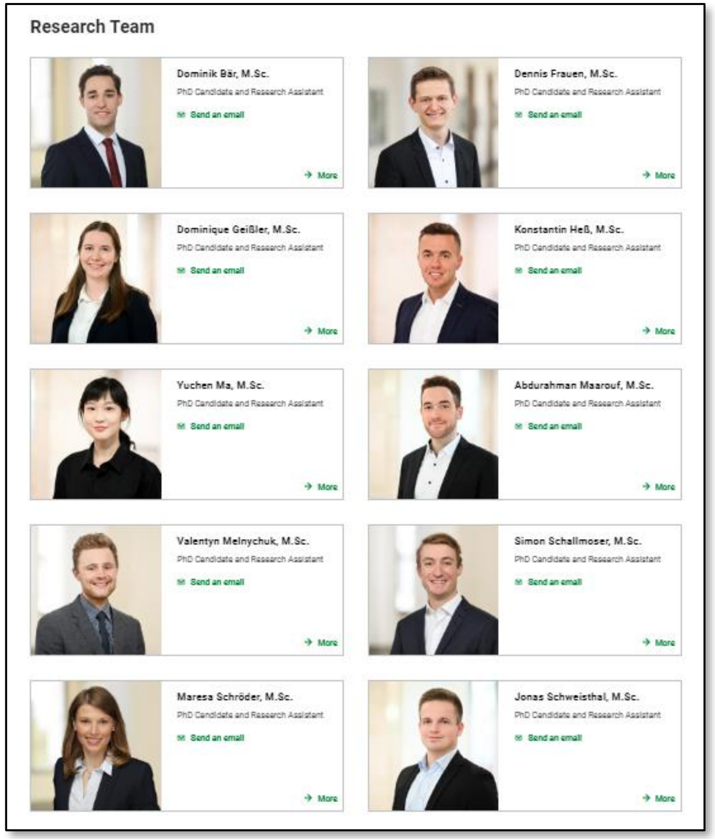
Causal ML for predicting treatment outcomes

Prof. Stefan Feuerriegel
& Valentyn Melnychuk

Institute of AI in Management
LMU Munich
https://www.ai.bwl.lmu.de

ABOUT OUR INSTITUTE

Solving real-world problems with artificial intelligence (AI)



What defines our research

1 **Information**

We solve management problems of **relevance** by using data science

2 **Innovation**

We develop **new** algorithms from the area of AI (statistics, computer science, etc.)

3 **Impact**

We evaluate the added value of our tools **rigorously** in management practice

OUR RESEARCH

We seek to publish in leading outlets from both artificial intelligence and domain applications

Artificial intelligence

- Thought leadership in applied AI ensures research that is **rigorous**
- Provides **state-of-the-art** performance

Example outlets

- NeurIPS, ICML, ICLR, …
- KDD
- ML4H, MLHC, CHIL

Application domains

- Contributes to research that is **relevant**
- Demonstrates **impact in practice**

Example outlets

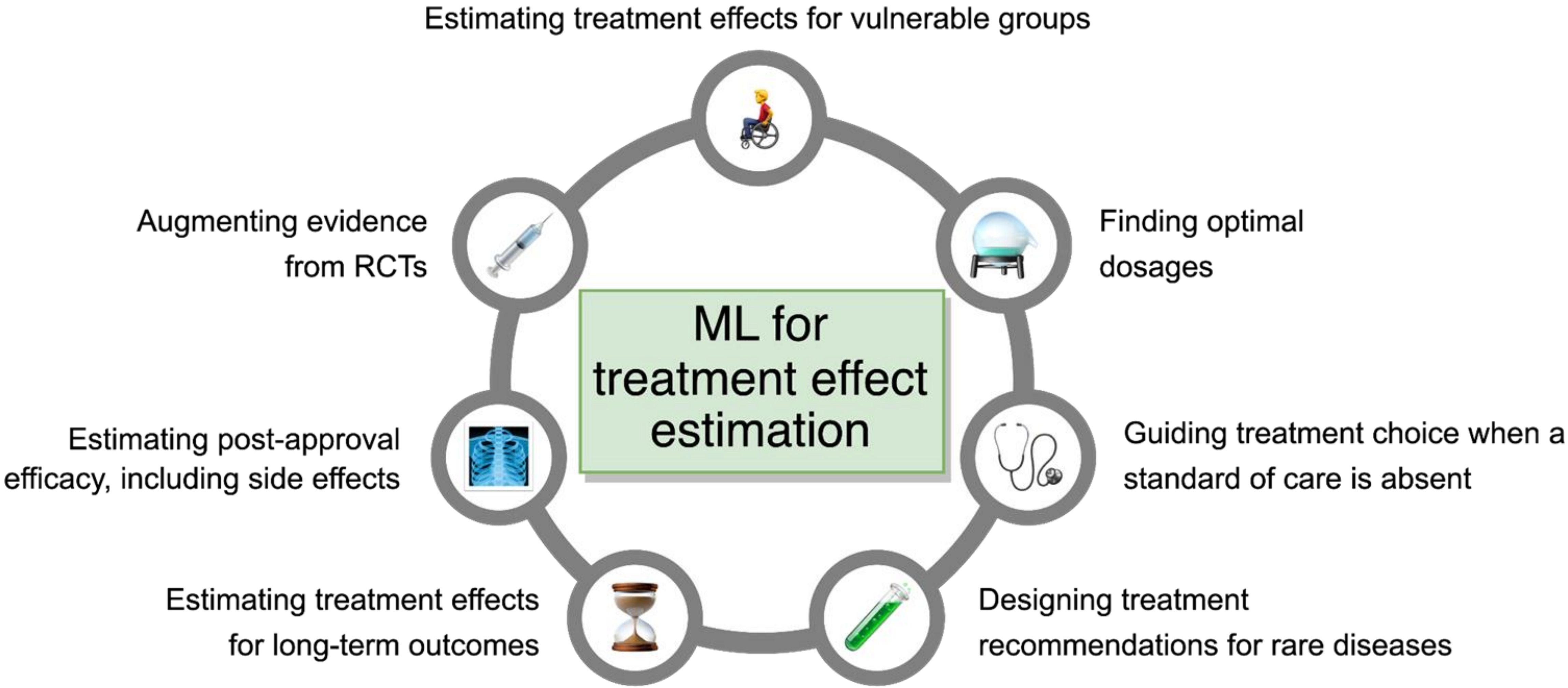
- Nature Medicine
- Lancet Digital Health
- Nature Sustainability
- Nature Human Behaviour
- PNAS

Research collaborations

TEXAS The University of Texas at Austin · NYU · UNIVERSITY OF OXFORD · HARVARD MEDICAL SCHOOL · Berkeley UNIVERSITY OF CALIFORNIA · Carnegie Mellon University · ETH

VISION

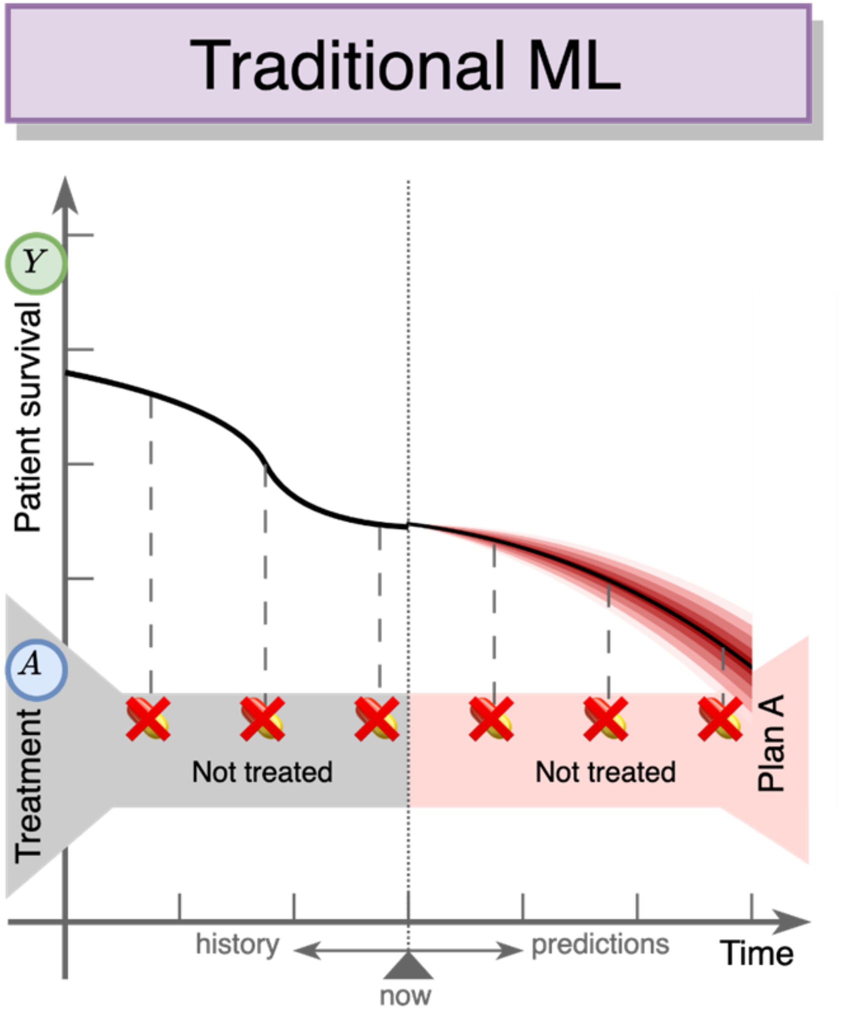
Promises of Causal ML

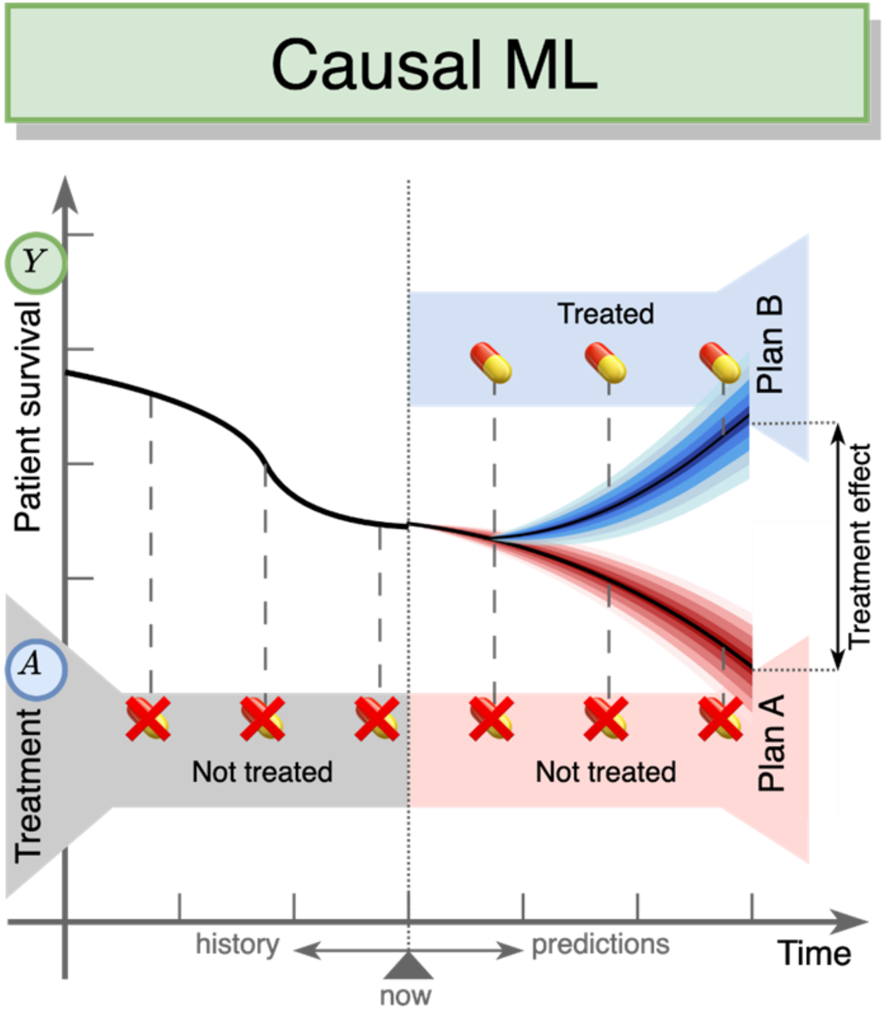


Why do we need Causal ML in medicine?

Reference:
Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I.S. and van der Schaar, M., 2024. **Causal machine learning for predicting treatment outcomes**. Nature Medicine, 30(4), pp.958-968.

Moving from diagnostics to therapeutics: Estimating treatment effects with ML





Estimated Average Treatment Effect of Psychiatric Hospitalization in Patients With Suicidal Behaviors

A Precision Treatment Analysis

Eric L. Ross, MD[1]; Robert M. Bossarte, PhD[2]; Steven K. Dobscha, MD[3]; et al

Author Affiliations | Article Information

JAMA Psychiatry. 2024;81(2):135-143. doi:10.1001/jamapsychiatry.2023.3994

Key Points

Question Can development of an individualized treatment rule identify patients presenting to emergency departments/urgent care with suicidal ideation or suicide attempts who are likely to benefit from psychiatric hospitalization?

Findings A decision analytic model found that hospitalization was associated with reduced suicide attempt risk among patients who attempted suicide in the past day but not among others with suicidality. Accounting for heterogeneity, suicide attempt risk was found to increase with hospitalization in 24% of patients and decrease in 28%.

Meaning Results of this study suggest that implementing an individualized treatment rule could identify many additional patients who may benefit from or be harmed by hospitalization.

TERMINOLOGY

Real-world data (RWD) vs. real-world evidence (RWE) to support medicine

The US Food and Drug Administration (FDA) defines [1,2,3]:

Real-world data (RWD)

- Data relating to patient health status and the delivery of healthcare
- **Examples:** electronic health records (EHRs), claims and billing activities, disease registries, …
- Naming: observational data (≠ experimental data)

Real-world evidence (RWE)

- Analysis of RWD regarding usage and effectiveness
- Vision: greater personalization of care
- Disclaimer: should not replace but augment RCTs

1) Real-World Evidence — Where Are We Now? https://www.nejm.org/doi/full/10.1056/NEJMp2200089
2) Real-World Evidence — What Is It and What Can It Tell Us? https://www.nejm.org/doi/full/10.1056/nejmsb1609216
3) Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness https://jamanetwork.com/journals/jama/fullarticle/2697359

TERMINOLOGY

Real-world data (RWD) vs. real-world evidence (RWE) to support medicine

The US Food and Drug Administration (FDA) defines [1,2,3]:

Real-world data (RWD)

- Data relating to patient health status and the delivery of healthcare
- **Examples:** electronic health records (EHRs), claims and billing activities, disease registries, …
- Naming: observational data (≠ experimental data)

- **Aim:** estimate treatment effectiveness
- **Challenges:** representativeness (selection bias), no proper randomization, …
- **Custom methodologies:** target trial emulation, **causal machine learning**, …

Real-world evidence (RWE)

- Analysis of RWD regarding usage and effectiveness
- Vision: greater personalization of care
- Disclaimer: should not replace but augment RCTs

1) Real-World Evidence — Where Are We Now? https://www.nejm.org/doi/full/10.1056/NEJMp2200089
2) Real-World Evidence — What Is It and What Can It Tell Us? https://www.nejm.org/doi/full/10.1056/nejmsb1609216
3) Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness https://jamanetwork.com/journals/jama/fullarticle/2697359

Application scenarios of RWD

RWD helps to guide decision-making (beyond RCTs):

1 … in the absence of a standard of care

- Specific subtypes of diseases with no standard of care yet (e.g., oncology)
- New or experimental drugs (e.g., orphan drugs, is Biontech vs. Moderna vaccine more effective for subcohort X?)

2 … in complex, high-dimensional decision problems

- Complex dosaging problems (e.g., chemotherapy, combi-treatments)

3 … when RCTs are unethical

- Vulnerable populations (pregnant women, children, severely ill, etc.) [1]

4 … when a greater personalization is desired

- Highly granular subpopulations that cannot be really placed in RCTs (e.g., women, above 60, with comorbidity X, Y & Z or generally specific patient trajectories)
  → maybe a subpopulations responds different for a specific drug, or a second line of treatment is more effective than the first line?
- Personalization based on genome data (e.g., precision medicine)

1) The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients https://jamanetwork.com/journals/jama/article-abstract/407990

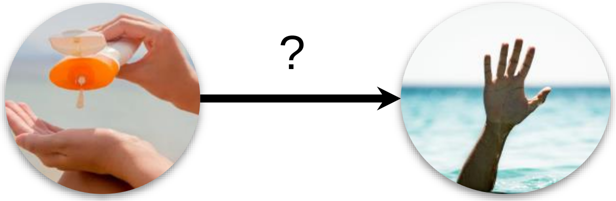
EXAMPLE

Real-world data (RWD) vs. real-world evidence (RWE) to support medicine

Why is getting a **meaningful** RWE challenging?

Real-world data (RWD)

- Observational data of
    - sunscreen usage (binary treatment)
    - number of drowning-related deaths (outcome)



- **Aim:** effect of sunscreen on the chance of drowning

Real-world evidence (RWE)

- Evidence: The higher the usage of sunscreen -> the more likely is the chance of drowning
- This is counterintuitive: Is there something we didn't account for?

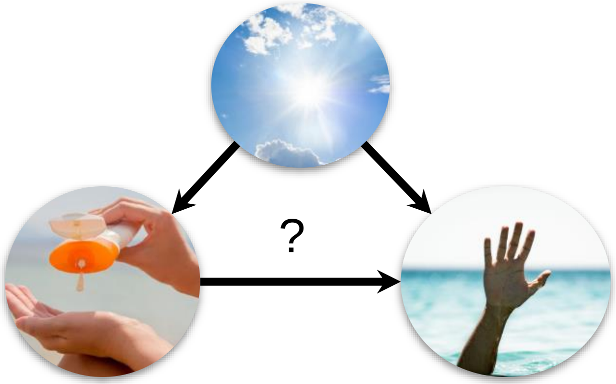
EXAMPLE

Real-world data (RWD) vs. real-world evidence (RWE) to support medicine

Why is getting a **meaningful** RWE challenging?

Real-world data (RWD)

- Observational data of
  - sunscreen usage (binary treatment)
  - number of drowning-related deaths (outcome)
  - **intensity of sunlight (covariates)**



- **Aim:** effect of sunscreen on the chance of drowning for **different intensities of sunlight**

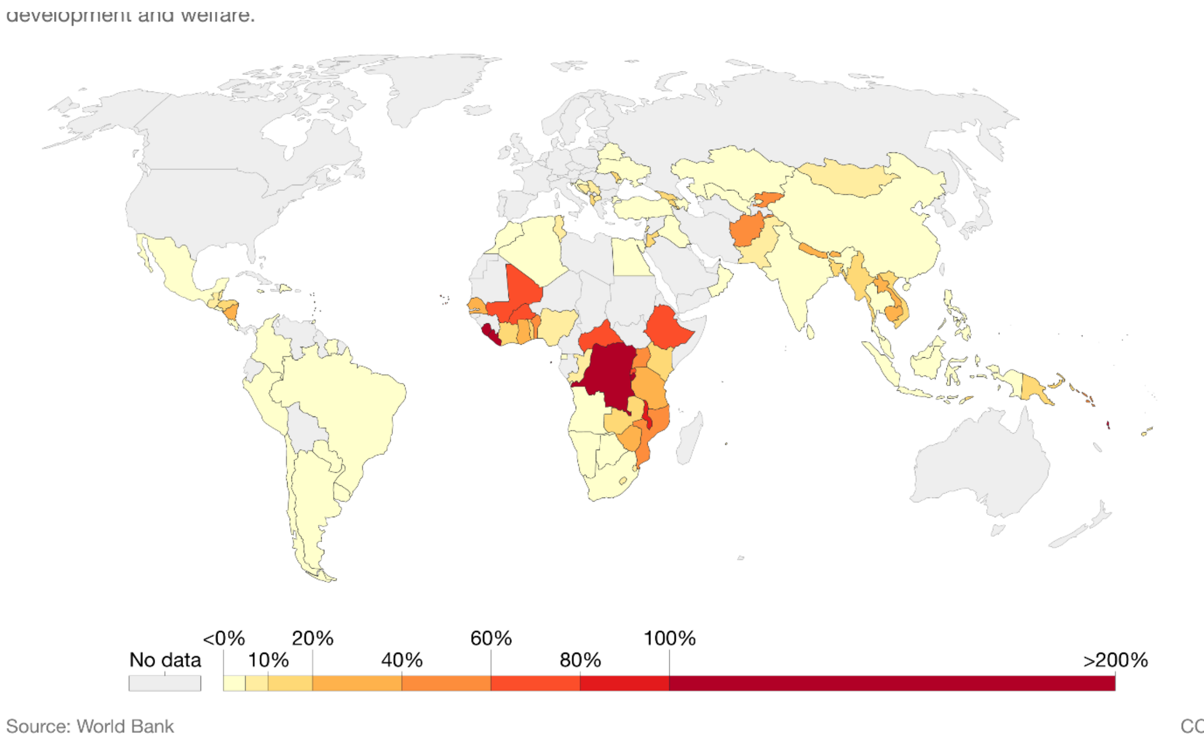
Real-world evidence (RWE)

- Evidence: no association between sunscreen usage and chance of drowning in each group of sunlight
- Comparing with the previous slide: Intensity of sunlight is a **confounder**

A tailored prediction method is needed to address selection bias

Predict probability of winning: 0.76

Rich countries also get more aid 0.35



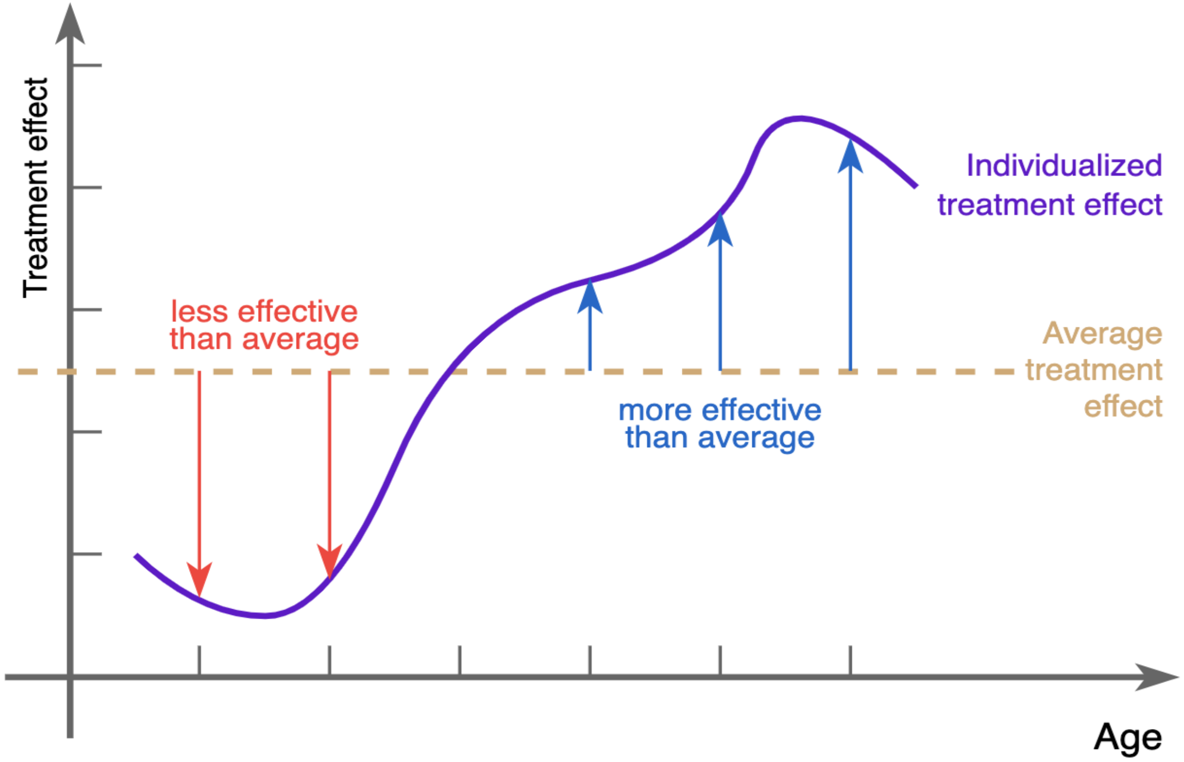
AIM

Understanding heterogeneity in the treatment effect

- Focus is often on **average** treatment effect (ATE)

- ATE is aggregated across the population
- ATE **cannot** tell whether a treatment works for some or not
  → e.g., medication works only for women but not for men, but RCT was done with all patients

- NB: both RCTs and target trial emulation focus on ATEs



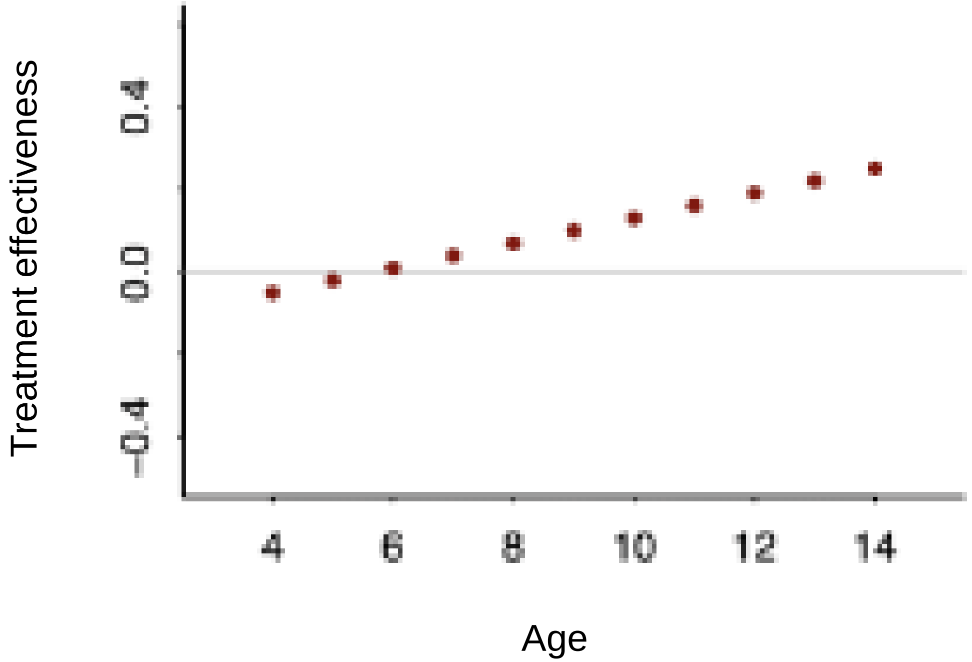
To personalize treatment recommendations, we need to understand the **individualized** treatment effect (ITE)

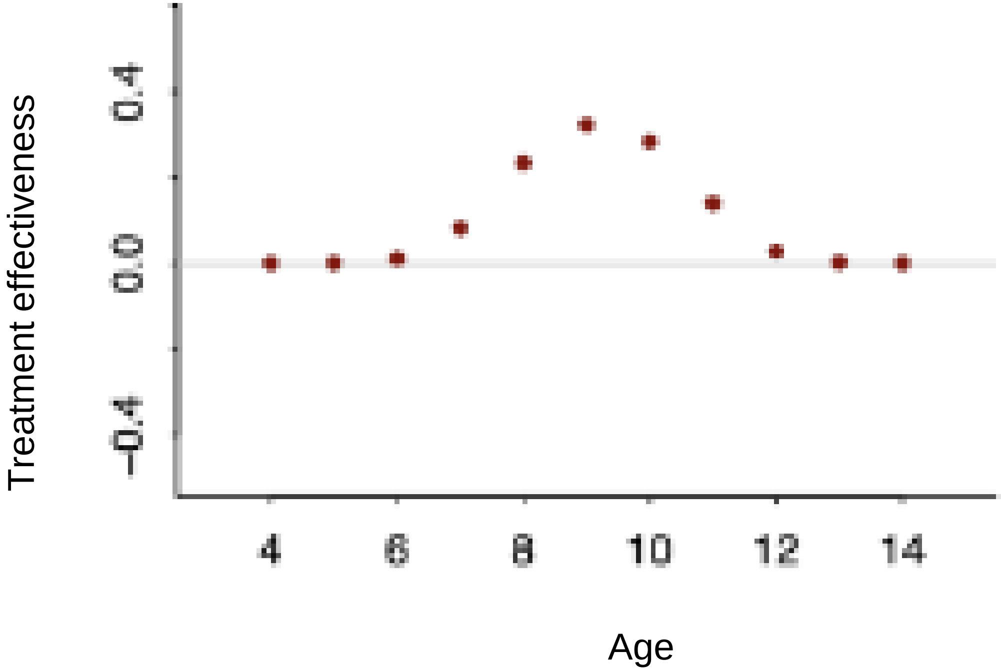
Illustrative

AIM

Why we need to go beyond the ATE and understand heterogeneity in the treatment effect

- All plots show the same average treatment effect (ATE)
- BUT: the medication is only effective some subpopulations





Short introduction
to causal machine learning

Reference:
Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I.S. and van der Schaar, M., 2024. Causal machine learning for predicting treatment outcomes. Nature Medicine, 30(4), pp.958-968.

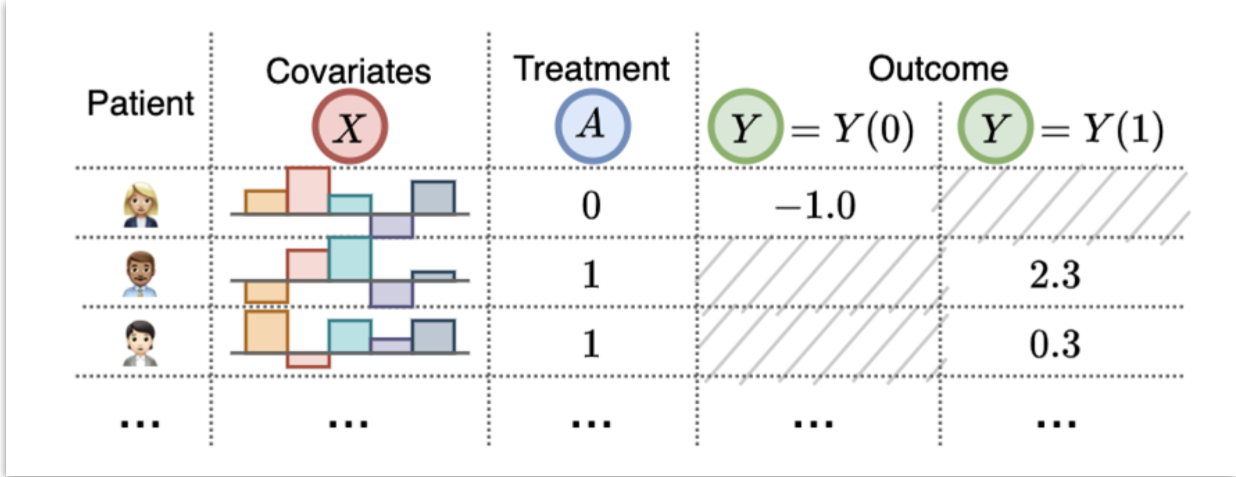
Estimating the potential outcomes of treatments

Problem formulation

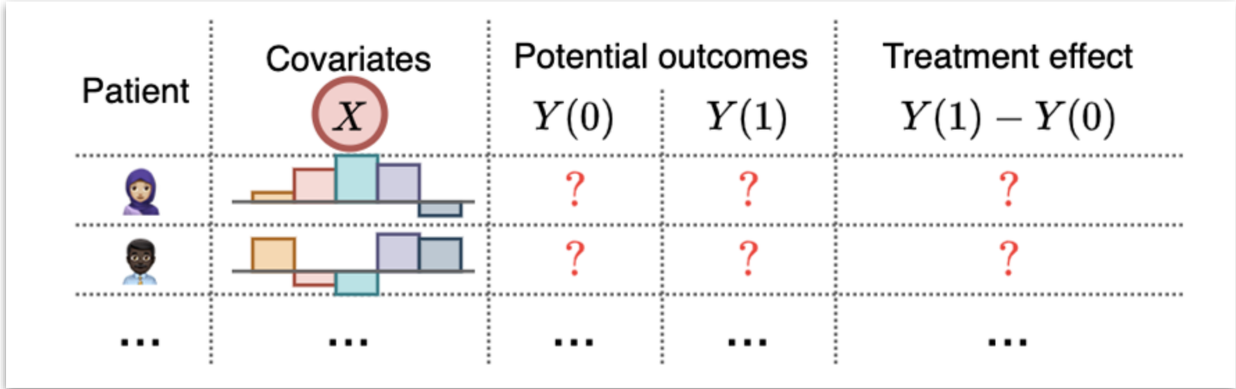
- Given i.i.d. observational dataset

$$\mathcal{D} = \{x_i, a_i, y_i\}_{i=1}^{n} \sim \mathbb{P}(X, A, Y)$$

  - X covariates
  - A (binary) treatments
  - Y continuous (factual) outcomes



- We want to identify & estimate treatment outcomes:
  - **treatment effects** $Y[1] - Y[0]$
  - **potential outcomes** (separately) $Y[0]$ $Y[1]$



- **Fundamental problem**: never observing both potential outcomes!

Ladder of causation

Pearl's layers of causation

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Causal Hierarchy Theorem: statistical inference for a layer requires the information from the same or higher layer. For the inference from lower layer data, we need to make **additional assumptions**.

[1] Elias Bareinboim et al. "On Pearl's hierarchy and the foundations of causal inference". In: Probabilistic and Causal Inference: The Works of Judea Pearl. Association for Computing Machinery, 2022, pp. 507–556.

PRIMER

Ladder of causation

Pearl's layers of causation

Traditional ML

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Causal Hierarchy Theorem: statistical inference for a layer requires the information from the same or higher layer. For the inference from lower layer data, we need to make **additional assumptions**.

[1] Elias Bareinboim et al. "On Pearl's hierarchy and the foundations of causal inference". In: Probabilistic and Causal Inference: The Works of Judea Pearl. Association for Computing Machinery, 2022, pp. 507–556.

PRIMER

Ladder of causation

Pearl's layers of causation

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a s about the electi
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

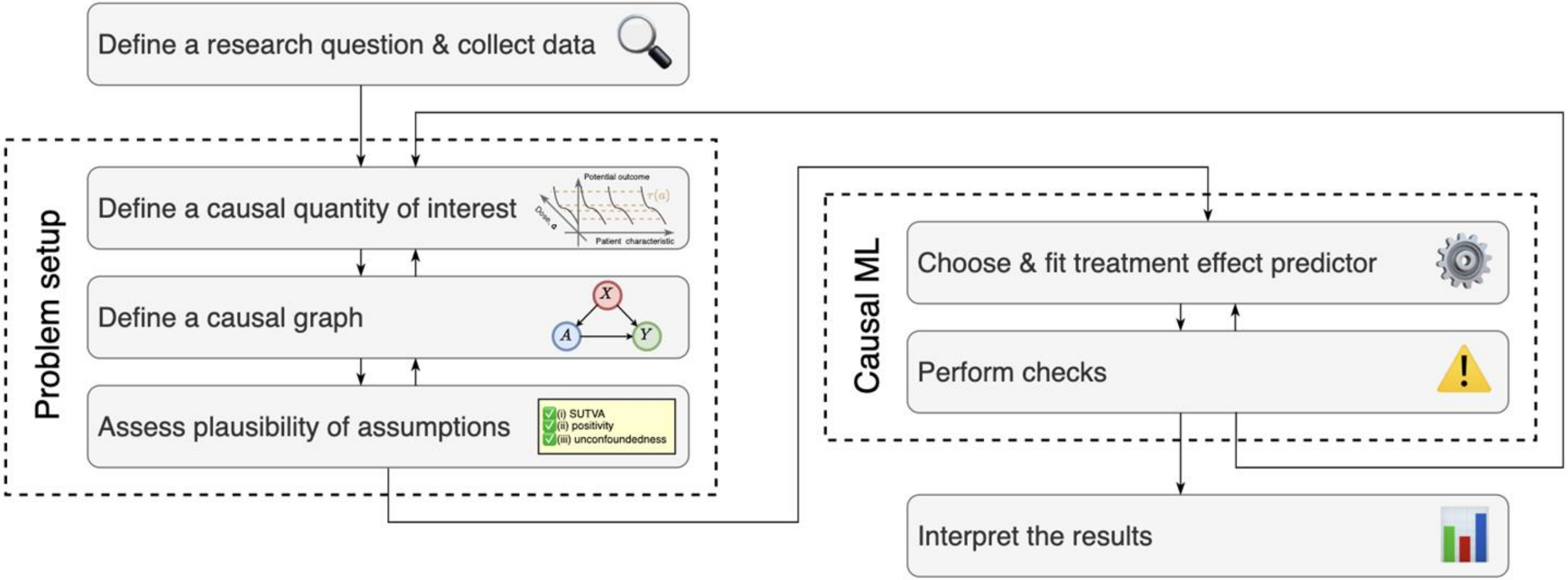
Causal ML

Causal Hierarchy Theorem: statistical inference for a layer requires the information from the same or higher layer. For the inference from lower layer data, we need to make **additional assumptions**.

[1] Elias Bareinboim et al. "On Pearl's hierarchy and the foundations of causal inference". In: Probabilistic and Causal Inference: The Works of Judea Pearl. Association for Computing Machinery, 2022, pp. 507–556.

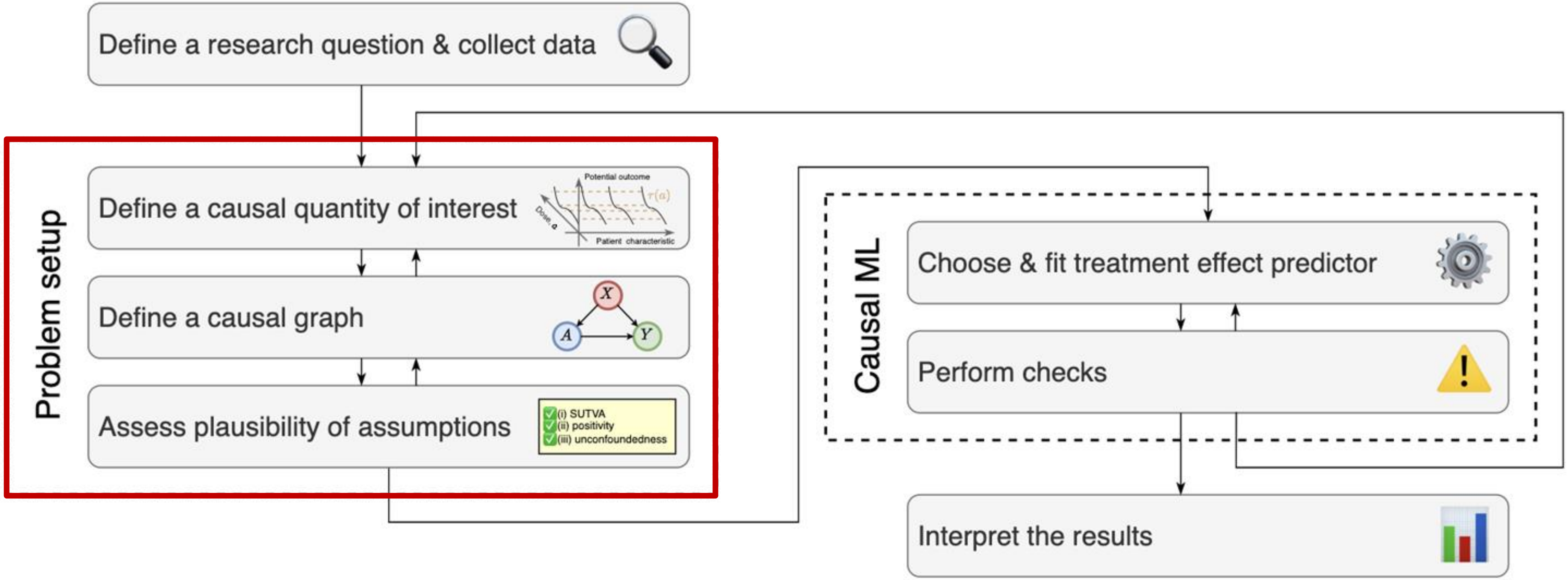
PRIMER

Causal ML Workflow



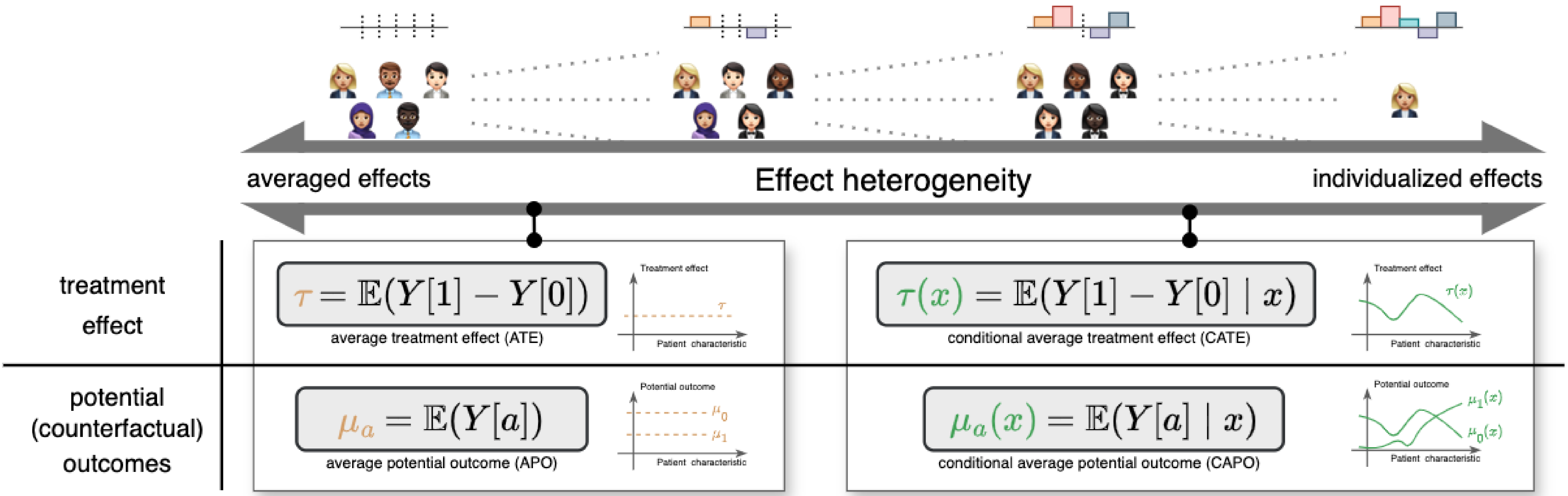
PRIMER

Causal ML Workflow



PROBLEM SETUP

Causal quantities of interest

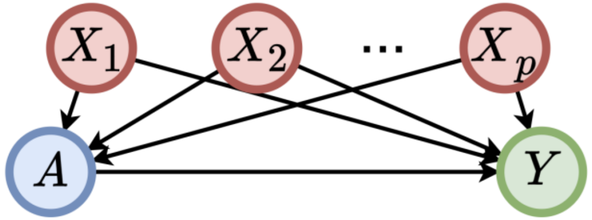


Example of a case study

Aim: estimate heterogeneous treatment effect of development aid on SDG outcomes

- Treatment *A*: development aid earmarked to end the HIV/AIDS epidemic
- Outcome *Y*: relative reduction in HIV infection rate
- Covariates *X*: control for differences in country characteristics

Causal graph



Causal quantity of interest

$$\boxed{\mu_a(x) = \mathbb{E}(Y[a] \mid x)}$$

conditional average potential outcome (CAPO)

Assumptions

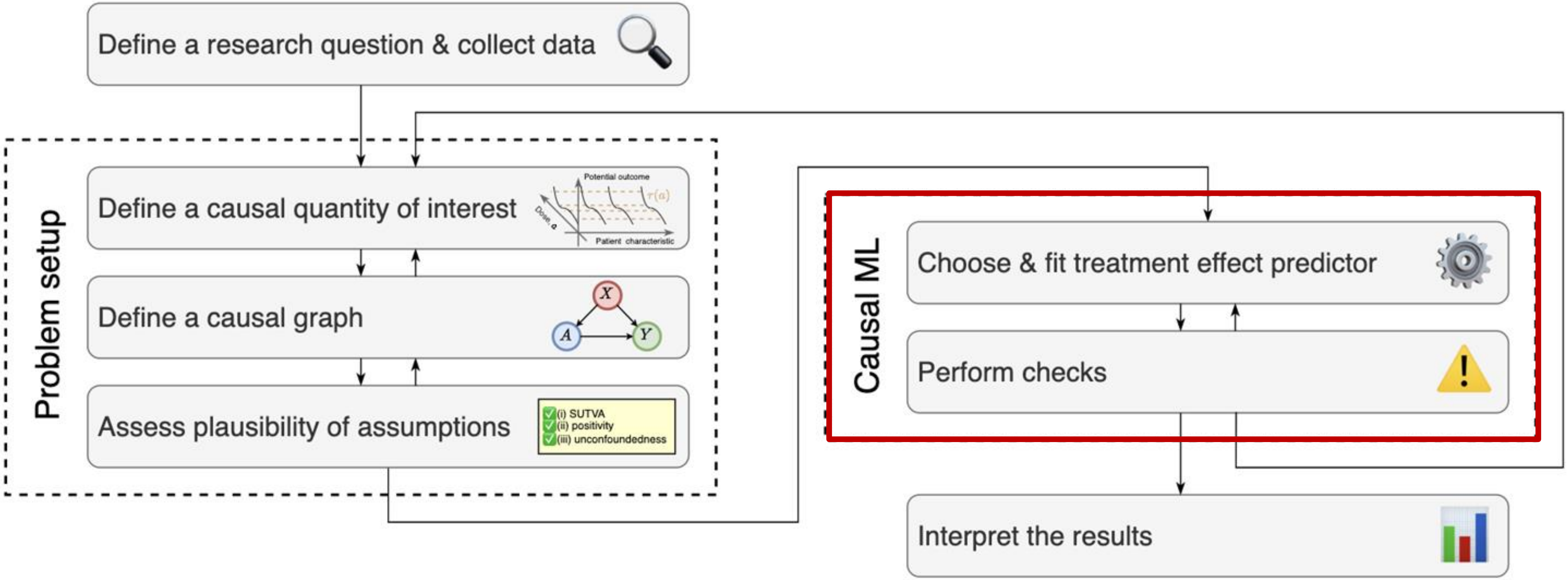
Consistency: $Y = Y(a)$ if $A = a$

Positivity: $0 < p(A = a \mid X = x) < 1, \forall a \in \mathcal{A}$

Ignorability: $Y(a) \perp\!\!\!\perp A \mid X = x \;, \forall a \in \mathcal{A}$

PRIMER

Causal ML Workflow



Challenges and open questions fitting an ML model

Challenges

$$\mu_a(x) = \mathbb{E}(Y[a] \mid x)$$

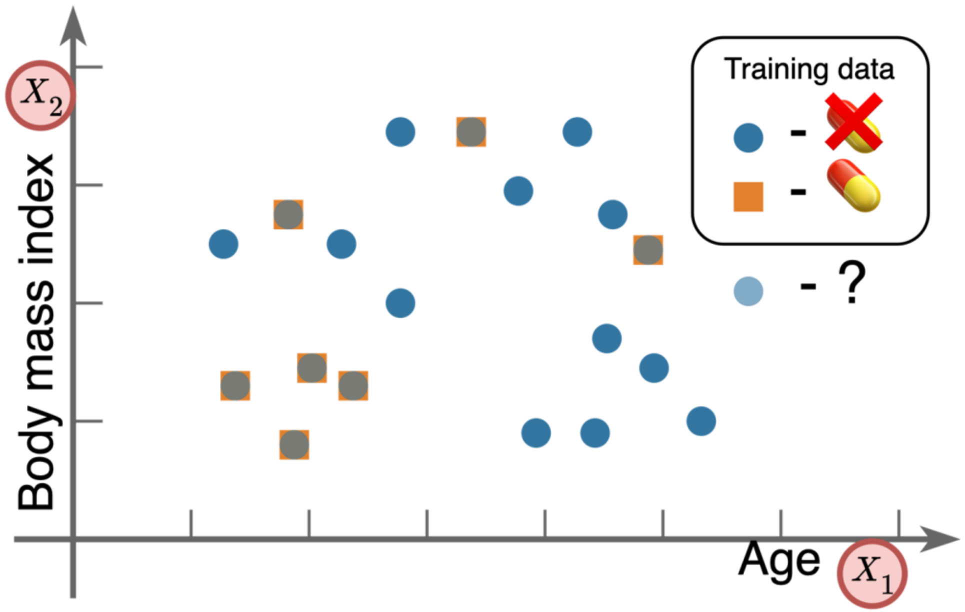
conditional average potential outcome (CAPO)

- **Selection bias**: some subpopulations are rarely treated

$$\tau(x) = \mathbb{E}(Y[1] - Y[0] \mid x)$$

conditional average treatment effect (CATE)

- **Selection bias**: some subpopulations are rarely treated



Open problems

Challenges and open questions fitting an ML model

Challenges

$$\mu_a(x) = \mathbb{E}(Y[a] \mid x)$$

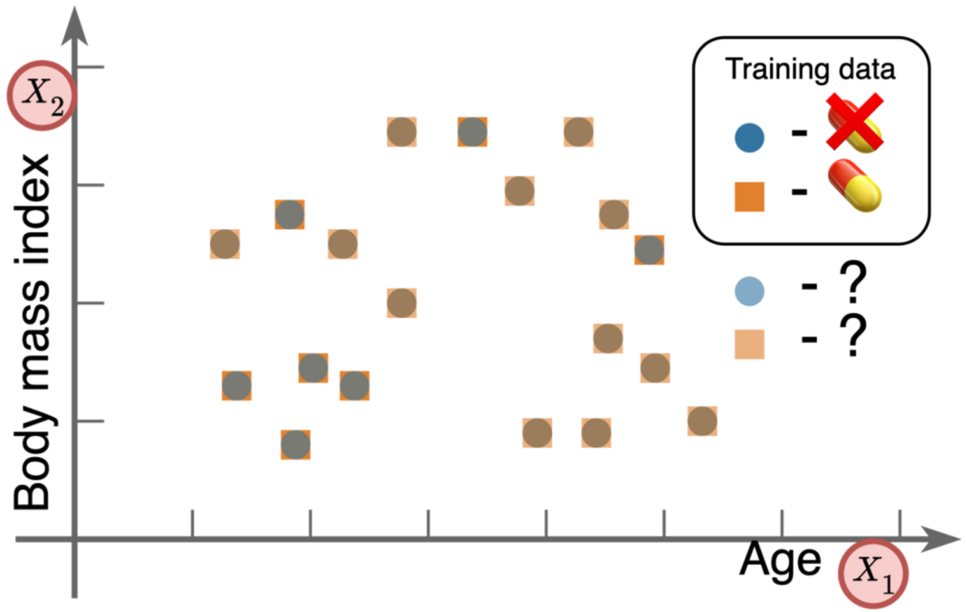
conditional average potential outcome (CAPO)

- **Selection bias**: some subpopulations are rarely treated

$$\tau(x) = \mathbb{E}(Y[1] - Y[0] \mid x)$$

conditional average treatment effect (CATE)

- **Selection bias**: some subpopulations are rarely treated
- **Fundamental problem**: never observing a difference of potential outcomes



Open problems

Challenges and open questions fitting an ML model

Challenges

$$\mu_a(x) = \mathbb{E}(Y[a] \mid x)$$

conditional average potential outcome (CAPO)

- **Selection bias**: some subpopulations are rarely treated

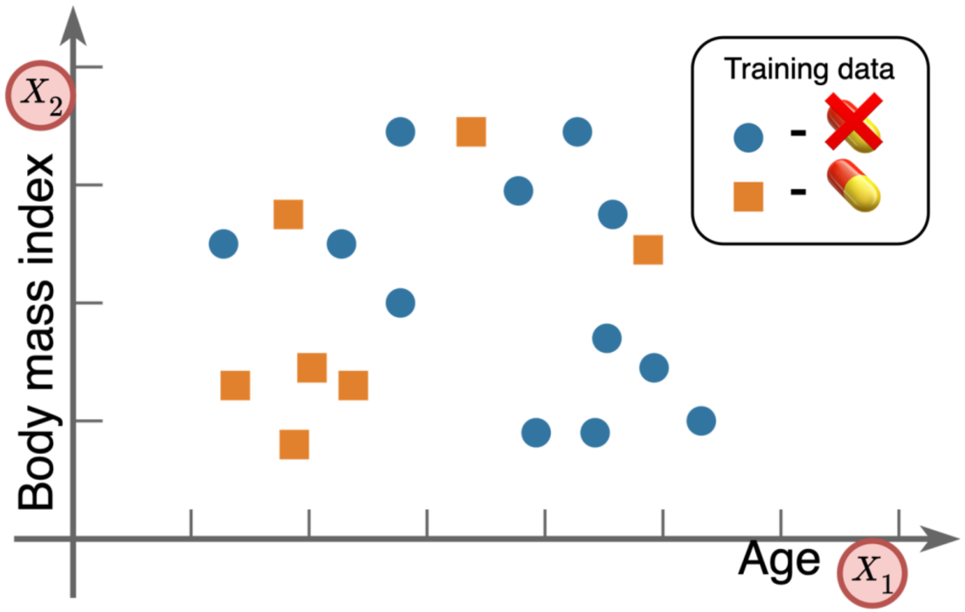
$$\tau(x) = \mathbb{E}(Y[1] - Y[0] \mid x)$$

conditional average treatment effect (CATE)

- **Selection bias**: some subpopulations are rarely treated
- **Fundamental problem**: never observing a difference of potential outcomes

Open problems

- How to effectively address selection bias?
- How to incorporate inductive biases, e.g., regularize CAPO / CATE models?



Methods

Meta-learners

- Meta-learners (Kunzel 2019) are model-agnostic methods for CATE estimation
- Can be used for treatment effect estimation in combination with an arbitrary ML model of choice (e.g., a decision tree, a neural network)

Model-based learners

- Model-specific methods make adjustments to existing ML models to address statistical challenges arising in treatment effect estimation
- Prominent **examples** are the causal tree (Athey 2016) and the causal forest (Wager 2018, Athey 2019)
- Others adapt representation learning to leverage neural networks (Shalit 2017, Shi 2019)

1. Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." Proceedings of the national academy of sciences 116.10 (2019): 4156-4165.
2. Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." Proceedings of the National Academy of Sciences 113.27 (2016): 7353-7360.
3. Athey, Susan, and Stefan Wager. "Estimating treatment effects with causal forests: An application." Observational studies 5.2 (2019): 37-51.
4. Shalit, Uri, Fredrik D. Johansson, and David Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." International conference on machine learning. PMLR, 2017.
5. Shi, Claudia, David Blei, and Victor Veitch. "Adapting neural networks for the estimation of treatment effects." Advances in neural information processing systems 32 (2019).

Methods

Meta-learners

One-stage learners

- "Plug-in learners": fit a **s**ingle regression model with a treatment as an input or **t**wo regression models for each treated and control sub-groups
- Examples: S-learner and T-learner

Two-stage learners

- Two-stages of learning: derive and estimate pseudo-outcomes as surrogates, which has the same expected value as the CATE
- Examples: DR-learner and R-learner

Model-based learners

- Model-specific methods make adjustments to existing ML models to address statistical challenges arising in treatment effect estimation
- Prominent **examples** are the causal tree (Athey 2016) and the causal forest (Wager 2018, Athey 2019)
- Others adapt representation learning to leverage neural networks (Shalit 2017, Shi 2019)

1. Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." Proceedings of the national academy of sciences 116.10 (2019): 4156-4165.
2. Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." Proceedings of the National Academy of Sciences 113.27 (2016): 7353-7360.
3. Athey, Susan, and Stefan Wager. "Estimating treatment effects with causal forests: An application." Observational studies 5.2 (2019): 37-51.
4. Shalit, Uri, Fredrik D. Johansson, and David Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." International conference on machine learning. PMLR, 2017.
5. Shi, Claudia, David Blei, and Victor Veitch. "Adapting neural networks for the estimation of treatment effects." Advances in neural information processing systems 32 (2019).

Methods

Meta-learners

One-stage learners

- "Plug-in learners": fit a **s**ingle regression model with a treatment as an input or **t**wo regression models for each treated and control sub-groups
- Examples: S-learner and T-learner

Two-stage learners

- Two-stages of learning: derive and estimate pseudo-outcomes as surrogates, which has the same expected value as the CATE
- Examples: DR-learner and R-learner

Model-based learners

- Model-specific methods make adjustments to existing ML models to address statistical challenges arising in treatment effect estimation
- Prominent **examples** are the causal tree (Athey 2016) and the causal forest (Wager 2018, Athey 2019)
- Others adapt representation learning to leverage neural networks (Shalit 2017, Shi 2019)

1. Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." Proceedings of the national academy of sciences 116.10 (2019): 4156-4165.
2. Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." Proceedings of the National Academy of Sciences 113.27 (2016): 7353-7360.
3. Athey, Susan, and Stefan Wager. "Estimating treatment effects with causal forests: An application." Observational studies 5.2 (2019): 37-51.
4. Shalit, Uri, Fredrik D. Johansson, and David Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." International conference on machine learning. PMLR, 2017.
5. Shi, Claudia, David Blei, and Victor Veitch. "Adapting neural networks for the estimation of treatment effects." Advances in neural information processing systems 32 (2019).

CAUSAL ML

Challenges and open questions fitting an ML model

Challenges

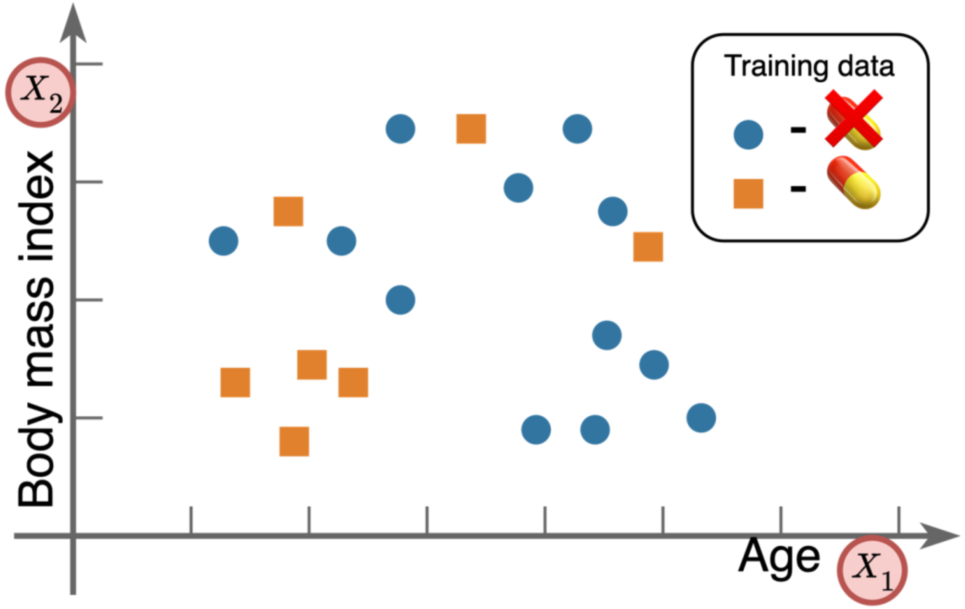
$$\mu_a(x) = \mathbb{E}(Y[a] \mid x)$$

conditional average potential outcome (CAPO)

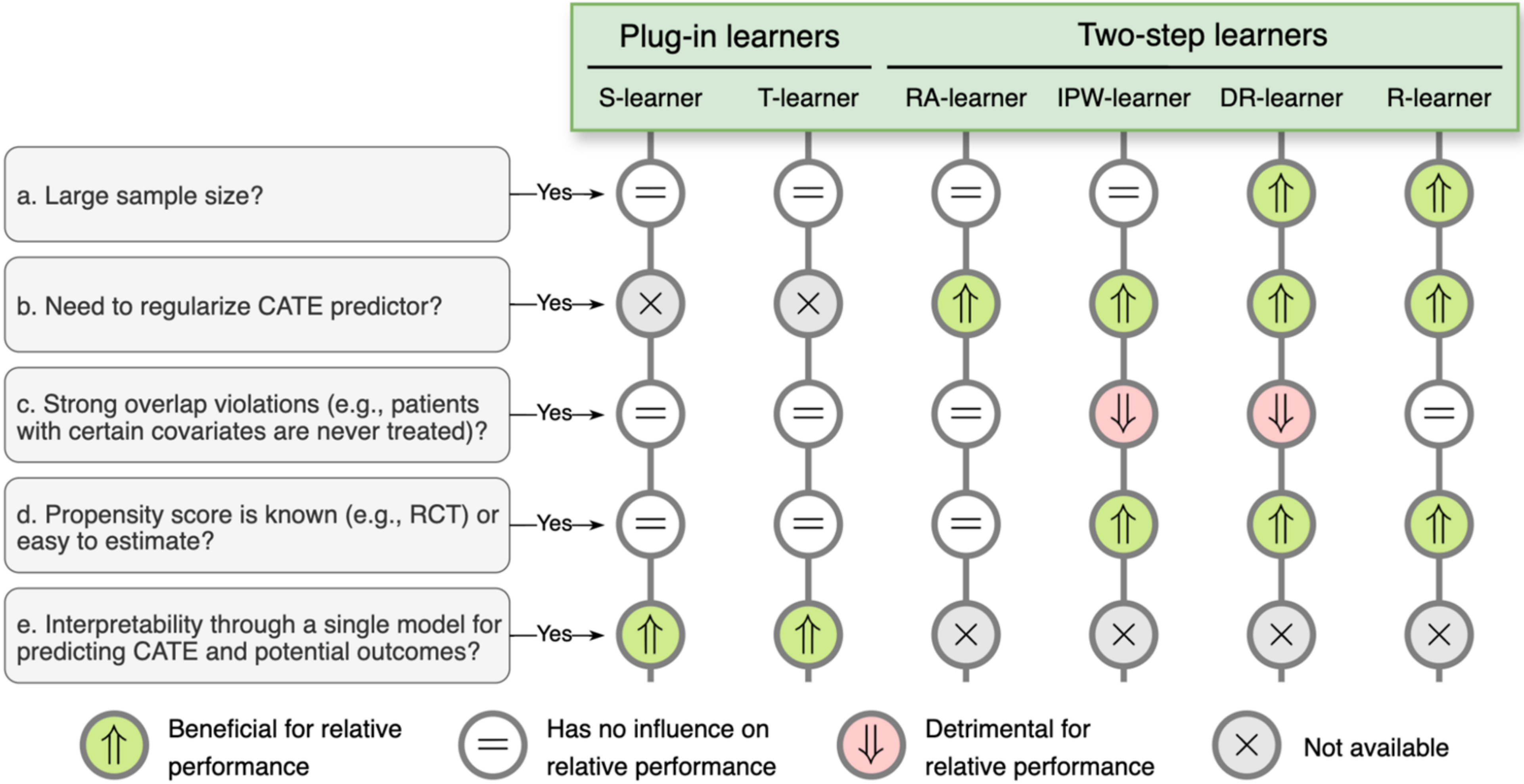
$$\tau(x) = \mathbb{E}(Y[1] - Y[0] \mid x)$$

conditional average treatment effect (CATE)

Open problems



Comparison of meta-learners



LMU

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF AI IN MANAGEMENT

Institute of AI in Management
Prof. Dr. Stefan Feuerriegel

http://www.ai.bwl.lmu.de

@stfeuerriegel stefan-feuerriegel

Artificial intelligence | Impact