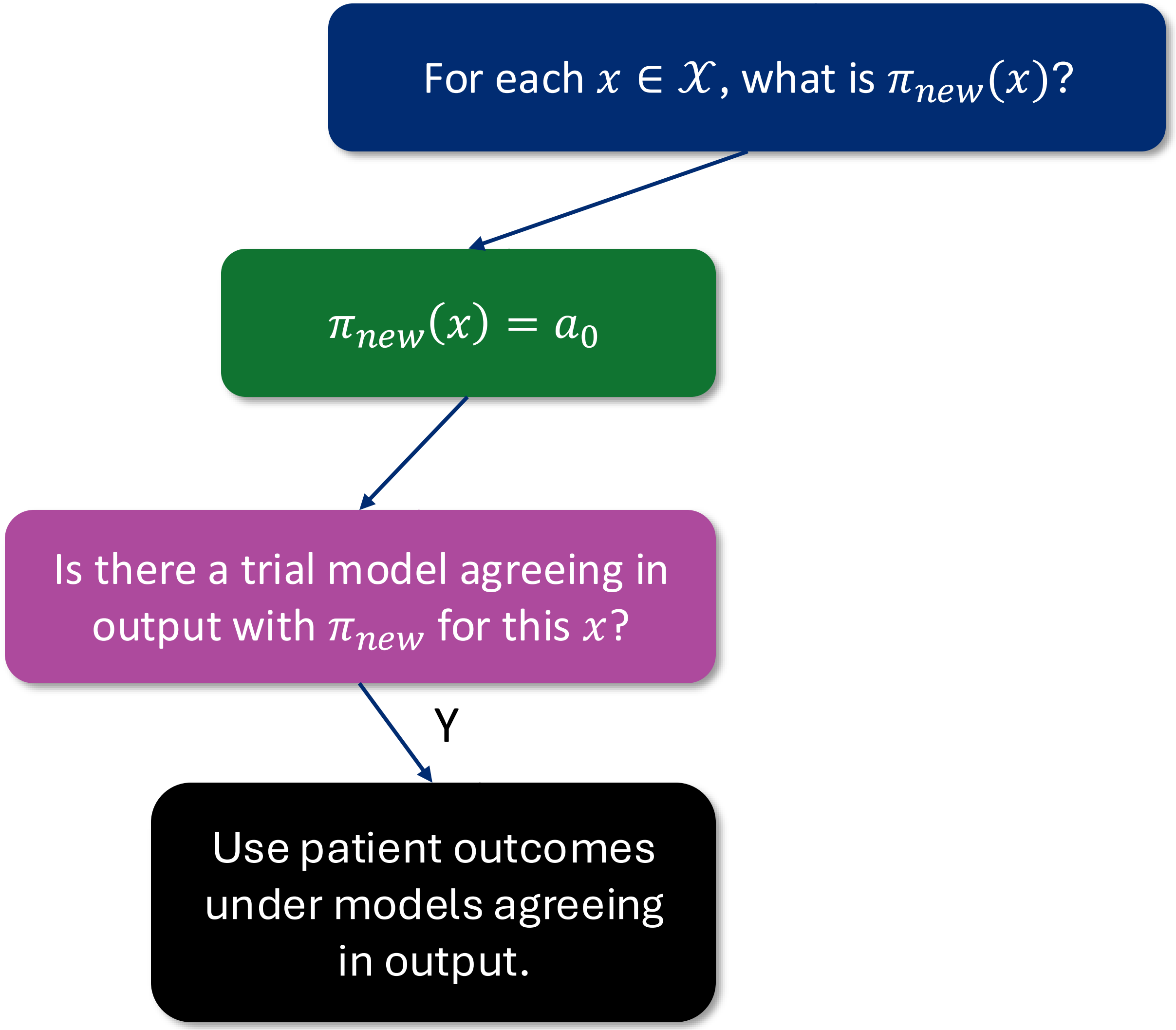
Lower Bound on Causal Impact



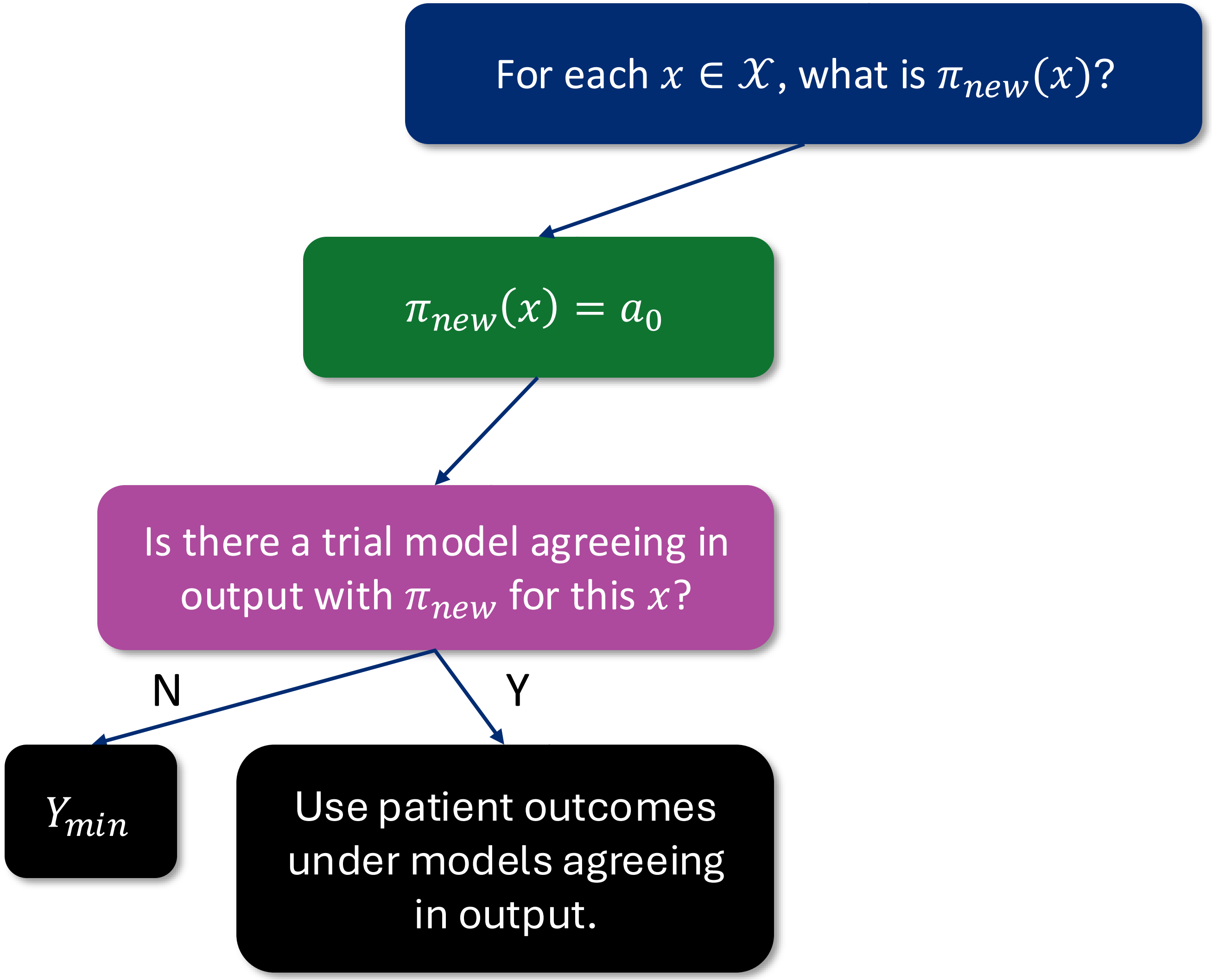
Lower Bound on Causal Impact



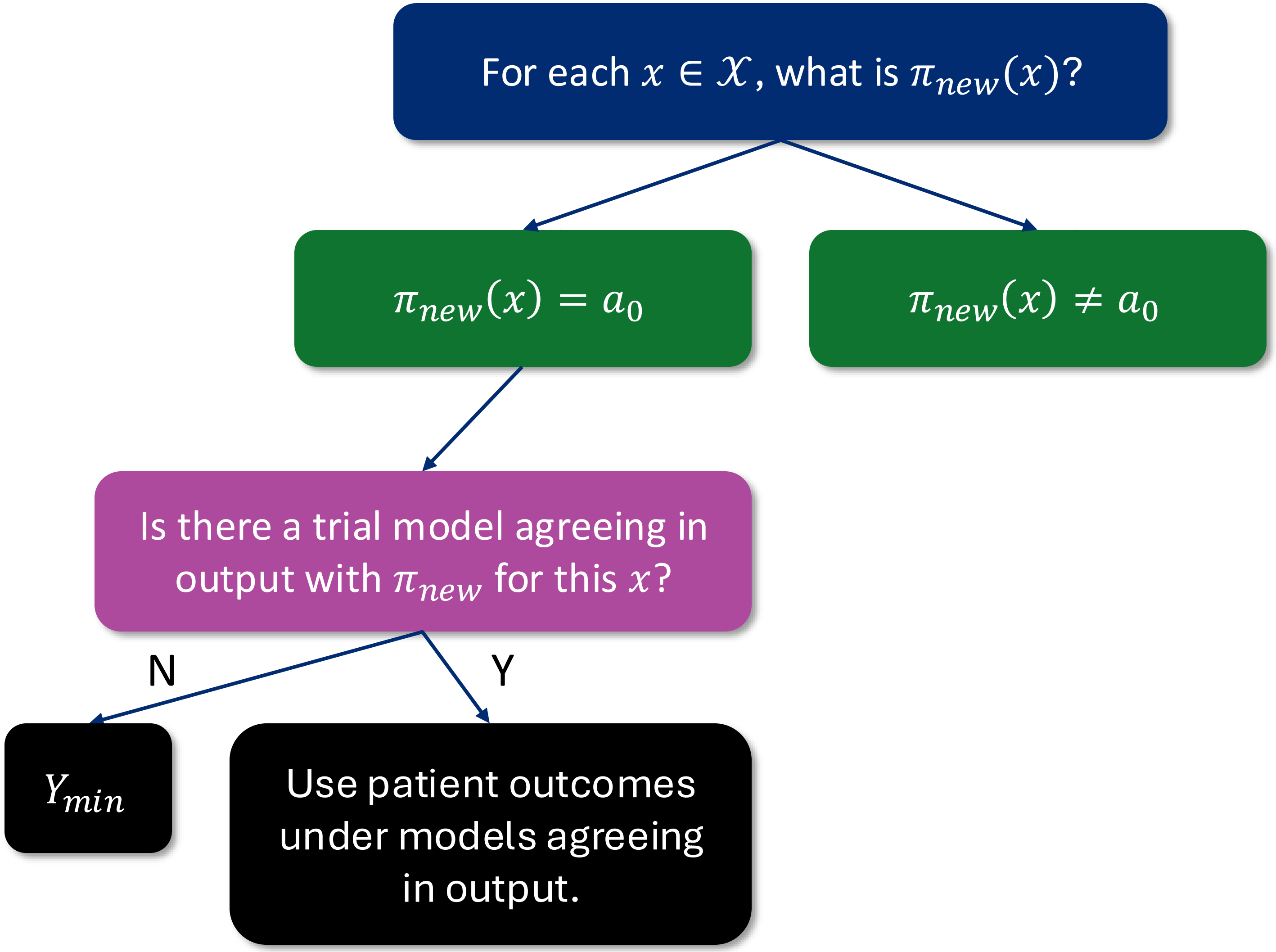
Lower Bound on Causal Impact



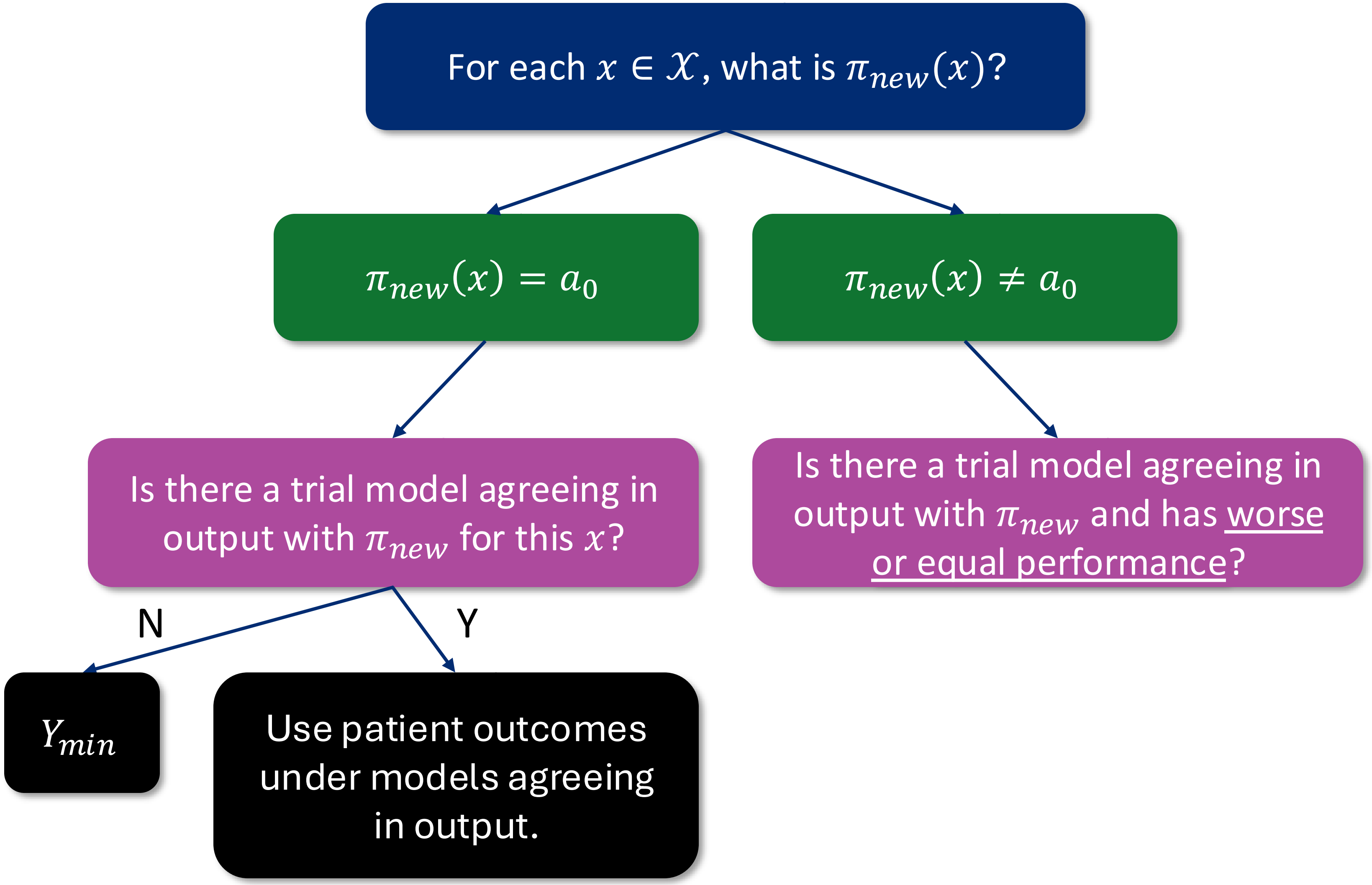
Lower Bound on Causal Impact



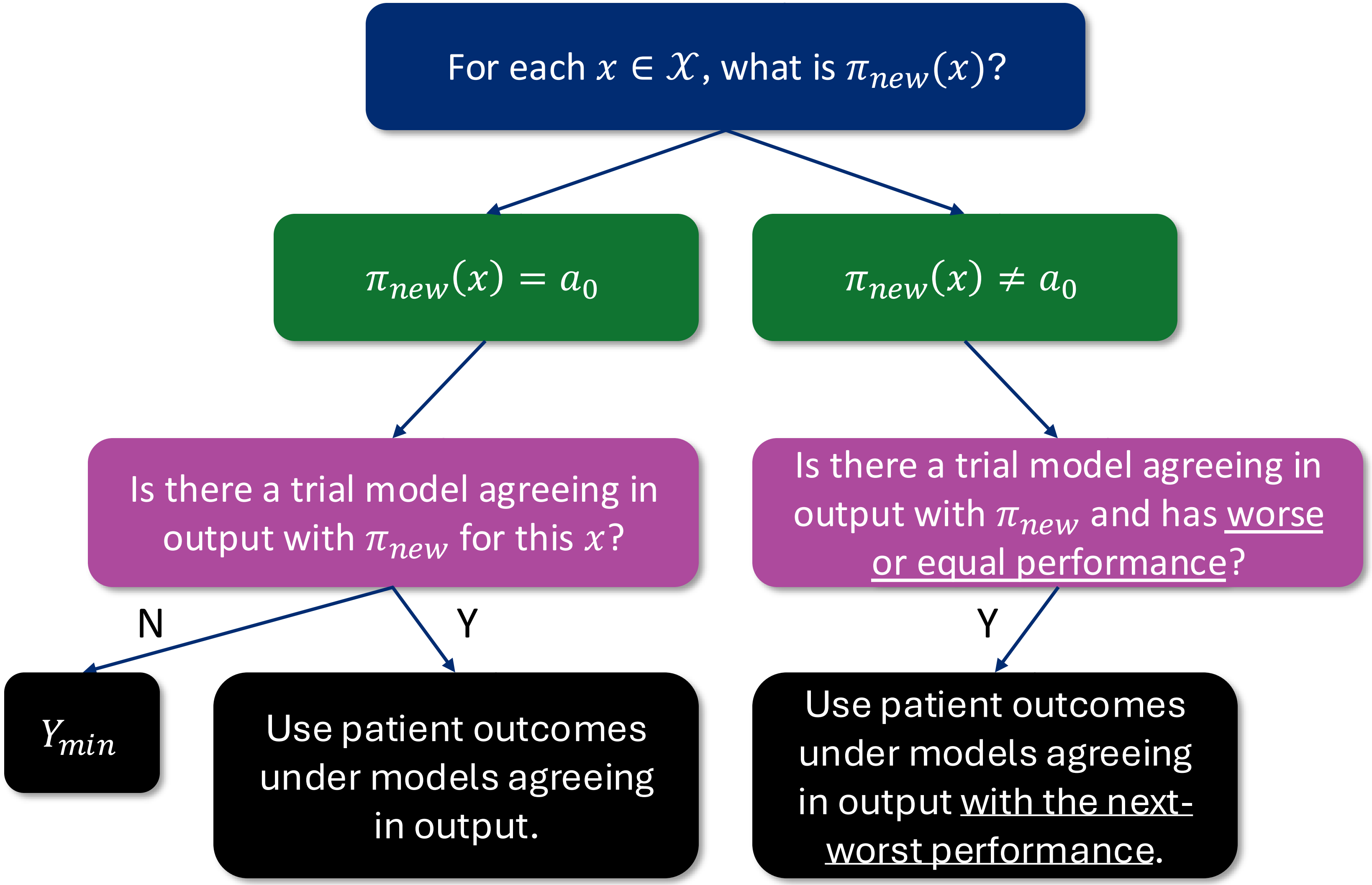
Lower Bound on Causal Impact



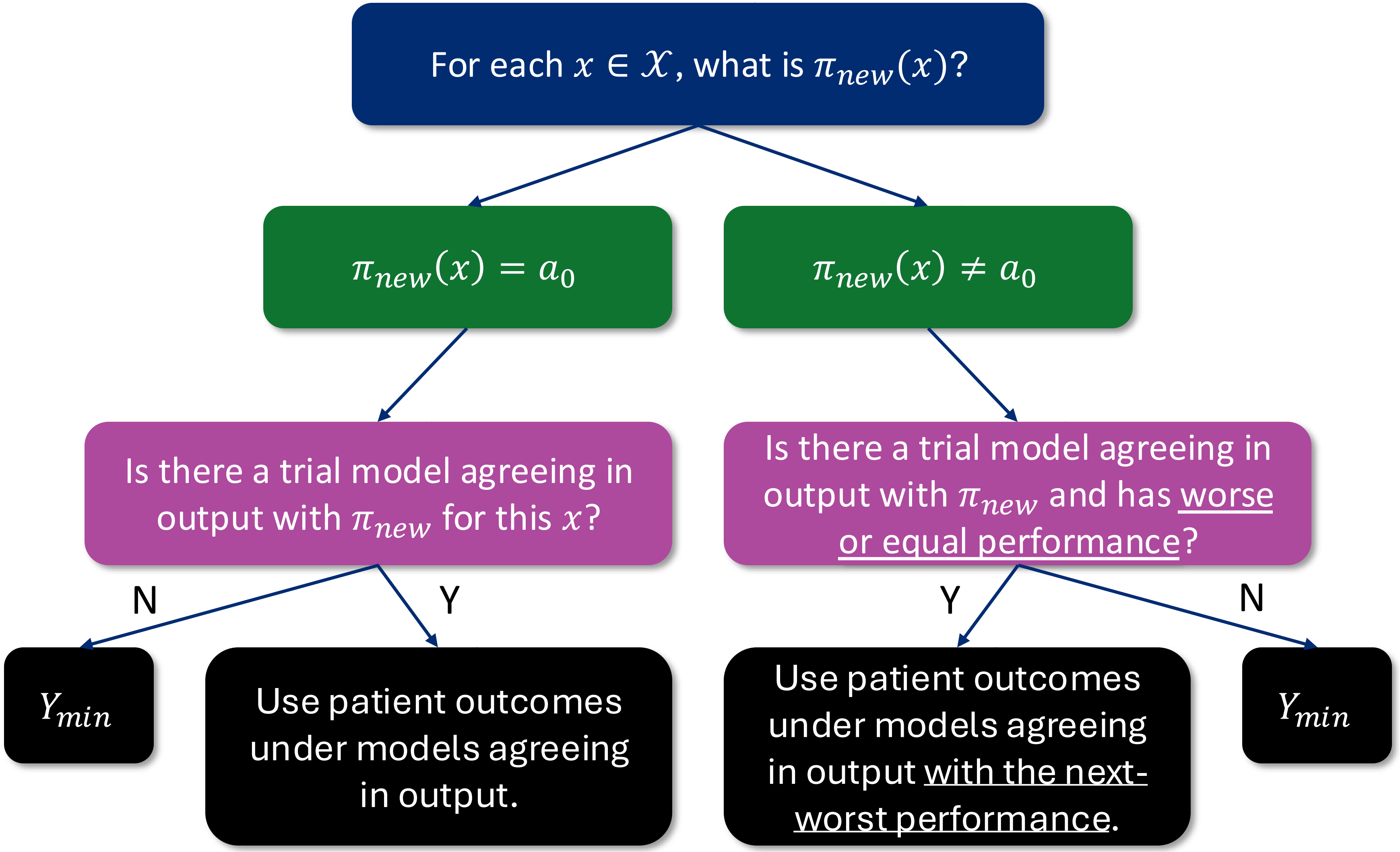
Lower Bound on Causal Impact



Lower Bound on Causal Impact



Lower Bound on Causal Impact



Lower / Upper Bounds on Causal Impact

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].

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

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].

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

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*

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

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

Lower / Upper Bounds on Causal Impact

- We further show that the lower / upper bounds are **tight**, i.e. they cannot be improved without further assumptions (Theorem 3.2).

Lower / Upper Bounds on Causal Impact

- We further show that the lower / upper bounds are **tight**, i.e. they cannot be improved without further assumptions (Theorem 3.2).
- We give inverse-probability weighted (IPW) estimators for the bounds with asymptotically valid confidence intervals (Proposition 3.4).

Lower / Upper Bounds on Causal Impact

- We further show that the lower / upper bounds are **tight**, i.e. they cannot be improved without further assumptions (Theorem 3.2).
- We give inverse-probability weighted (IPW) estimators for the bounds with asymptotically valid confidence intervals (Proposition 3.4).
- The more "similar" that the trialed models in the RCT are to the new model in coverage and performance metric, the tighter the bounds will be.

Summary

Summary

- We propose a framework and method for **estimating / bounding the causal impact of deploying a new ML model** from RCT data where the new model was never trialed.

Summary

- We propose a framework and method for **estimating / bounding the causal impact of deploying a new ML model** from RCT data where the new model was never trialed.
- Our bounds rely on assumptions, but these **assumptions are falsifiable using RCT data**, given that multiple models were trialed.

Summary

- We propose a framework and method for **estimating / bounding the causal impact of deploying a new ML model** from RCT data where the new model was never trialed.
- Our bounds rely on assumptions, but these **assumptions are falsifiable using RCT data**, given that multiple models were trialed.
- One implication of results: **trial multiple models in cluster RCTs**. This allows for falsification of assumptions and alleviates challenges related to coverage and performance.

Summary

- We propose a framework and method for **estimating / bounding the causal impact of deploying a new ML model** from RCT data where the new model was never trialed.
- Our bounds rely on assumptions, but these **assumptions are falsifiable using RCT data**, given that multiple models were trialed.
- One implication of results: **trial multiple models in cluster RCTs**. This allows for falsification of assumptions and alleviates challenges related to coverage and performance.
- Potential use case: **bounding causal impacts of model updates** before trialing new model updates in RCTs.

Summary

- We propose a framework and method for **estimating / bounding the causal impact of deploying a new ML model** from RCT data where the new model was never trialed.
- Our bounds rely on assumptions, but these **assumptions are falsifiable using RCT data**, given that multiple models were trialed.
- One implication of results: **trial multiple models in cluster RCTs**. This allows for falsification of assumptions and alleviates challenges related to coverage and performance.
- Potential use case: **bounding causal impacts of model updates** before trialing new model updates in RCTs.
- A step towards reliable re-use of RCT data evaluating ML models.

Thank you for your attention!

- Please join us at our poster this afternoon from 16:00-18:30!

Please scan the QR code for the arXiv link to our paper.

Backup Slides

Inverse Probability Weighted-Style Estimators

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

Notation: Partitions of Model Sets

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].

$$\mathbf{\Pi}^e(x) := \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x)\}$$ (1)

$$\mathbf{\Pi}^e_{\leq}(x) := \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x), f_M(\pi) \leq f_M(\pi_e)\}$$

$$\mathbf{\Pi}^e_{\geq}(x) := \{\pi \in \mathit{\Pi} \mid \pi(x) = \pi_e(x), f_M(\pi) \geq f_M(\pi_e)\}$$

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*.

Notation: Partitions of Model Sets

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].

$$\mathbf{\Pi}^e(x) := \{\pi \in \Pi \mid \pi(x) = \pi_e(x)\}$$ (1)

$$\mathbf{\Pi}^e_{\leq}(x) := \{\pi \in \Pi \mid \pi(x) = \pi_e(x), f_M(\pi) \leq f_M(\pi_e)\}$$ (2)

$$\mathbf{\Pi}^e_{\geq}(x) := \{\pi \in \Pi \mid \pi(x) = \pi_e(x), f_M(\pi) \geq f_M(\pi_e)\}$$

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.

Notation: Partitions of Model Sets

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].

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

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

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

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 <u>equal or worse</u> performance.
3. Subset of models that agree in output with the new model for a patient *x* AND has <u>equal or better</u> performance.

Notation: Partitions of Model Sets

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].

$$\boldsymbol{\Pi}^e(x) := \{\pi \in \Pi \mid \pi(x) = \pi_e(x)\} \quad (1)$$

$$\boldsymbol{\Pi}^e_{\leq}(x) := \{\pi \in \Pi \mid \pi(x) = \pi_e(x), f_M(\pi) \leq f_M(\pi_e)\} \quad (2)$$

$$\boldsymbol{\Pi}^e_{\geq}(x) := \{\pi \in \Pi \mid \pi(x) = \pi_e(x), f_M(\pi) \geq f_M(\pi_e)\} \quad (3)$$

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

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

$$\tilde{\boldsymbol{\Pi}}^e_{\geq}(x) := \arg\min_{\pi \in \boldsymbol{\Pi}^e_{\geq}(x)} f_M(\pi)$$

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.

Lower / Upper Bounds on Causal Impact

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*

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

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

(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.

Lower / Upper Bounds on Causal Impact

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*

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

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

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.

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

Lower / Upper Bounds on Causal Impact

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*

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

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

(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.

Lower / Upper Bounds on Causal Impact

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*

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

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

(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.

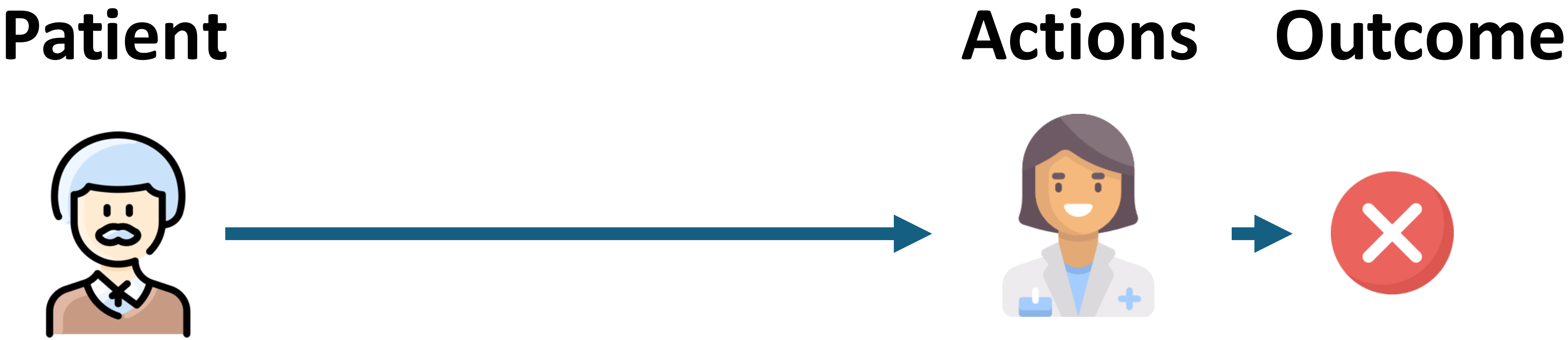
(4) Otherwise, lower bound by the lowest possible value of the outcome.

Randomized Controlled Trials (RCTs) help us compare between two scenarios

Compare no deployment of ML model vs. deployment of ML model.

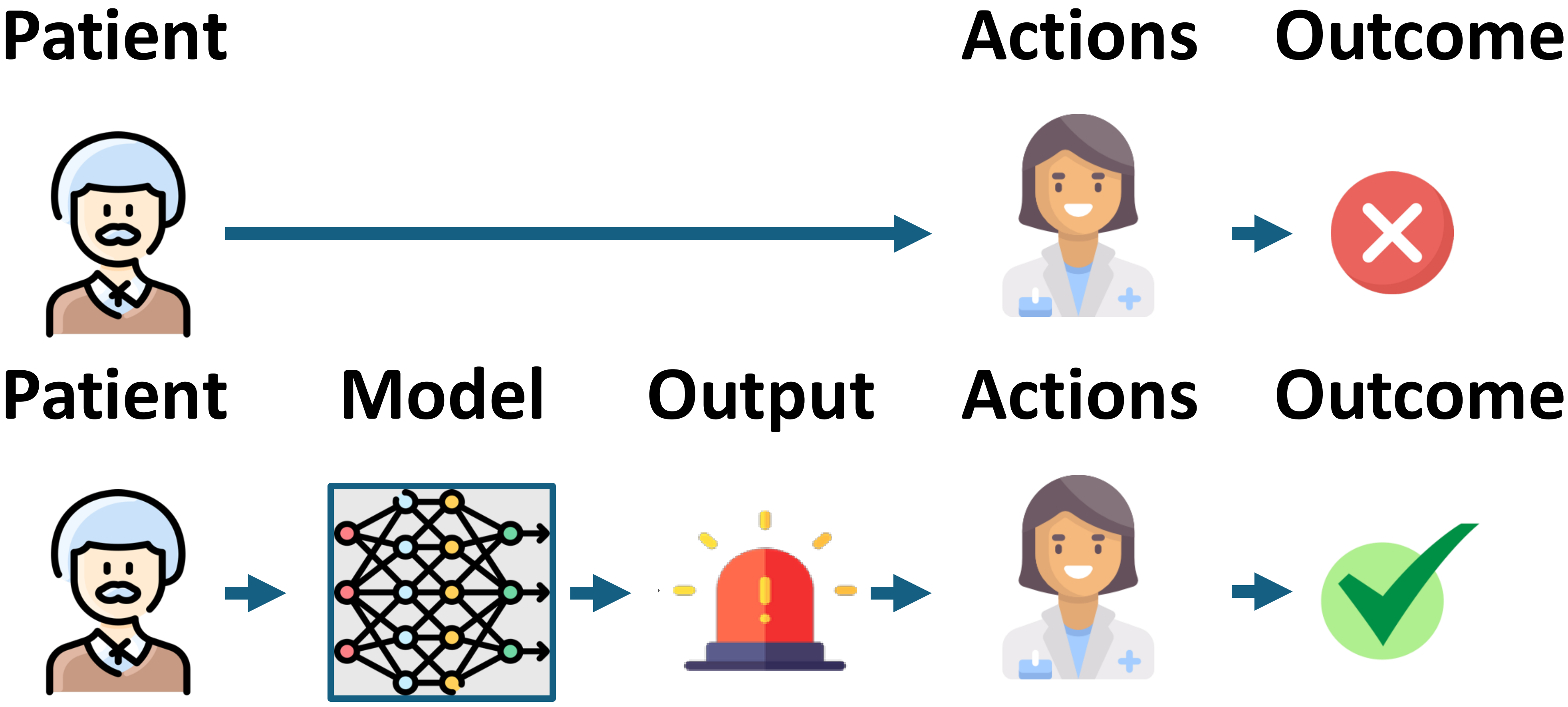
Randomized Controlled Trials (RCTs) help us compare between two scenarios

Compare *no deployment of ML model* vs. deployment of ML model.

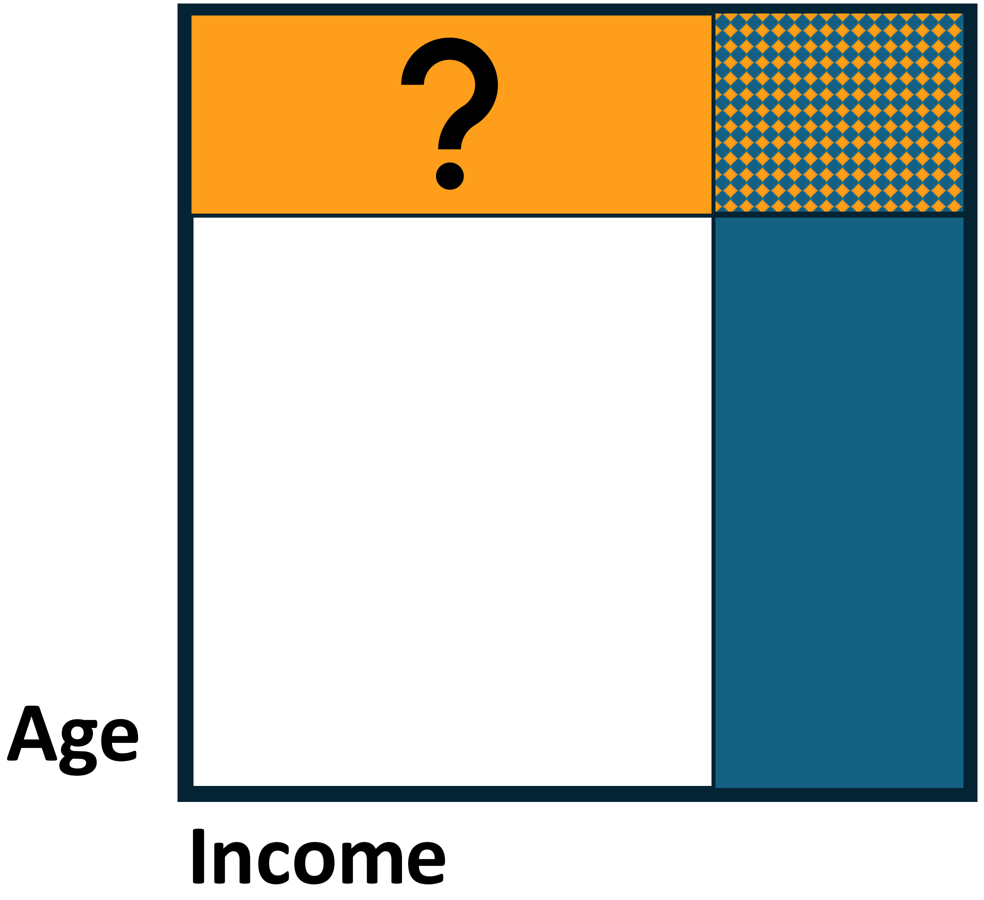


Randomized Controlled Trials (RCTs) help us compare between two scenarios

Compare no deployment of ML model vs. *deployment of ML model*.

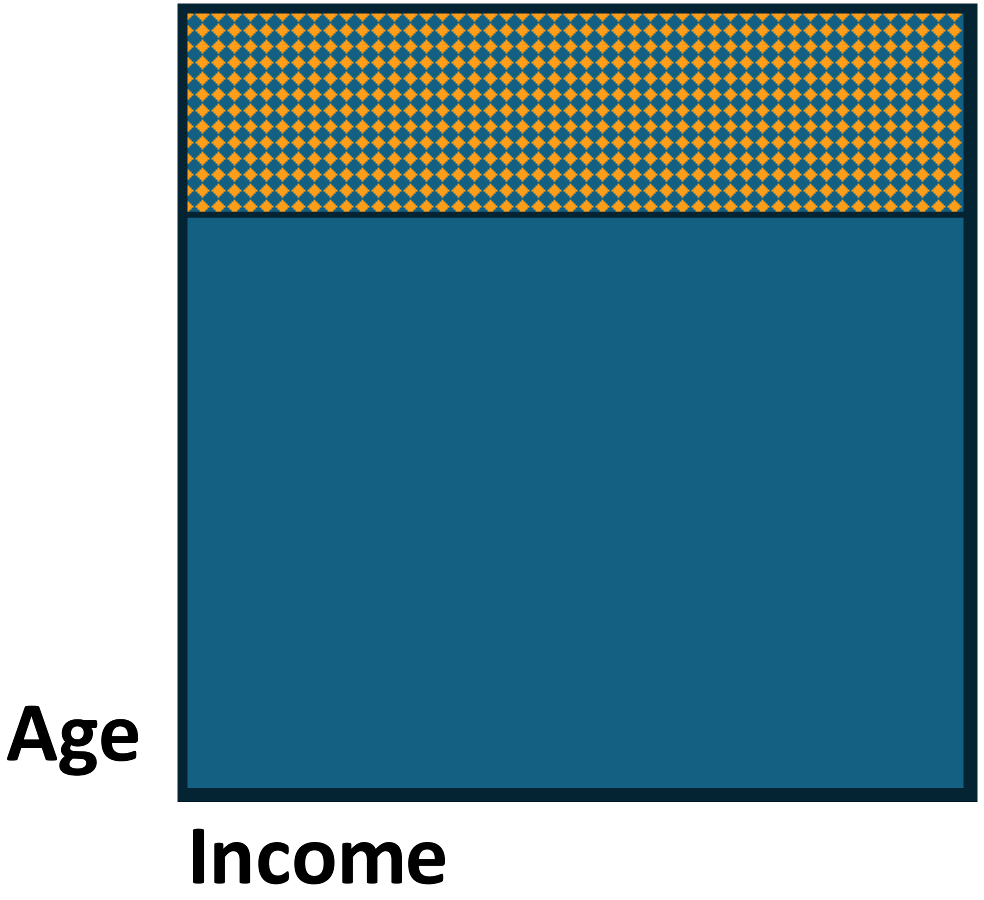


Why Require the Performance Assumption?



Recall the coverage challenge.

Why Require the Performance Assumption?



Recall the coverage challenge.

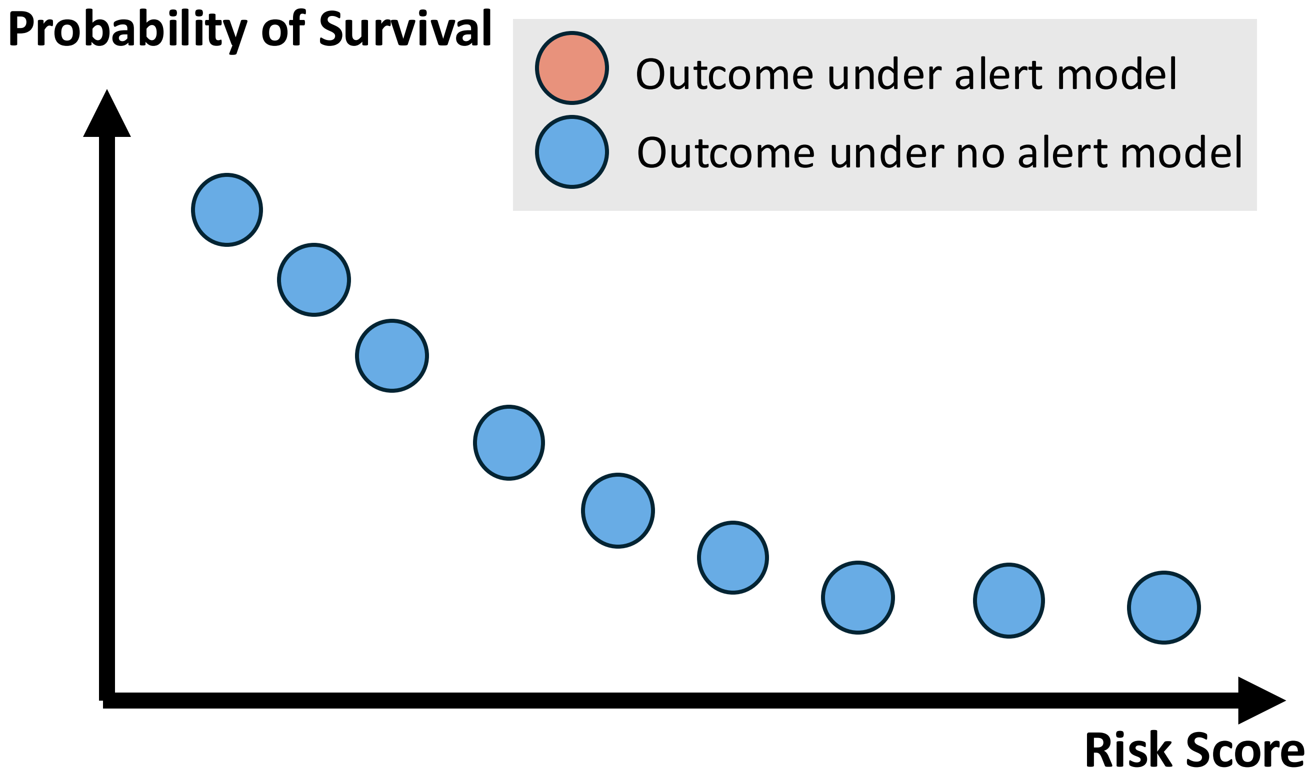
A simple solution: have the trialed model (blue) cover the whole square.

Why Require the Performance Assumption?

- Why is trialing a model that always alerts and a model that never alerts (the control arm) a bad idea?

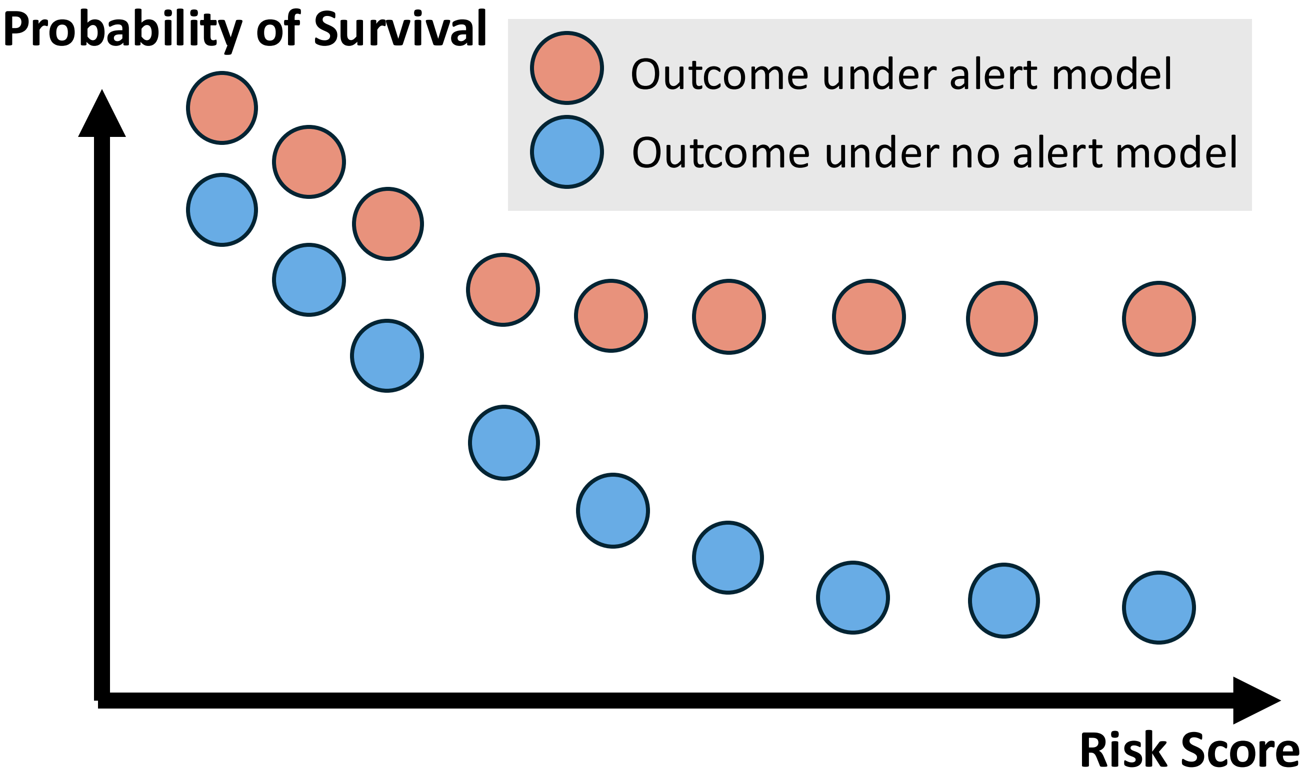
Why Require the Performance Assumption?

- Why is trialing a model that always alerts and a model that never alerts (the control arm) a bad idea?



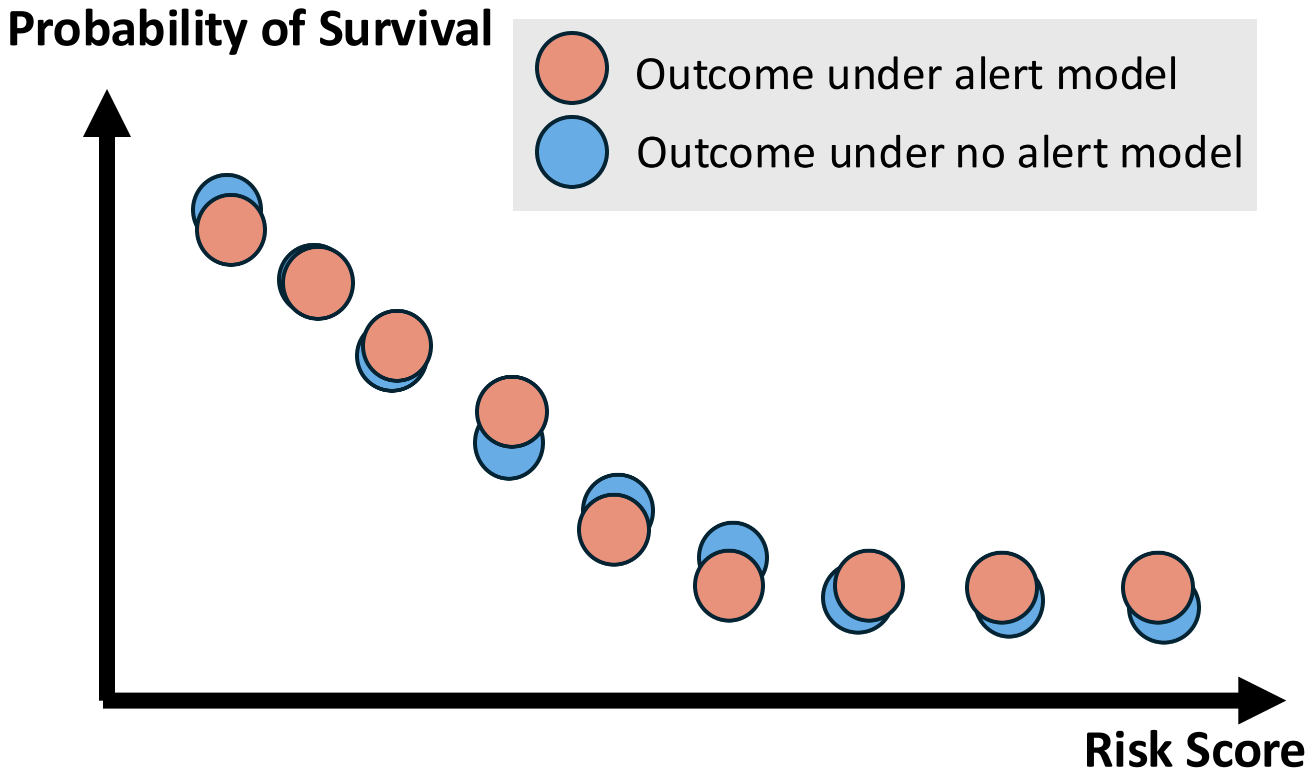
Why Require the Performance Assumption?

- Why is trialing a model that always alerts and a model that never alerts (the control arm) a bad idea?



Why Require the Performance Assumption?

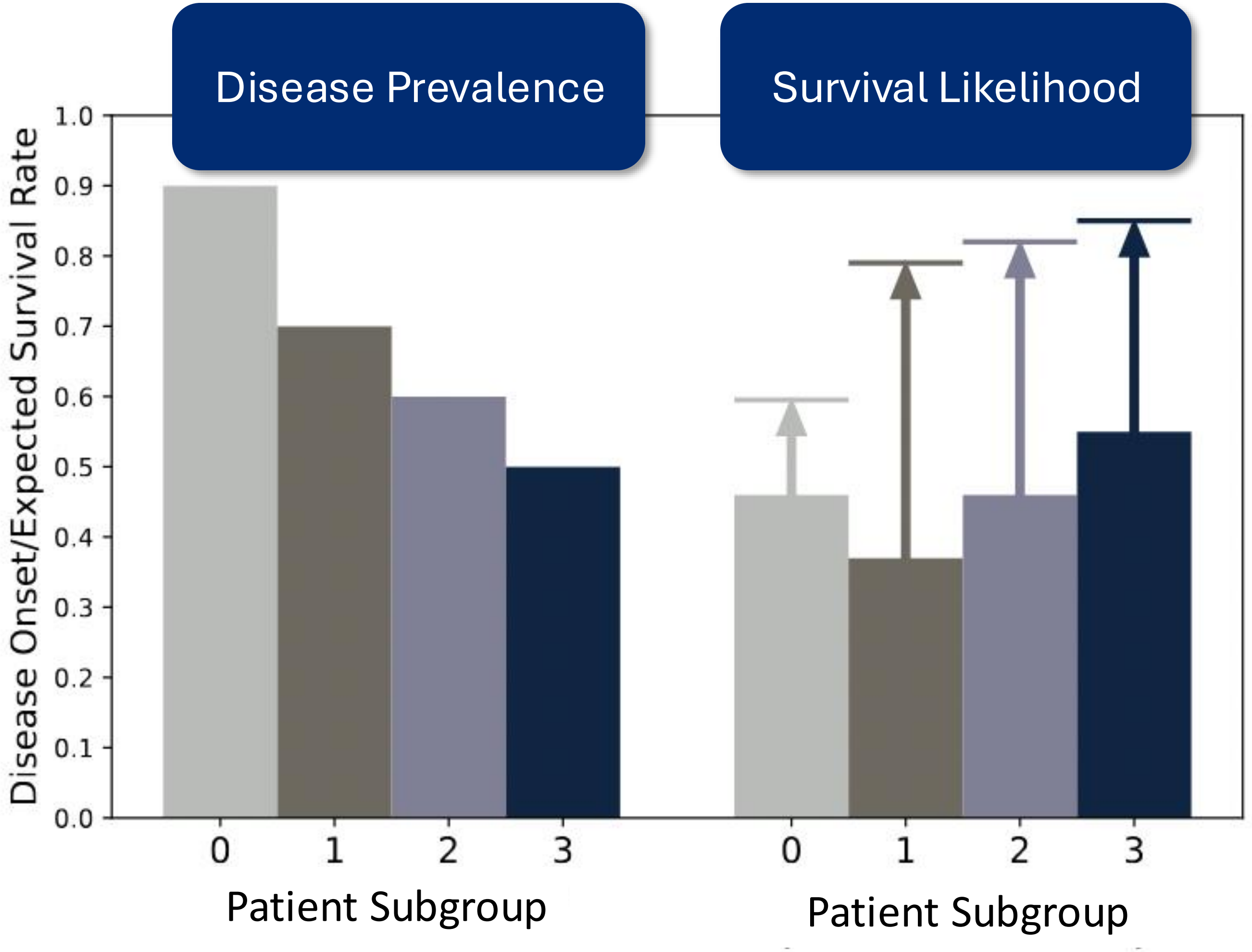
- Why is trialing a model that always alerts and a model that never alerts (the control arm) a bad idea?
- This "always alert" model will likely have minimal impact due to its poor performance.



Synthetic Simulation Study

Setup

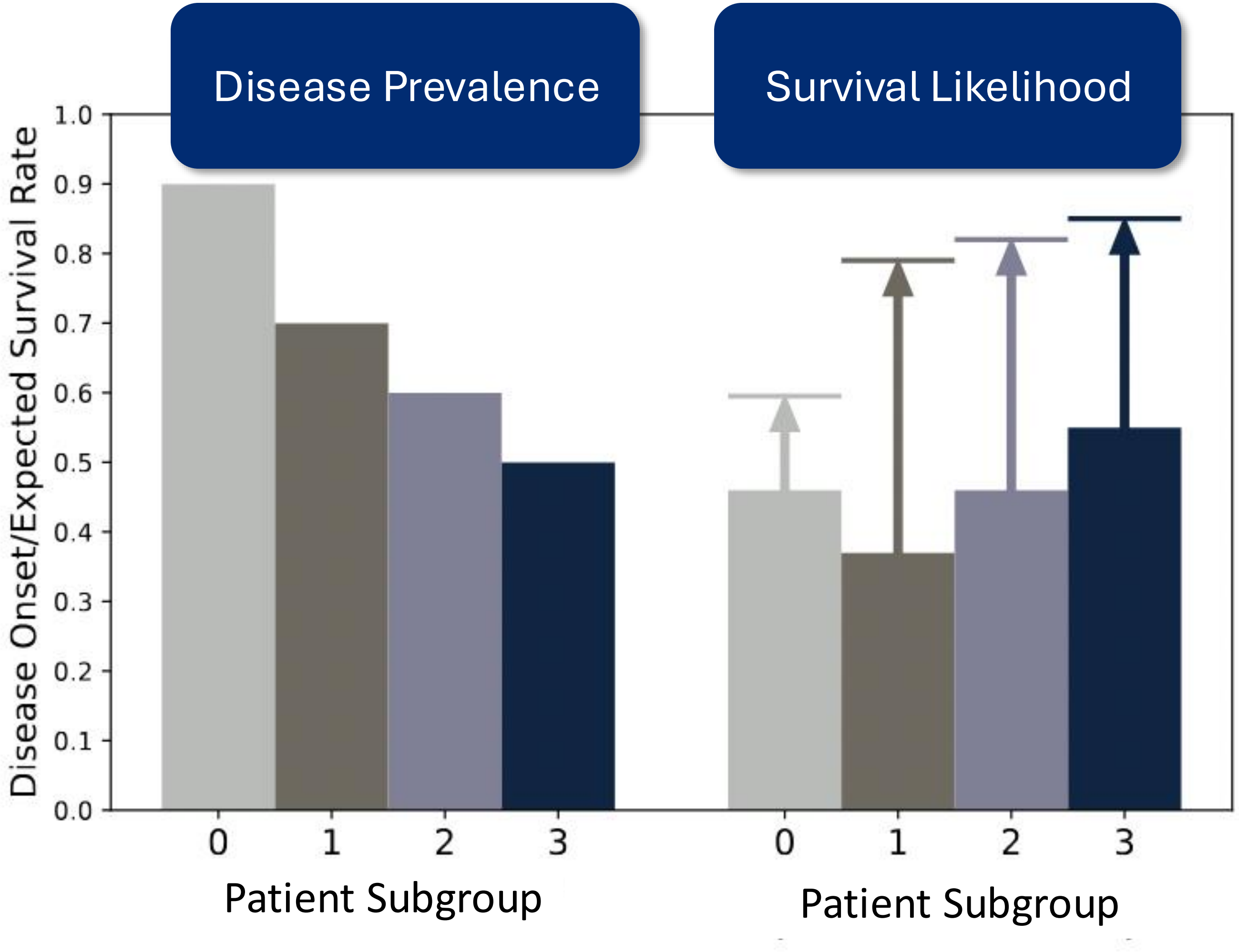
- Four types of patients with varying likelihoods of developing disease and survival rates.



Synthetic Simulation Study

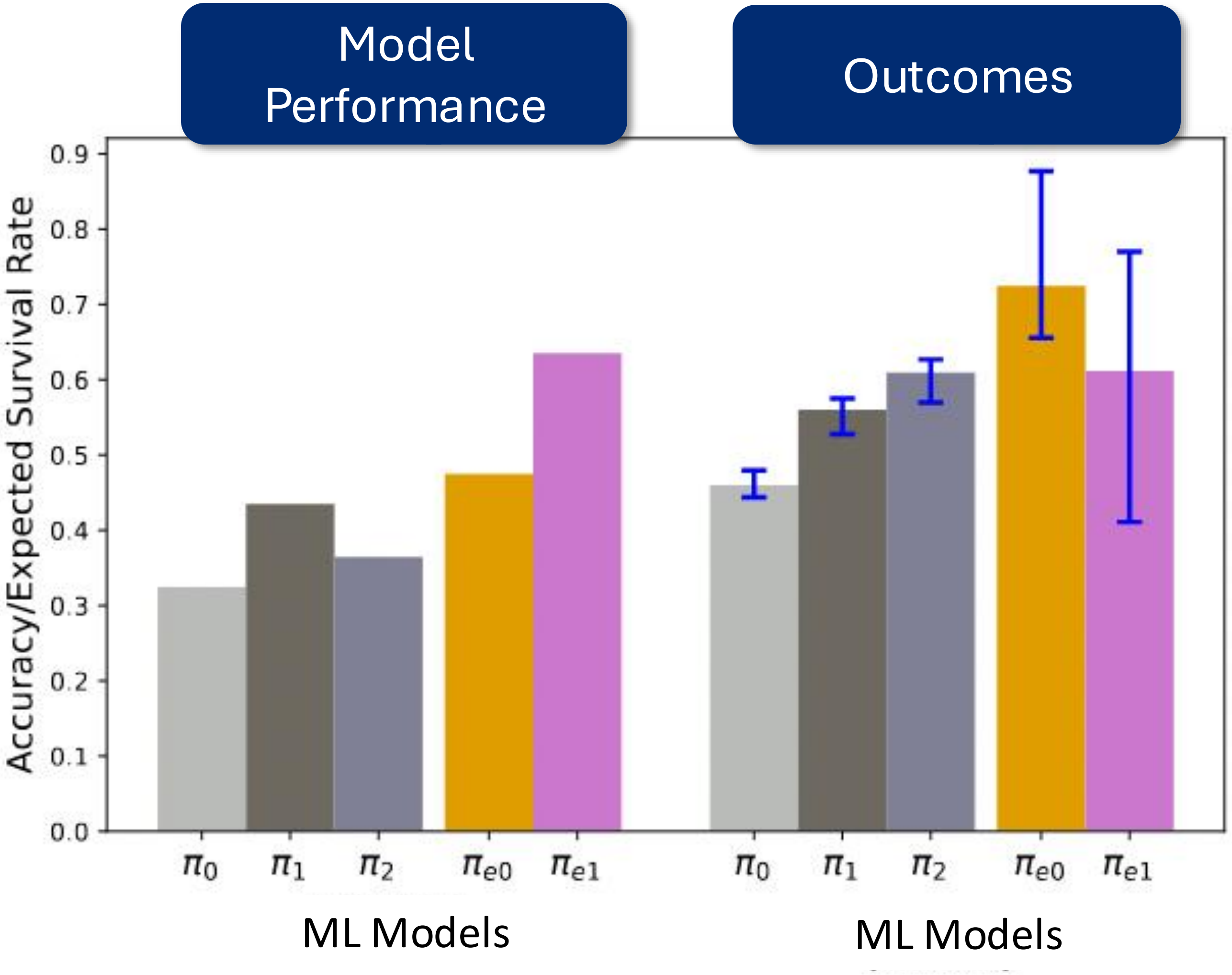
Setup

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



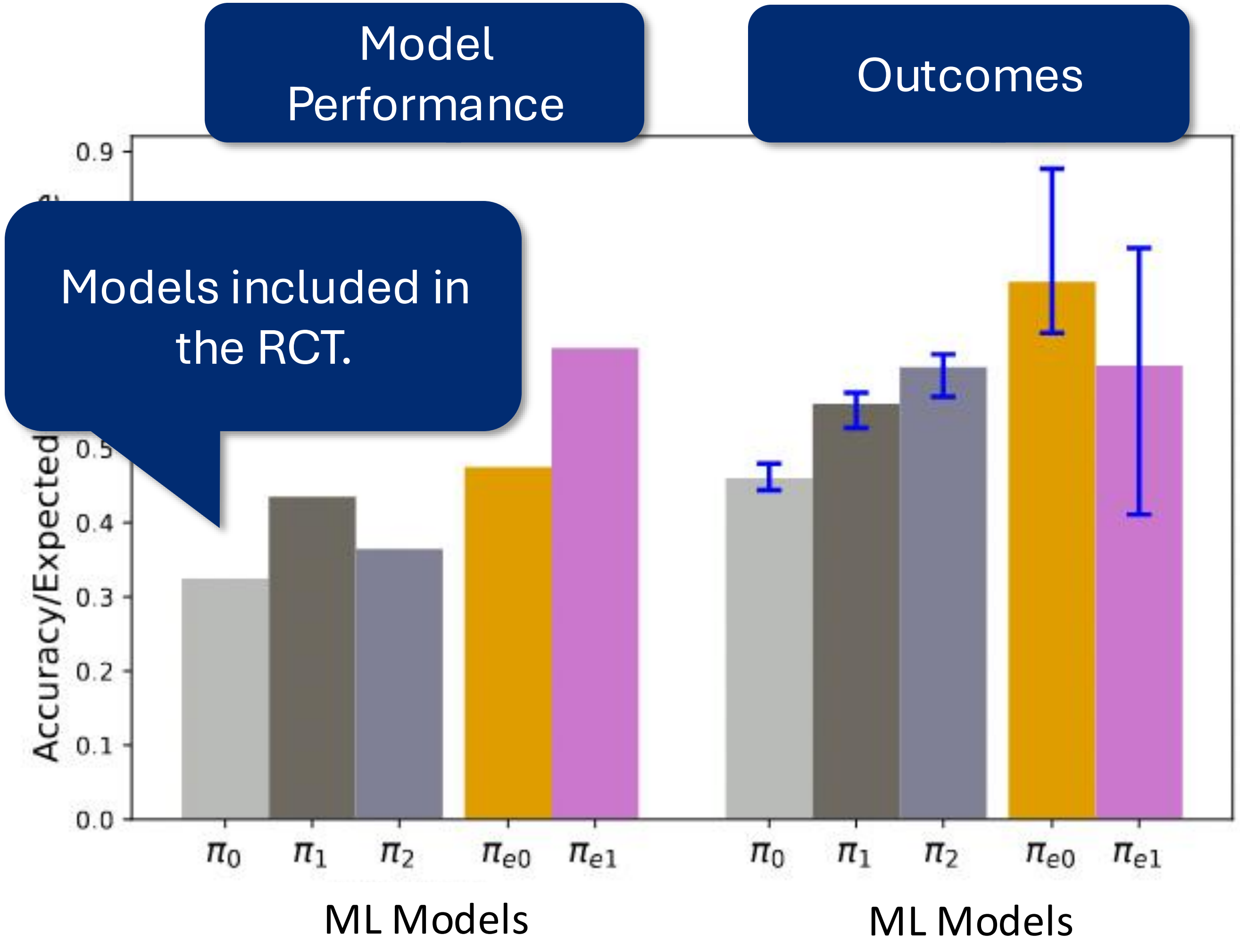
Synthetic Simulation Study

Results



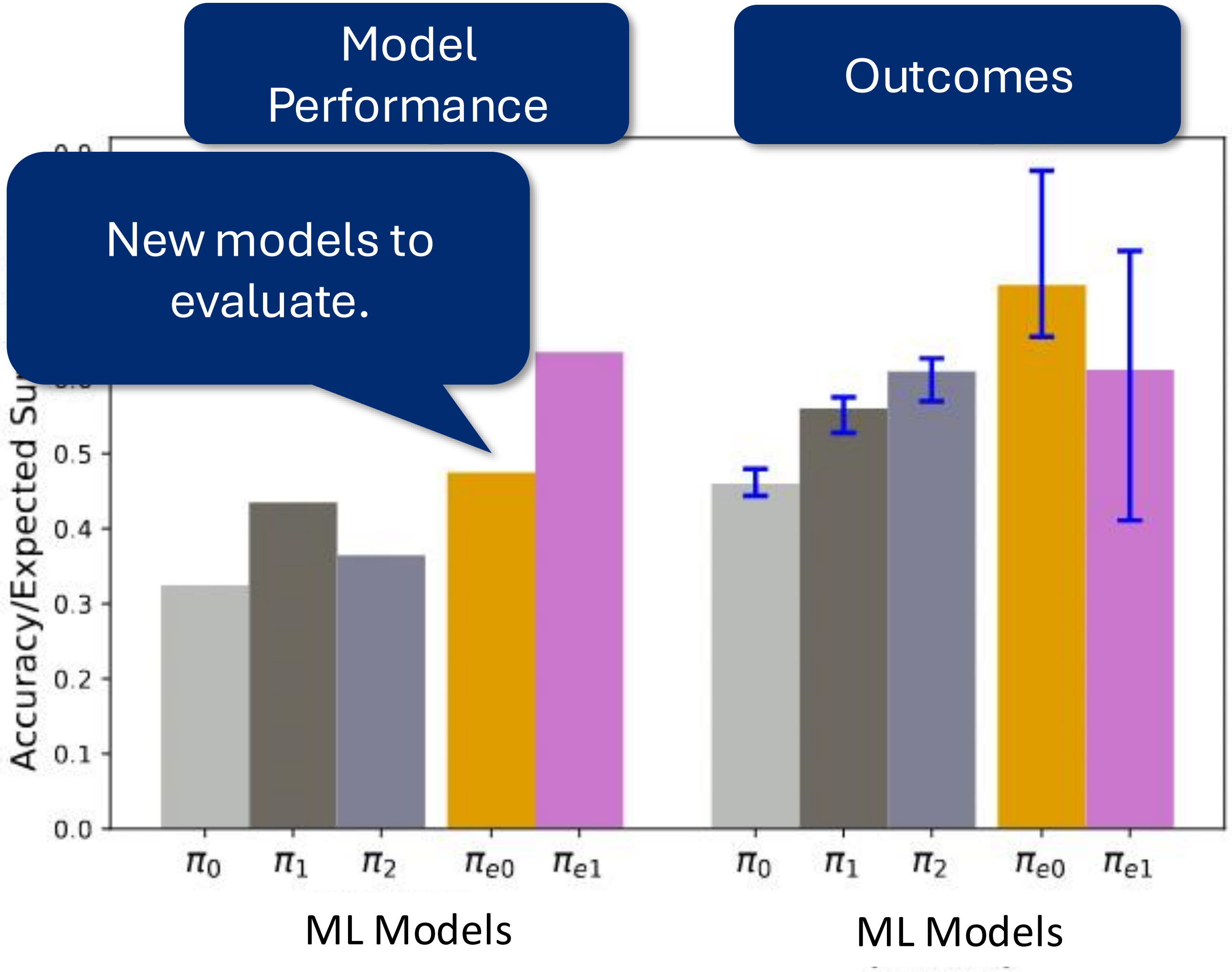
Synthetic Simulation Study

Results



Synthetic Simulation Study

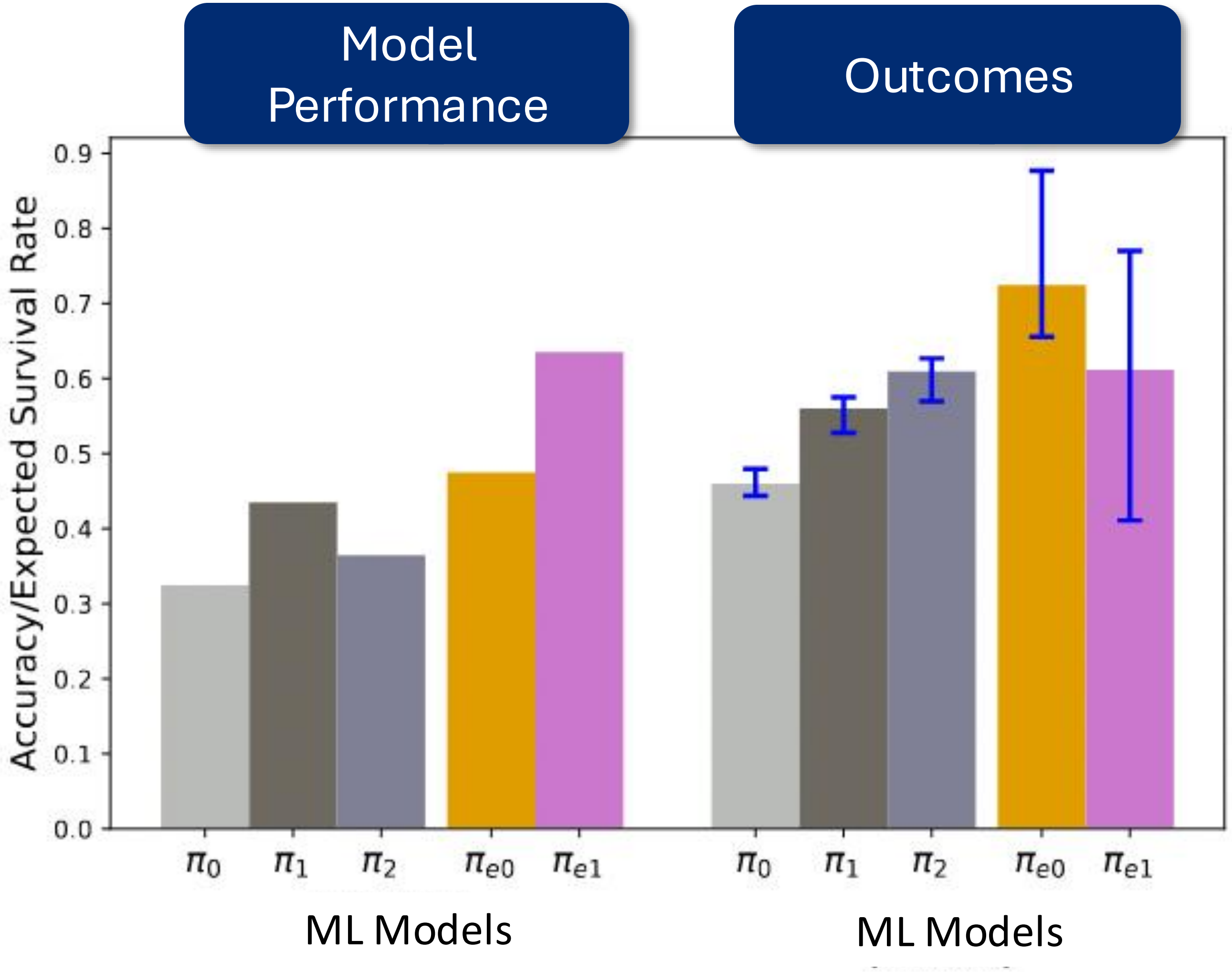
Results



Synthetic Simulation Study

Results

- Model performance is the raw accuracy of the model in predicting disease onset.
- Bars indicate ground truth, and intervals indicate statistical uncertainty.

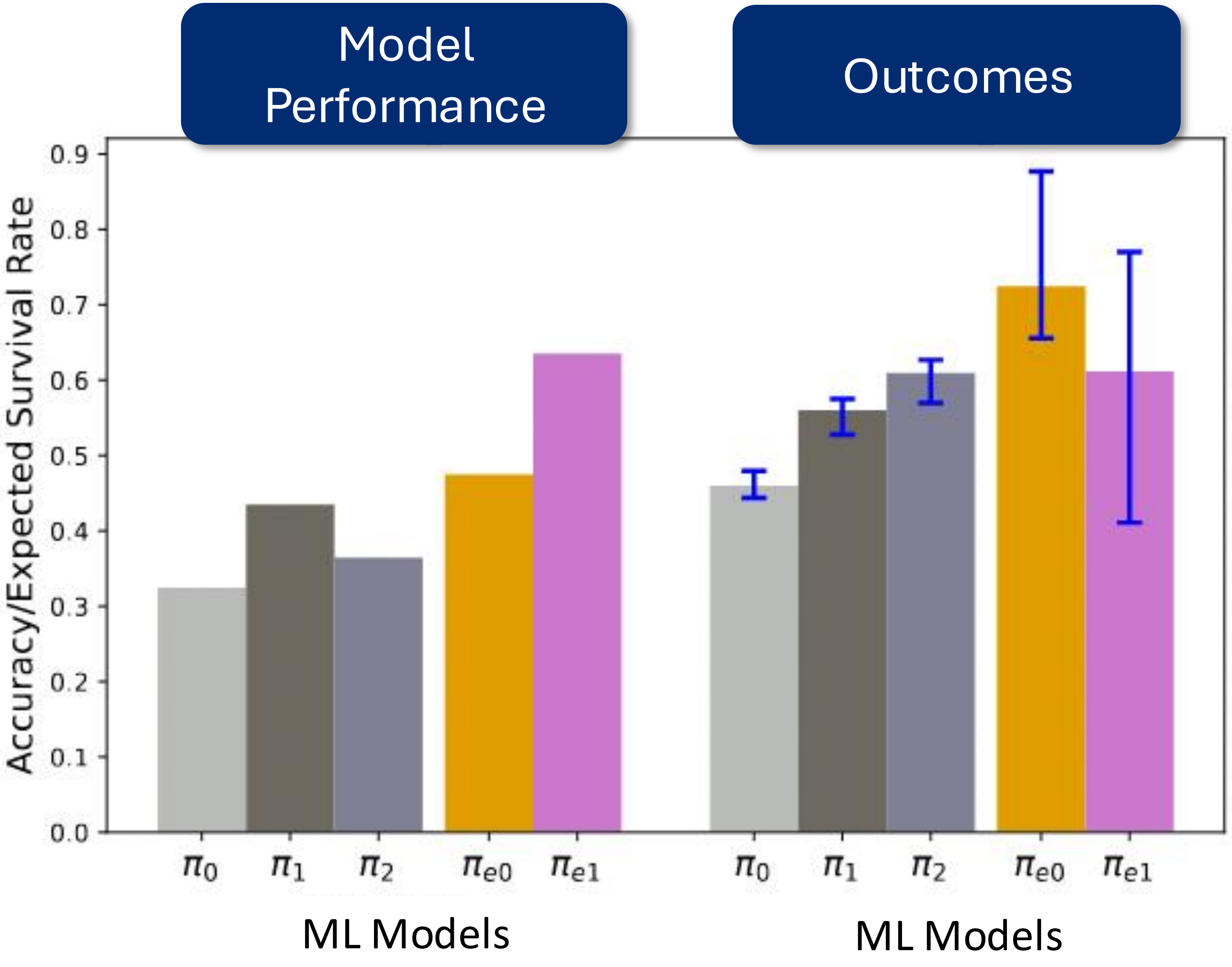


Synthetic Simulation Study

Results

- Model performance is the raw accuracy of the model in predicting disease onset.
- Bars indicate ground truth, and intervals indicate statistical uncertainty.

Model accuracy is not indicative of **causal impact**.



Machine Learning (ML) Models as Medical Devices

Artificial intelligence and machine learning models are increasingly deployed in high-risk domains such as healthcare.

Artificial Intelligence and Machine Learning in Software as a Medical Device

Artificial intelligence (AI) and machine learning (ML) technologies have the potential to transform health care by deriving new and important insights from the vast amount of data generated during the delivery of health care every day. Medical device manufacturers are using these technologies to innovate their products to better assist health care providers and improve patient care. The complex and dynamic processes involved in the development, deployment, use, and maintenance of AI technologies benefit from careful management throughout the medical product life cycle.

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 June 29th, 2025.

Need for More RCTs of ML/AI Models

NEJM AI

NEJM AI 2024;1(11)
DOI: 10.1056/AIe2400881

EDITORIAL

We Need More Randomized Clinical Trials of AI

David Ouyang, M.D.,[1] and Joseph Hogan, Sc.D.[2]

Received: September 5, 2024; Accepted: September 10, 2024; Published: October 18, 2024

Abstract

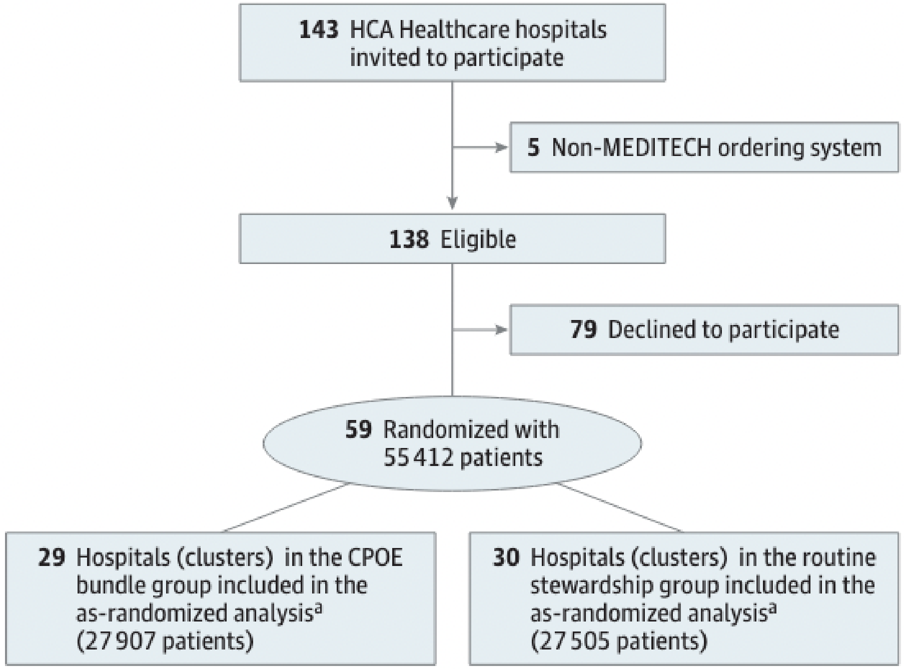
In the first prospective clinical trial of artificial intelligence (AI) assistance in stress echocardiography, there was no difference in diagnostic accuracy between AI assistance and standard-of-care assessment. There is significant value in conducting prospective clinical trials of AI, and there are lessons on implementation to be learned from this study.

David Ouyang and Joseph Hogan. We need more randomized clinical trials of AI, 2024. URL: https://ai.nejm.org/doi/pdf/10.1056/AIe2400881.

Recent RCTs of ML/AI Models

Example: INSPIRE trial for improving antibiotic prescriptions using model-driven best-practice alerts.

Figure 1. Hospital Recruitment and Randomization in the INSPIRE Urinary Tract Infection Trial



MEDITECH is a hospital electronic health record system. CPOE indicates computerized provider order entry and INSPIRE, Intelligent Stewardship Prompts to Improve Real-time Empiric antibiotic selection.

[a]All analyses are as-randomized because all hospitals remained in the trial until end of intervention (no hospital withdrawals after enrollment). There was a median (IQR) of 2364 (1277-2963) patients per hospital in the CPOE bundle group and 2008 (1365-3064) in the routine stewardship group.

S. K. Gohil et al. 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.

Recent RCTs of ML/AI Models

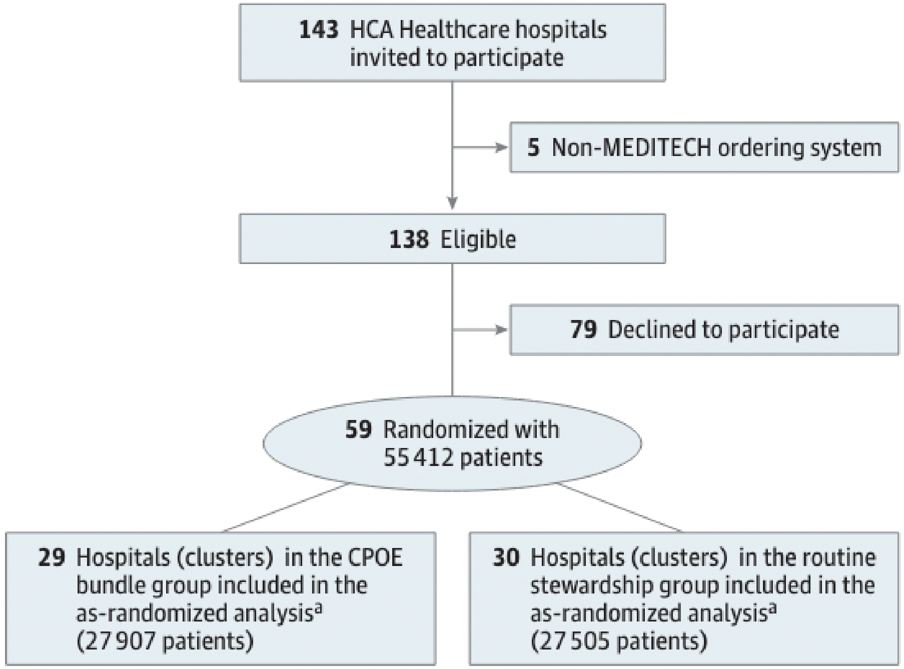
Example: INSPIRE trial for improving antibiotic prescriptions using model-driven best-practice alerts.

Important Features

- **Cluster RCT:** Randomizes hospitals to ML model vs. control.
- **Outcomes:** Compares clinical outcomes between treatment/control groups to assess the impact of model deployment.

S. K. Gohil et al. 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.

Figure 1. Hospital Recruitment and Randomization in the INSPIRE Urinary Tract Infection Trial



MEDITECH is a hospital electronic health record system. CPOE indicates computerized provider order entry and INSPIRE, Intelligent Stewardship Prompts to Improve Real-time Empiric antibiotic selection.

[a] All analyses are as-randomized because all hospitals remained in the trial until end of intervention (no hospital withdrawals after enrollment). There was a median (IQR) of 2364 (1277-2963) patients per hospital in the CPOE bundle group and 2008 (1365-3064) in the routine stewardship group.