Disentangling Treatment Assignment Bias in Counterfactual Outcome Prediction and Biomarker Identification

First Authors: Michael Vollenweider[1], Manuel Schürch[1,2,4] [michavol@ethz.ch, manuel_schurch@dfci.harvard.edu]
Co-Authors: Chiara Rohrer[1], Gabriele Gut[2,3], Michael Krauthammer[2,3] , Andreas Wicki[2,3]
[1]ETHZ: Federal Institute of Technology Zurich; [2]University of Zurich; [3]University Hospital Zurich; [4]Harvard University

Vollenweider, Michael, et al. "Learning Personalized Treatment Decisions in Precision Medicine: Disentangling Treatment Assignment Bias in Counterfactual Outcome Prediction and Biomarker Identification." *Machine Learning for Health (ML4H)*. PMLR, 2025.

Outline

Outline

1. Treatment Effect Prediction & Selection Bias

Outline

1. Treatment Effect Prediction & Selection Bias

2. Information-Theoretic Modelling of Bias

Outline

1. Treatment Effect Prediction & Selection Bias

3. Simulations & Experiments

2. Information-Theoretic Modelling of Bias

Outline

1. Treatment Effect Prediction & Selection Bias

2. Information-Theoretic Modelling of Bias

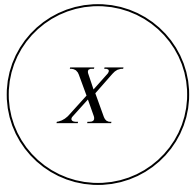
3. Simulations & Experiments

4. Results

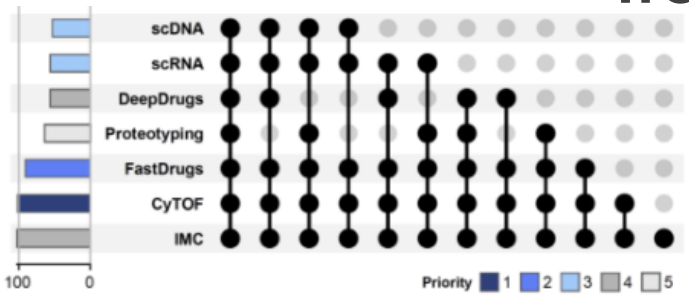
Treatment Effect Model

Treatment Effect Model

Patient Data



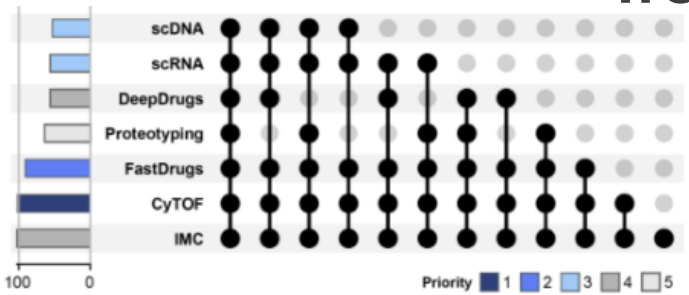
Treatment Effect Model

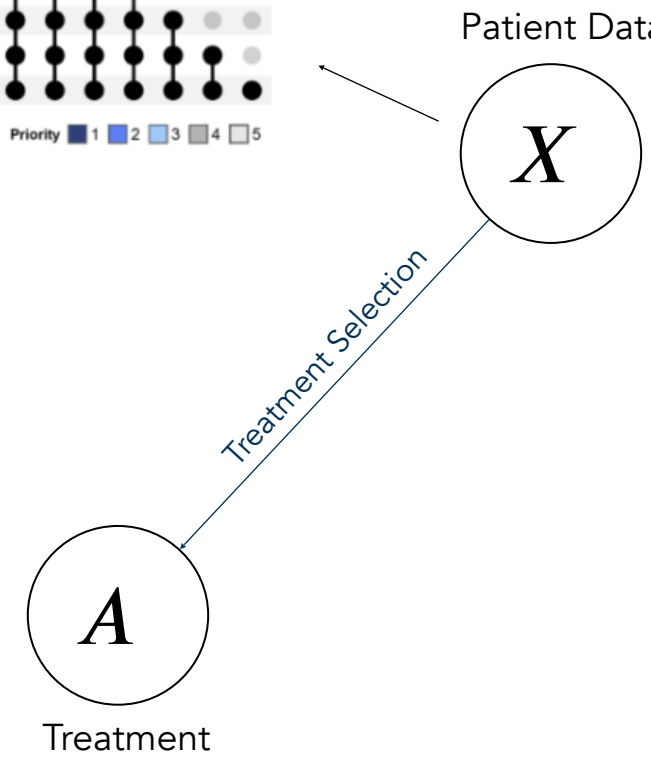


Patient Data

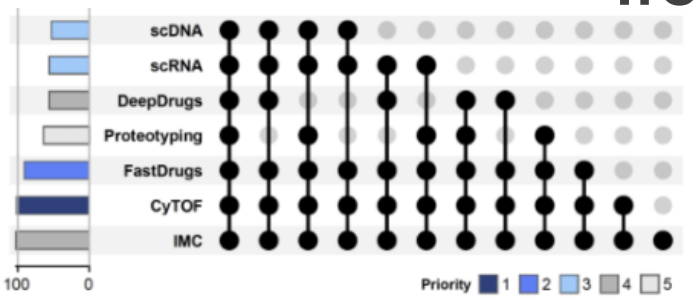
X

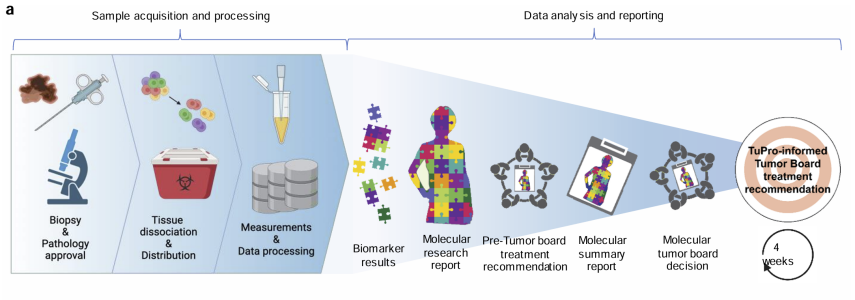
Treatment Effect Model

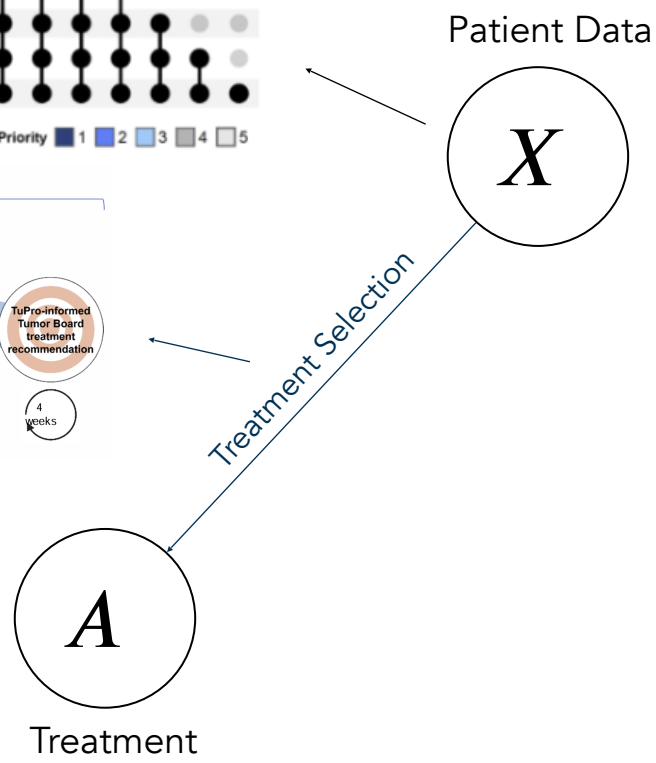




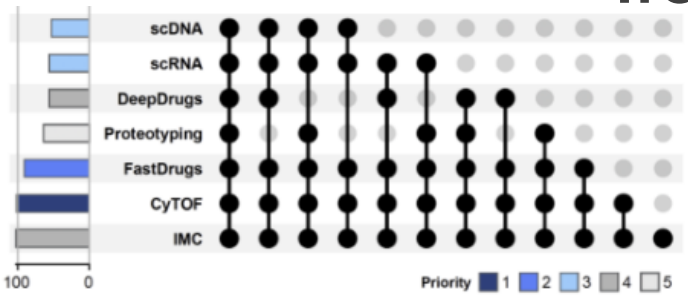
Treatment Effect Model

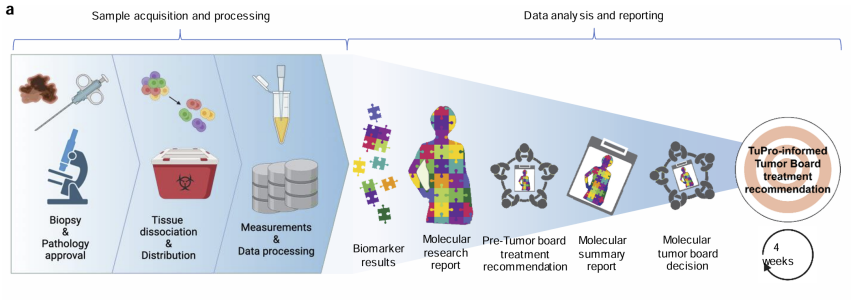


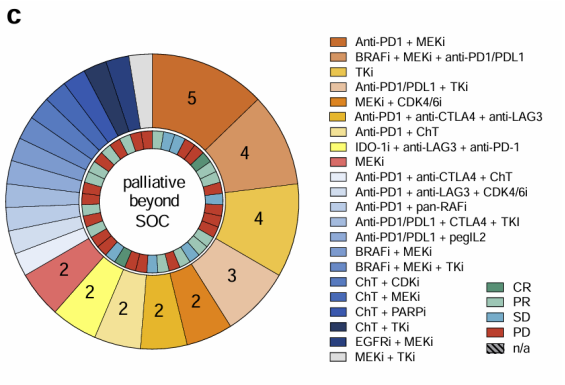




Treatment Effect Model







Patient Data

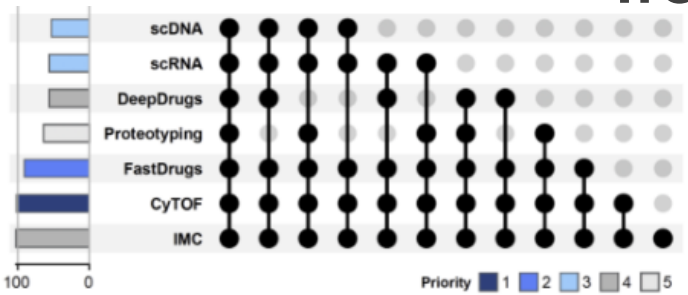
X

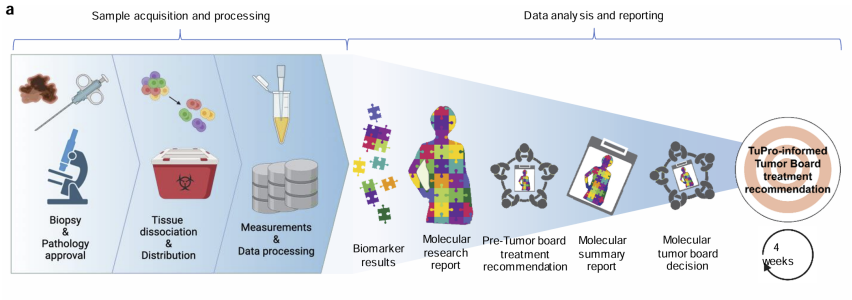
Treatment Selection

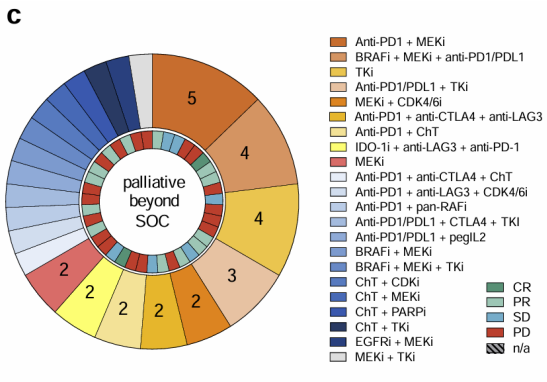
A

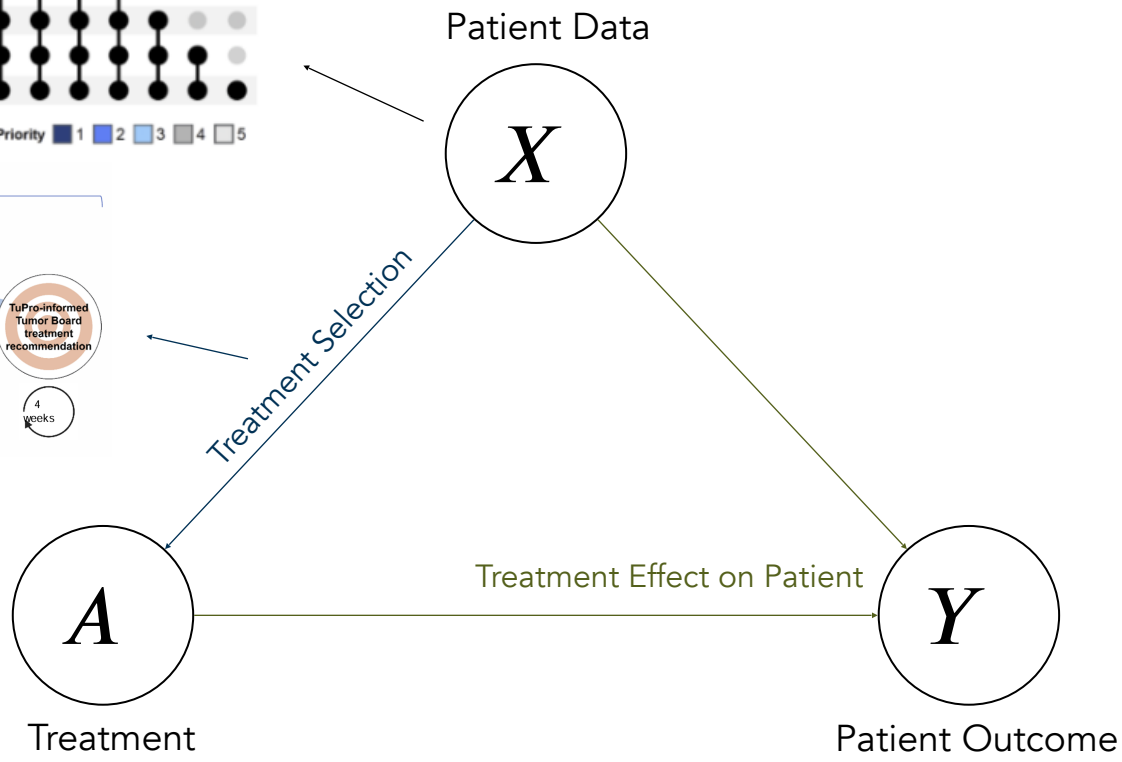
Treatment

Treatment Effect Model

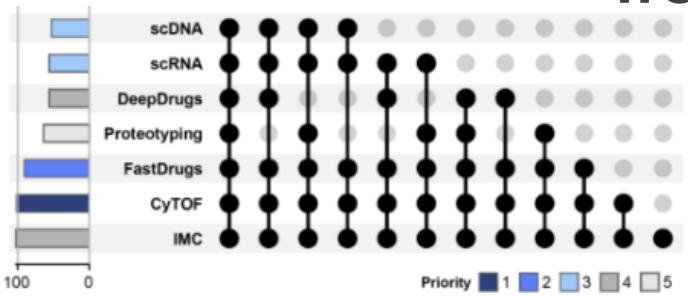


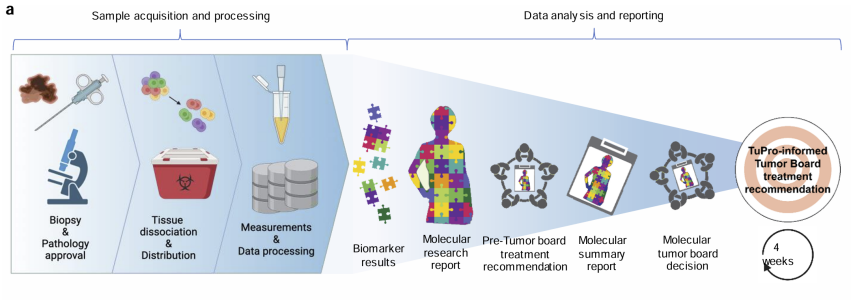


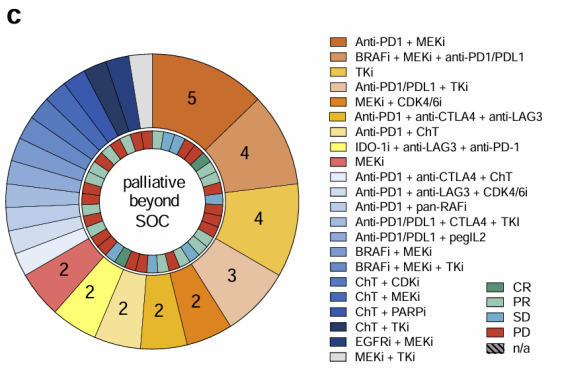


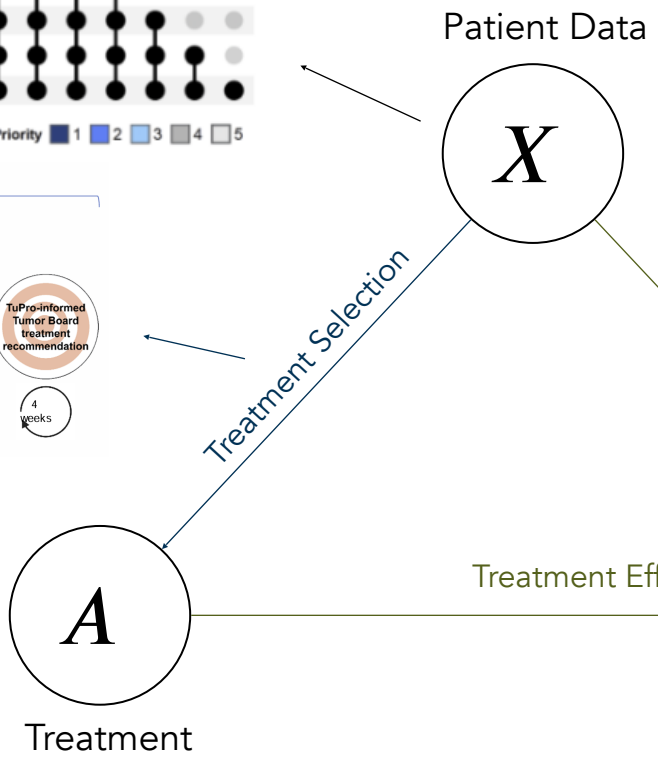


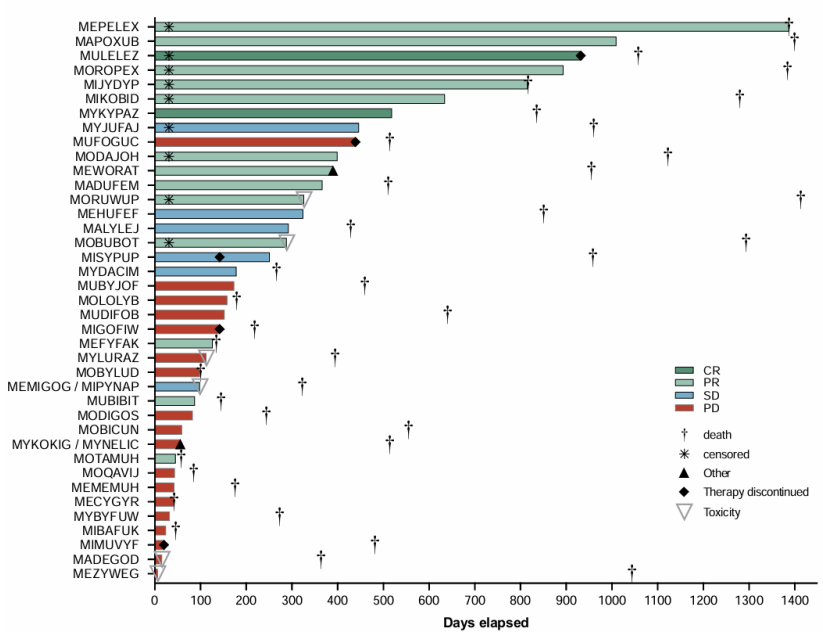
Treatment Effect Model











CATE Prediction

Potential Outcomes Framework by Rubin (2005):

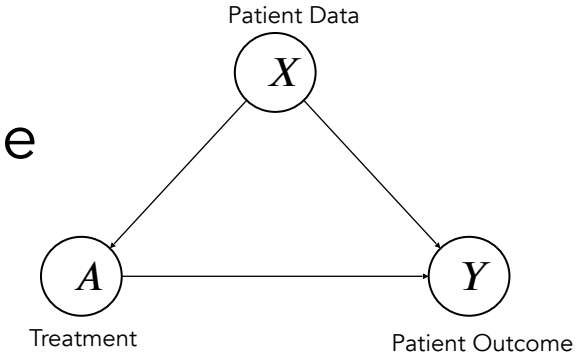
Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331, 2005.

CATE Prediction

Potential Outcomes Framework by Rubin (2005):

Conditional Expected Potential Outcome

$$\mu_a(x) = E[Y_i^a \mid X_i = x]$$



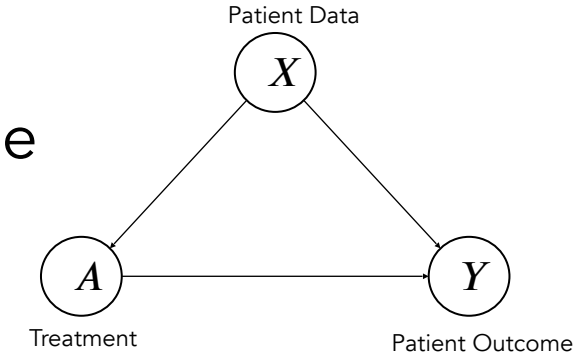
Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331, 2005.

CATE Prediction

Potential Outcomes Framework by Rubin (2005):

Conditional Expected Potential Outcome

$$\mu_a(x) = E[Y_i^a \mid X_i = x]$$

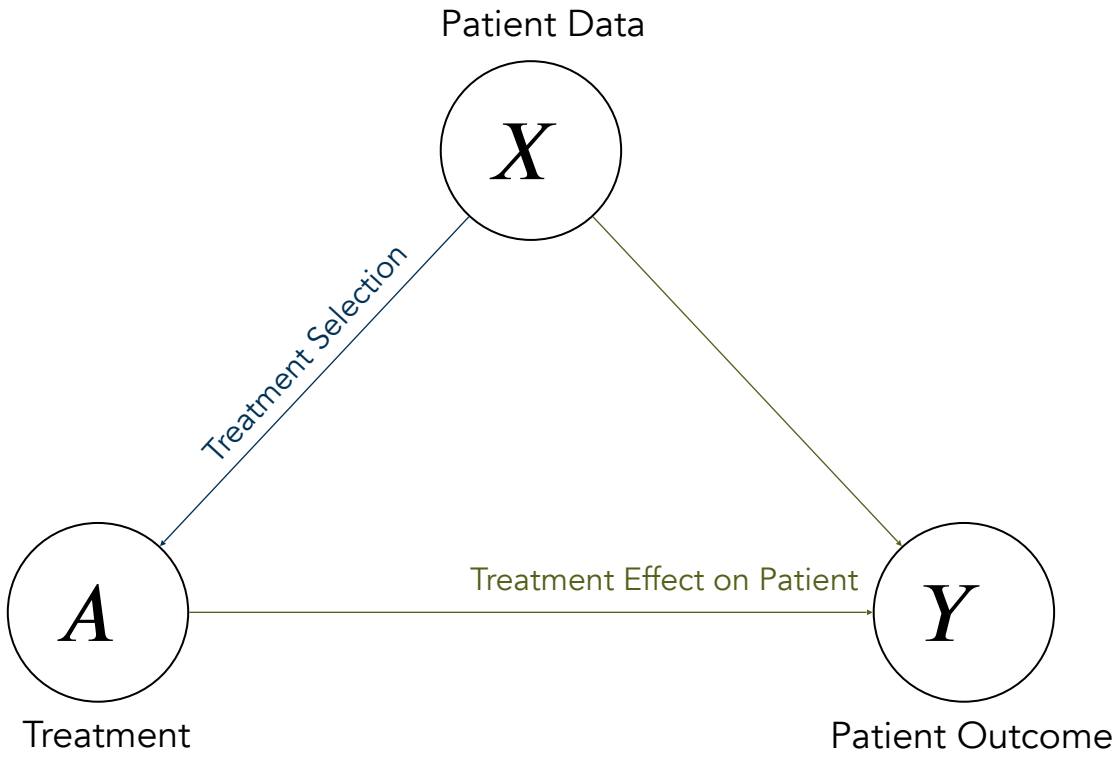


Conditional Average Treatment Effect - CATE

$$\tau(x) = E[Y_i^1 - Y_i^0 \mid X_i = x] = \mu_1(x) - \mu_0(x)$$

Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331, 2005.

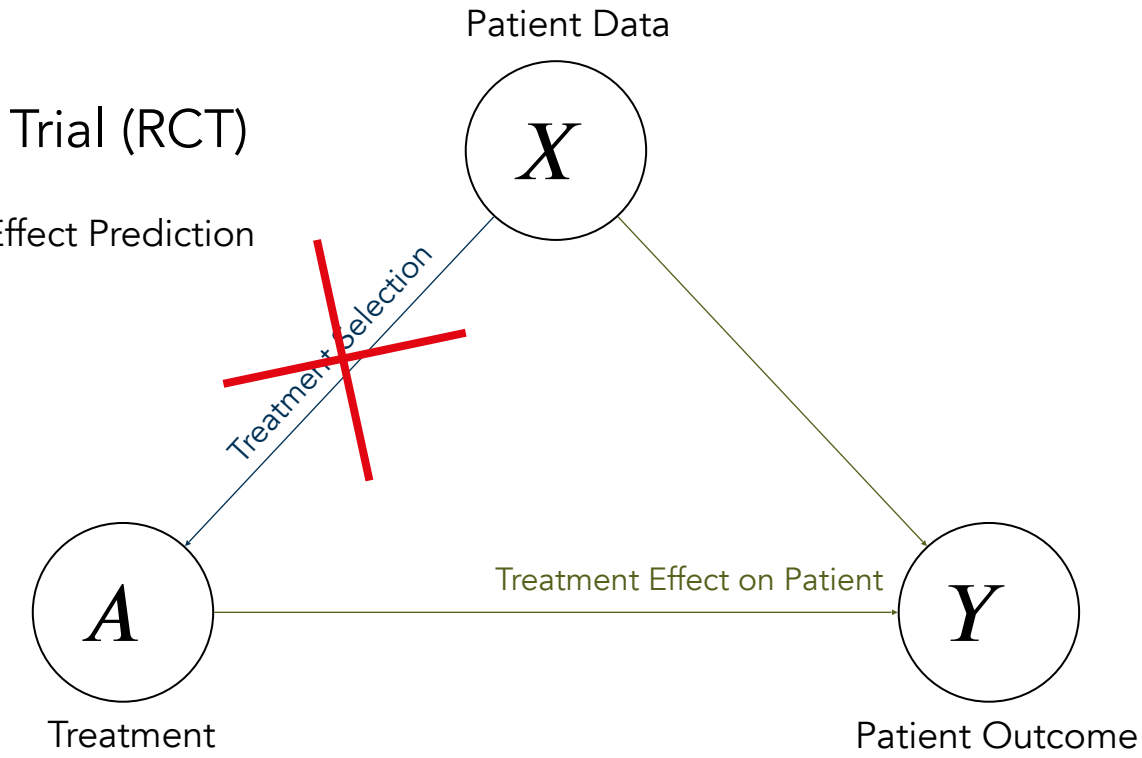
Treatment Effect Model



Treatment Effect Model

Randomized Controlled Trial (RCT)

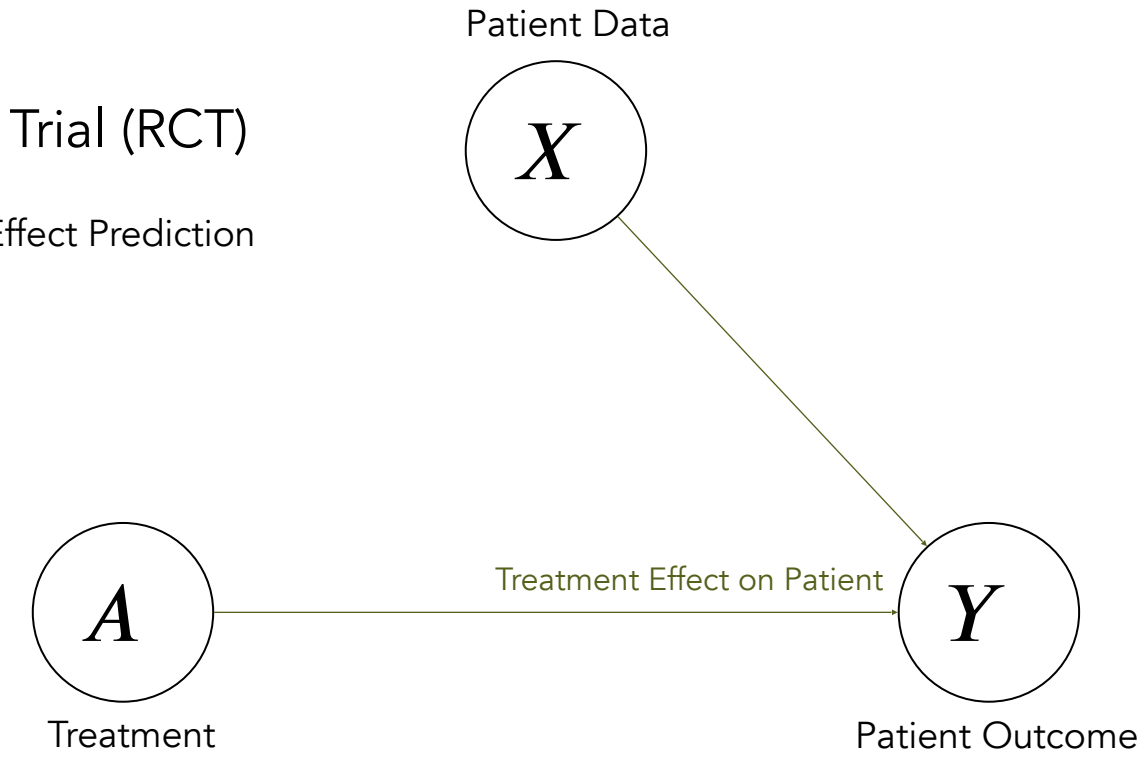
→ Gold Standard for Treatment Effect Prediction



Treatment Effect Model

Randomized Controlled Trial (RCT)

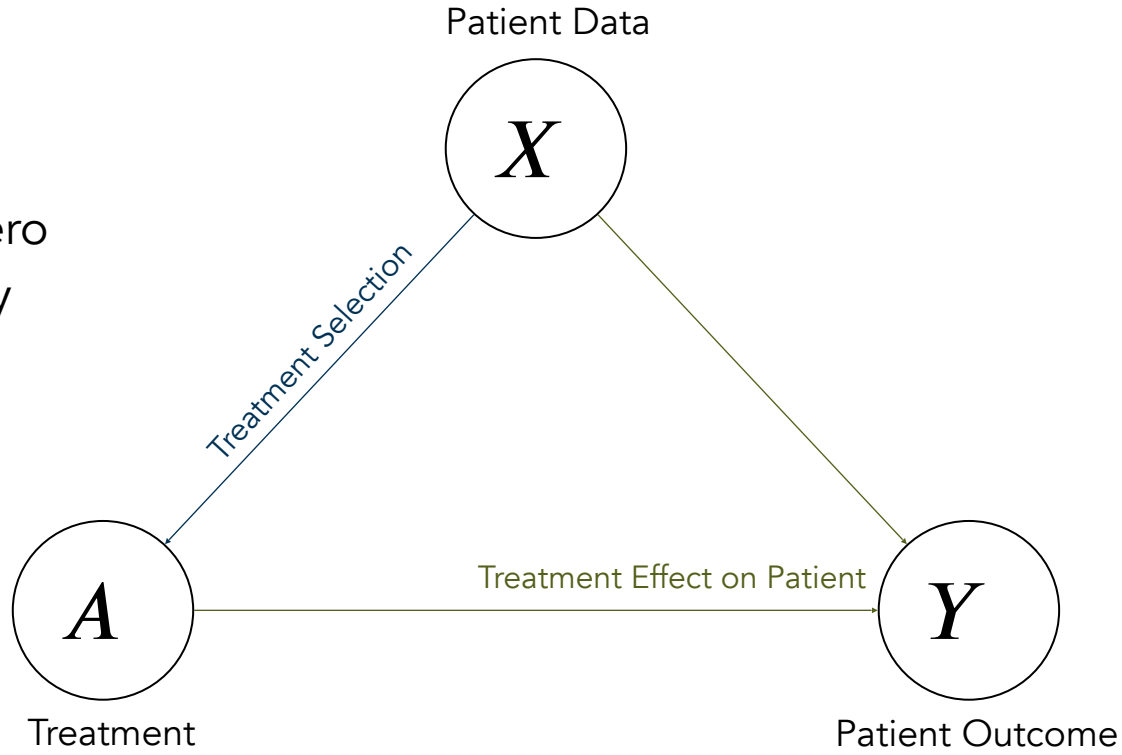
→ Gold Standard for Treatment Effect Prediction



Treatment Effect Model

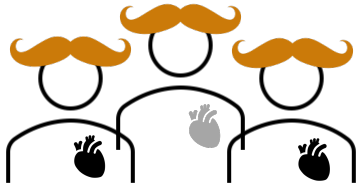
Overlap Assumption:

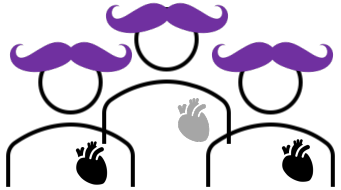
"Each patient has a non-zero probability of receiving any treatment"
(Rosenbaum and Rubin, 1983)



Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Types of Bias





Types of Bias

Treatment selection
based on hair color.

Treatment 1

Treatment 2

Types of Bias

Treatment selection
based on hair color.

Treatment 1

Treatment 2

Overlap Assumption not Fulfilled!

Types of Bias

Treatment selection
based on hair color.

Treatment 1

Treatment 2

Overlap Assumption not Fulfilled!

But does it matter?

Types of Bias

Treatment selection
based on hair color.

Treatment 1

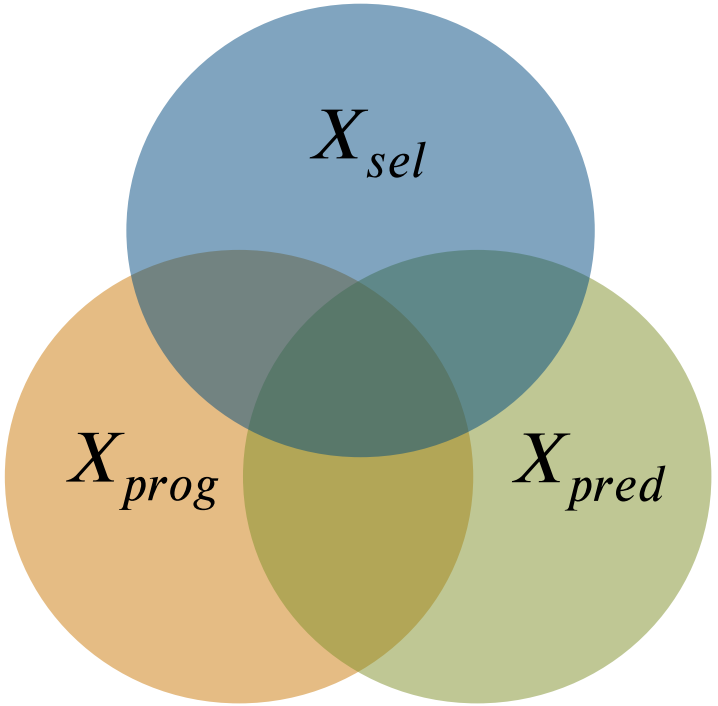
Treatment 2

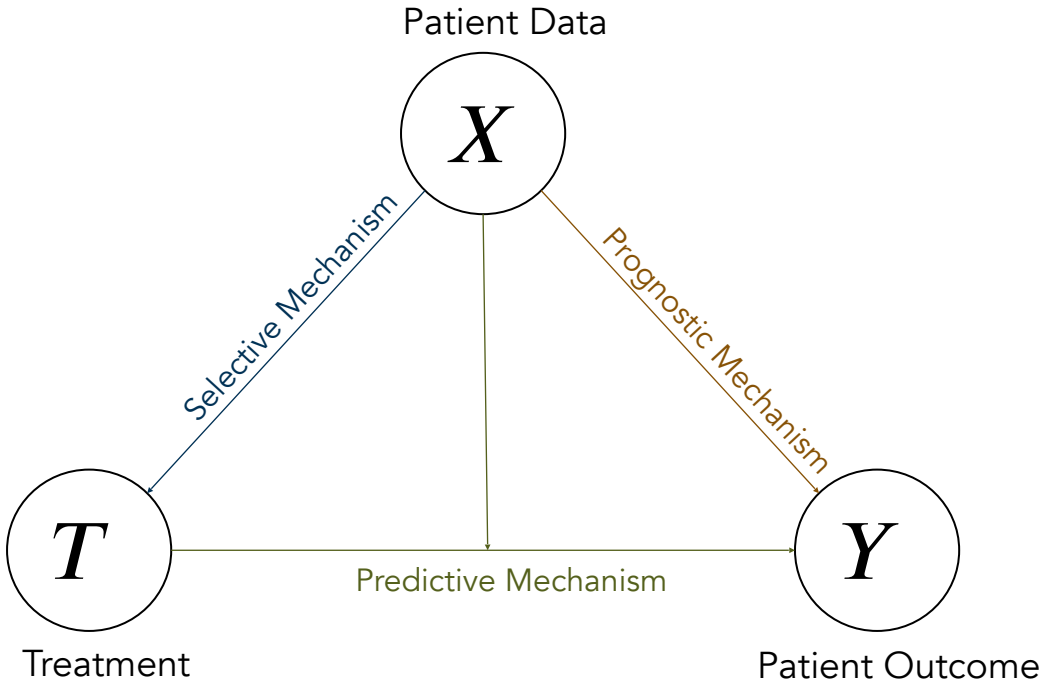
Overlap Assumption not Fulfilled!

But does it matter?

With regard to ‘relevant’ features,
the study is still randomized.

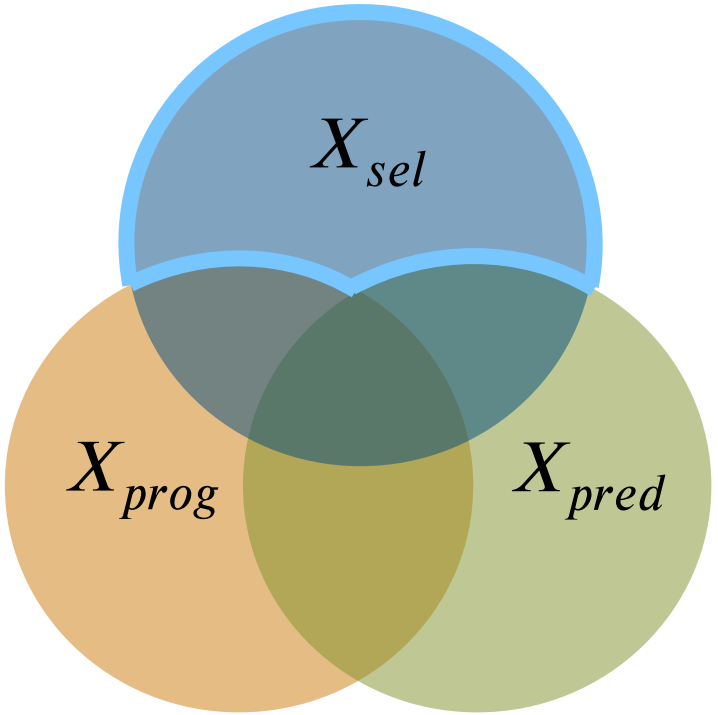
Irrelevant Features

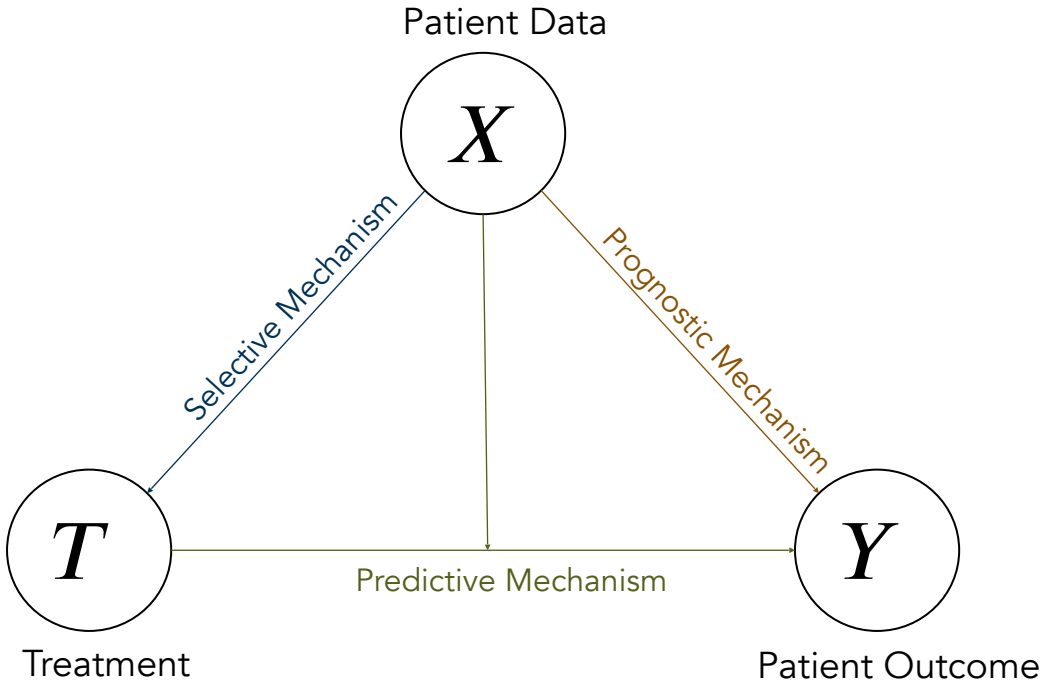




Irrelevant Features

Features that do not have any impact on the patient outcome but still introduce treatment selection bias





How can we formalize the type and strength of treatment selection bias?

Data Model

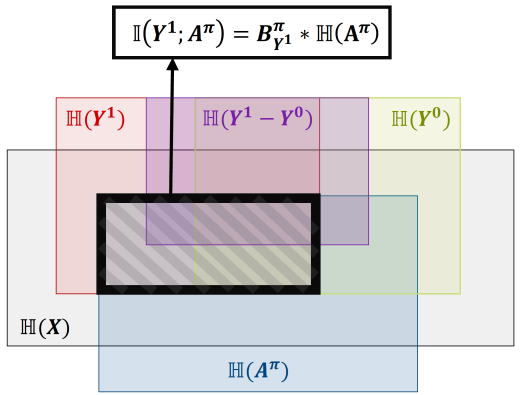
Definition 1 (Z-Bias) *The Z-bias for any treatment assignment* $A^\pi \sim \pi(X)$ *and random variable* Z *is defined as*

$$B_Z^\pi = \frac{\mathbb{I}(A^\pi; Z)}{\mathbb{H}[A^\pi]},$$

which measures the degree of bias of treatment assignment A^π *with respect to* Z.

- **Observable Bias** B_X^π
  Quantifies how much information all patient characteristics provide about treatment selection. This bias can also be seen to quantify the degree of violation of the overlap assumption in causal inference.
- **Outcome Biases** B_{Y0}^π and B_{Y1}^π
  Describe to what extent the potential outcomes of the treatments $a = 0$ and $a = 1$ determine the treatment decision.
- **Treatment Effect Bias** B_{Y1-Y0}^π
  Quantifies how much the treatment decision is biased with respect to the true individualized treatment effect.

Data Model

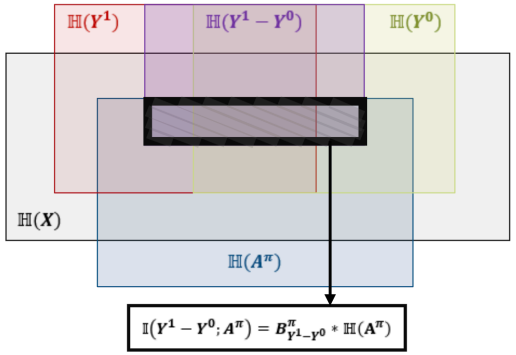


Definition 1 (Z-Bias) *The Z-bias for any treatment assignment* $A^\pi \sim \pi(X)$ *and random variable* Z *is defined as*

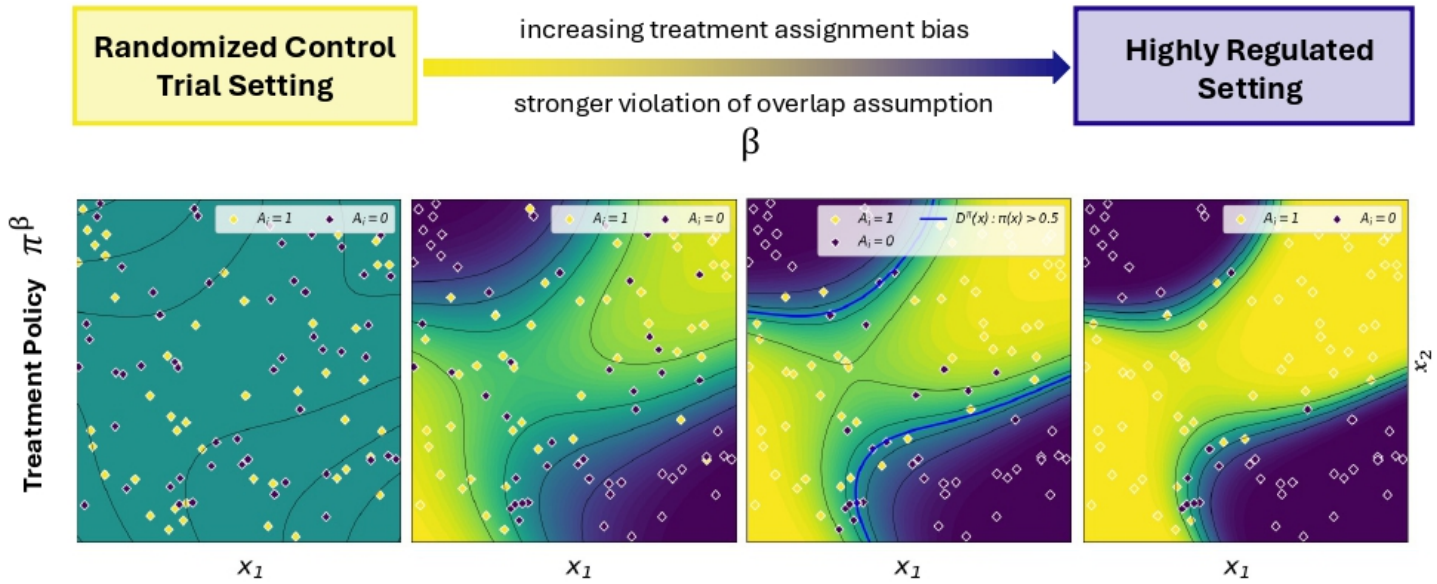
$$B^\pi_Z = \frac{\mathbb{I}(A^\pi; Z)}{\mathbb{H}[A^\pi]},$$

which measures the degree of bias of treatment assignment A^π *with respect to* Z.

- **Observable Bias** B^π_X
  Quantifies how much information all patient characteristics provide about treatment selection. This bias can also be seen to quantify the degree of violation of the overlap assumption in causal inference.
- **Outcome Biases** B^π_{Y0} and B^π_{Y1}
  Describe to what extent the potential outcomes of the treatments $a = 0$ and $a = 1$ determine the treatment decision.
- **Treatment Effect Bias** B^π_{Y1-Y0}
  Quantifies how much the treatment decision is biased with respect to the true individualized treatment effect.



Policy Simulation

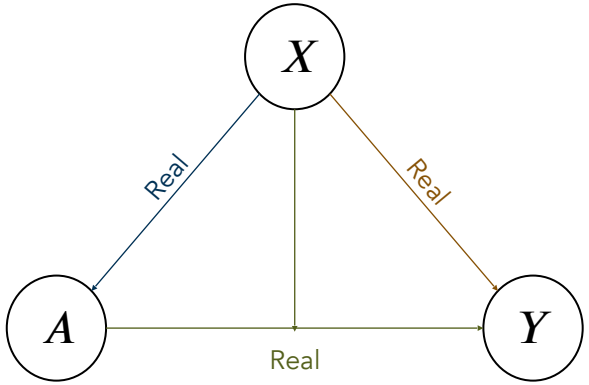


Definition 4 (Z-Policy) *For a given random variable Z and parameter β, the Z-policy is defined as*

$$\pi_Z^\beta(x) := \sigma(\beta Z(x)), \quad \sigma(x) := \frac{e^x}{1+e^x}$$

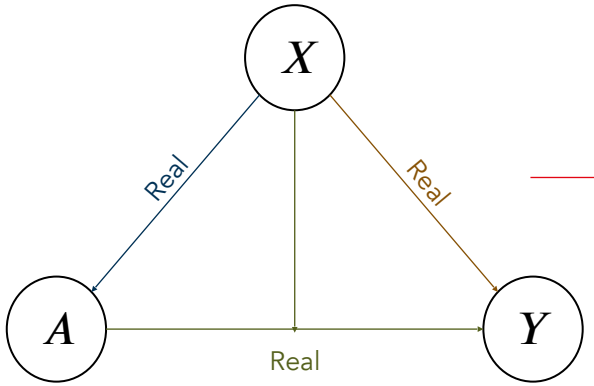
Simulations

No Simulation
(Clinical Data)



Simulations

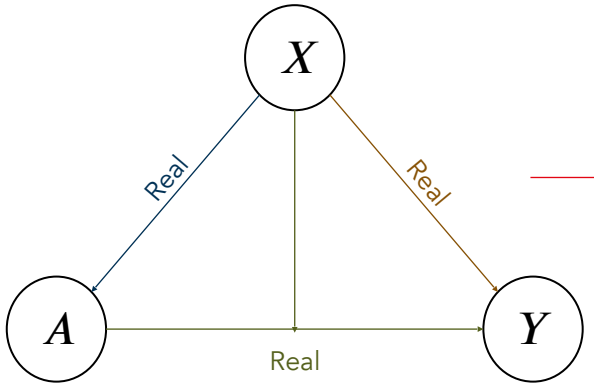
No Simulation
(Clinical Data)



No ground-truth effect available!

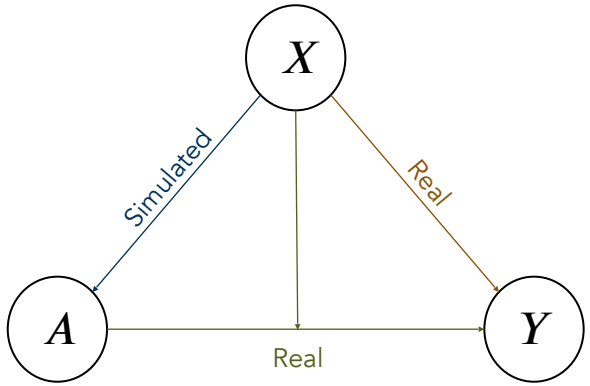
Simulations

No Simulation
(Clinical Data)



No ground-truth effect available!

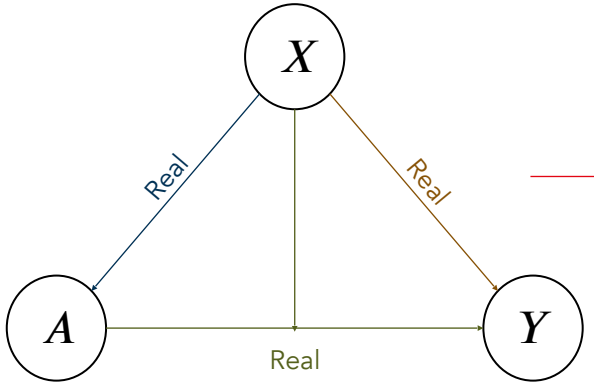
A-Simulation



Less Realistic
More Evaluation

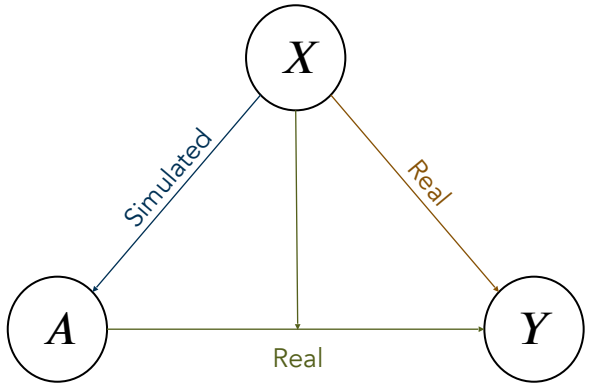
Simulations

No Simulation
(Clinical Data)

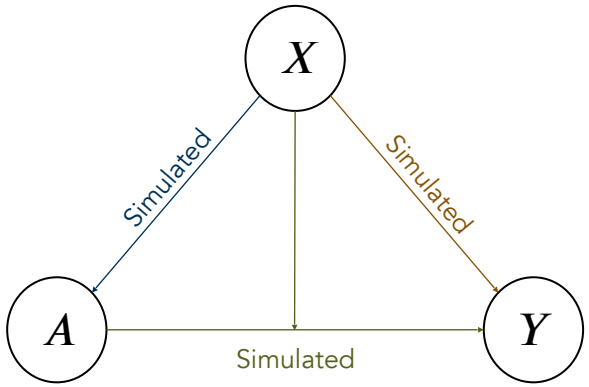


No ground-truth effect available!

A-Simulation



AY-Simulation

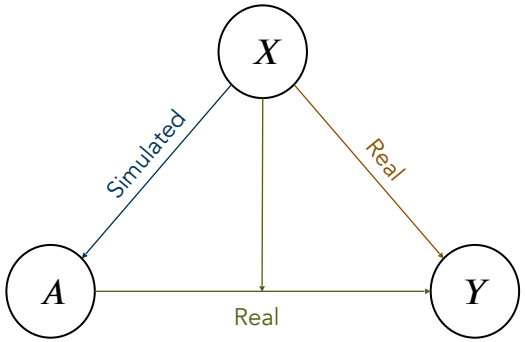


Less Realistic
More Evaluation

A-Simulation

‘patients’ … ~1000 cell lines as

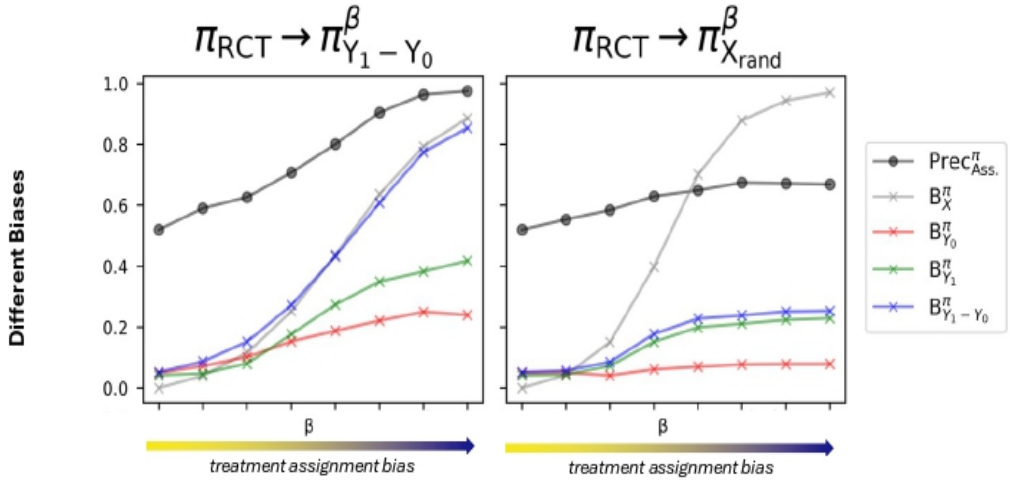
‘treatments’ … Drugs or CRISPR knock-outs as ‘treatments’



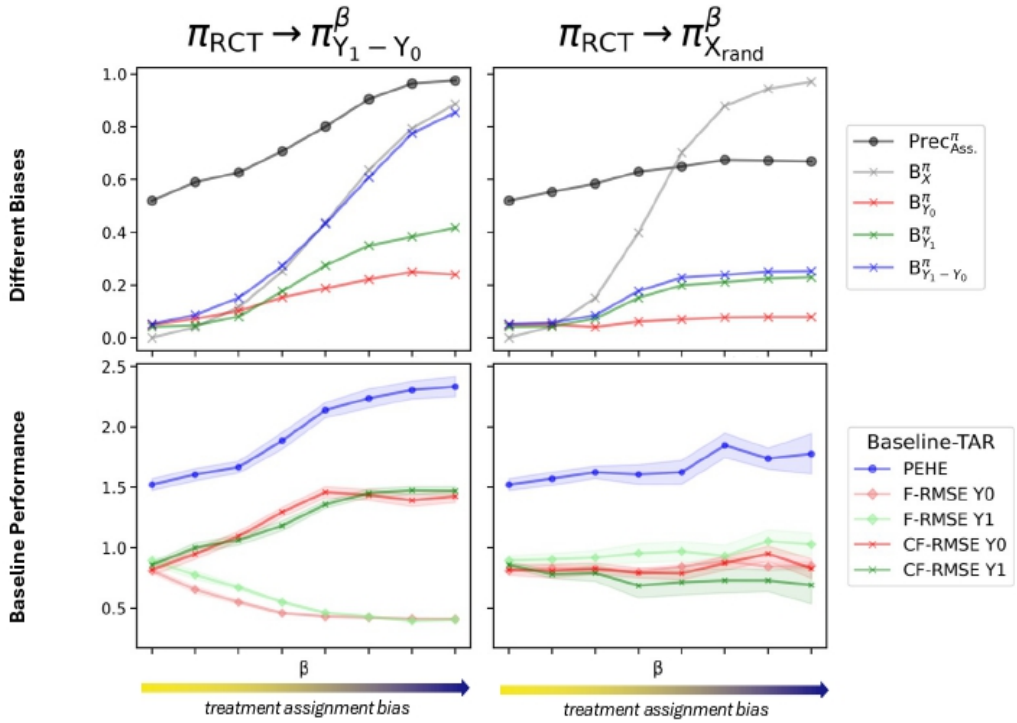
Results – Drug Screen

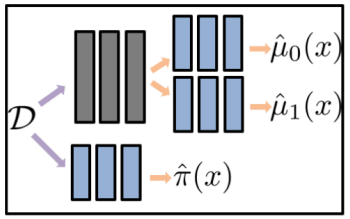
Definition 4 (Z-Policy) *For a given random variable Z and parameter β, the Z-policy is defined as*

$$\pi_Z^\beta(x) := \sigma(\beta Z(x)), \quad \sigma(x) := \frac{e^x}{1+e^x}$$



Results – Drug Screen



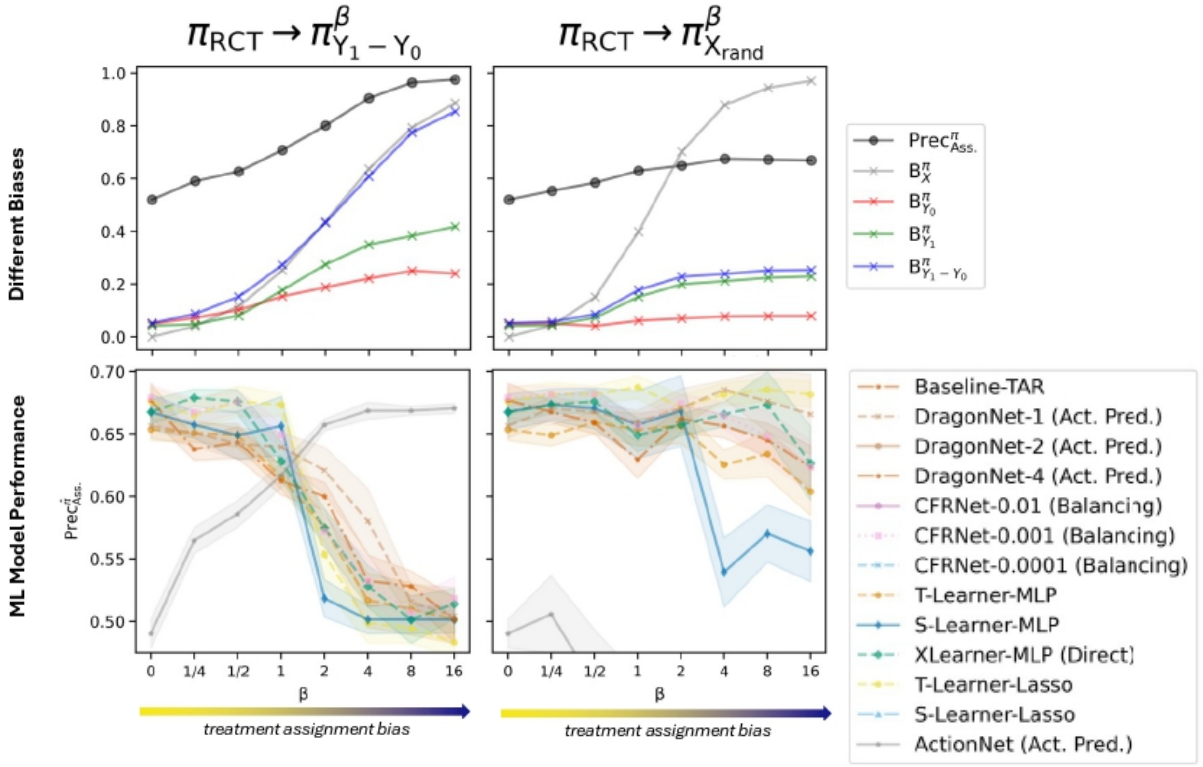


TARNet (SNet-1)

$$PEHE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}(x_i) - \tau(x_i))^2}, \qquad \tau(x) = E[Y_i^1 - Y_i^0 \mid X_i = x]$$

Alicia Curth and Mihaela Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

Results – Drug Screen



Conclusion

- Formalizing Treatment Biases
  Modeled various types of treatment assignment biases using mutual information.

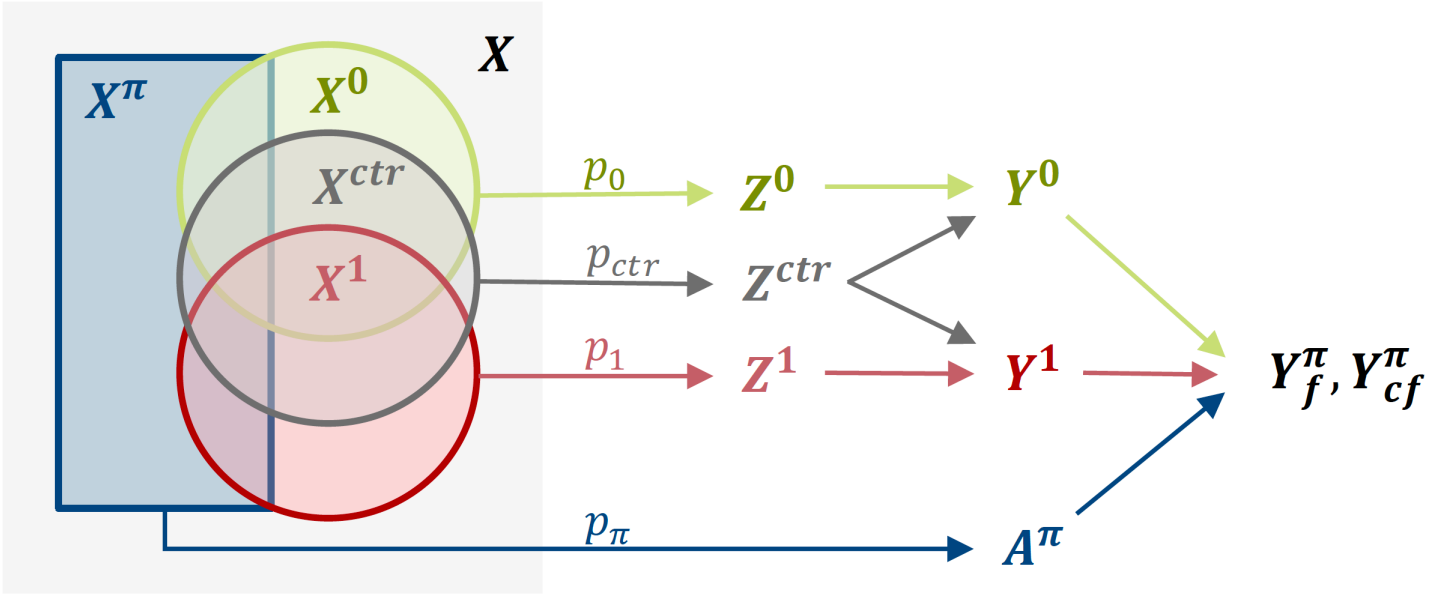
- Realistic Data Benchmarks
  Introduced novel biological datasets for robust evaluations.

- Bias Impact Analysis
  Provide simulation framework for how different biases affect counterfactual predictions and biomarker identification.

- Overlap Assumption
  Demonstrated that the violation of the overlap assumption is not necessarily detrimental.

- Model Development
  Highlighted the need for bias-aware ML models in precision medicine applications.

Data Model



$$A^{\pi} \sim \pi(X^{\pi})$$
$$Z^{a} \sim p_a(X^{a})$$
$$Z^{ctr} \sim p_{ctr}(X^{ctr})$$

$$Y^{a} := Z^{ctr} + Z^{a}$$
$$Y_f^{\pi} := A^{\pi} * Y^{1} + (1 - A^{\pi}) * Y^{0}$$
$$Y_{cf}^{\pi} := A^{\pi} * Y^{0} + (1 - A^{\pi}) * Y^{1}$$

Data Model

Proposition 2 *Under the non-confounding assumption, we have*

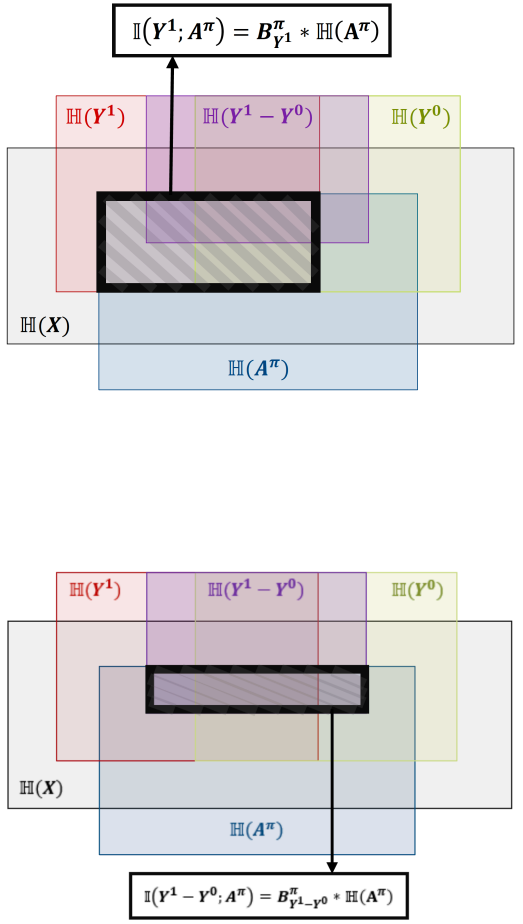
$$B_X^\pi \geq B_{Y^0,Y^1}^\pi \geq B_{Y^0}^\pi, B_{Y^1}^\pi, B_{Y^1-Y^0}^\pi.$$

Proposition 3 *Z-bias relates to the overlap assumption (OA) as follows*

$$B_X^\pi = 1 \implies \textit{OA is violated.}$$

Definition 4 (Z-Policy) *For a given random variable Z and parameter β, the Z-policy is defined as*

$$\pi_Z^\beta(x) := \sigma(\beta Z(x)),\ \sigma(x) := \frac{e^x}{1+e^x}$$



Datasets

PRISM repurposing drug screens. (DepMap and Kocak, 2024), performed on 877 cell lines according to Corsello et al. (2020) record the effect of hundreds of drugs on various cancer cell lines. The goal is to identify new therapeutic uses for existing drugs through a multiplexed approach. For our use case we focused on the measurements for two drugs, imatinib and az-628. Imatinib is a well-known tyrosine kinase inhibitor used mainly for chronic myeloid leukemia (CML) and gastrointestinal stromal tumors (GISTs). Az-628 is a selective inhibitor of RAF kinase, a key player in the MAPK/ERK signaling pathway, which is often involved in cancer development. As covariates we use RNA transcriptomics data, retaining the 200 most correlated features as covariates.

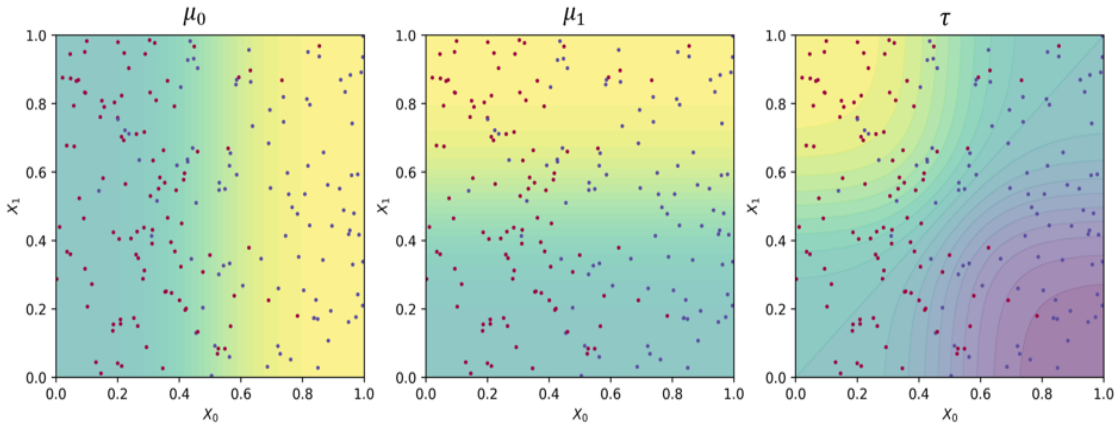
CRISPR knock-out screens. (DepMap, Broad, 2021) were conducted on 1067 cell lines following the methodology described by Behan et al. (2019), and targeted a myriad of genes of interest. They provide critical insights into gene dependencies by knocking out specific genes to observe their effects on cancer cell viability. As the two treatment options in our setting, we used the knock-out of genes EGFR (Epidermal Growth Factor Receptor) and KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog), which are crucial in cell signaling pathways related to cancer, with mutations or overexpression frequently observed in various cancers. Again, the most correlated 200 RNA transcriptomics features were used as covariates for our analysis. We refer to these datasets as A-CRISPR and A-DRUG.

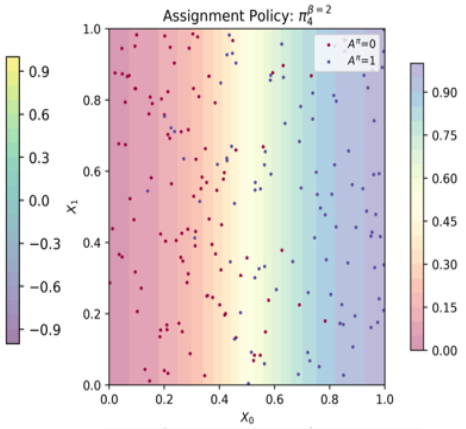
Fiona M Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, et al. Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature*, 568(7753):511–516, 2019.

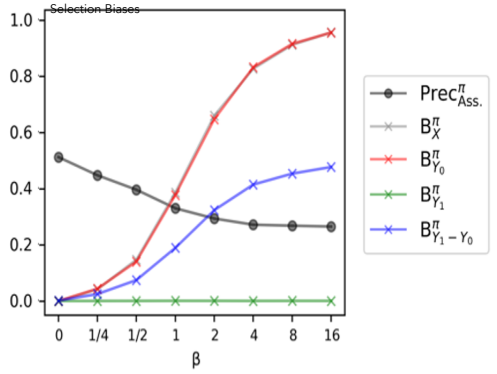
Steven M. Corsello, Rameen T. Nagari, Ryan D. Spangler, and et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, February 2020. doi: 10.1038/s43018-019-0018-6.

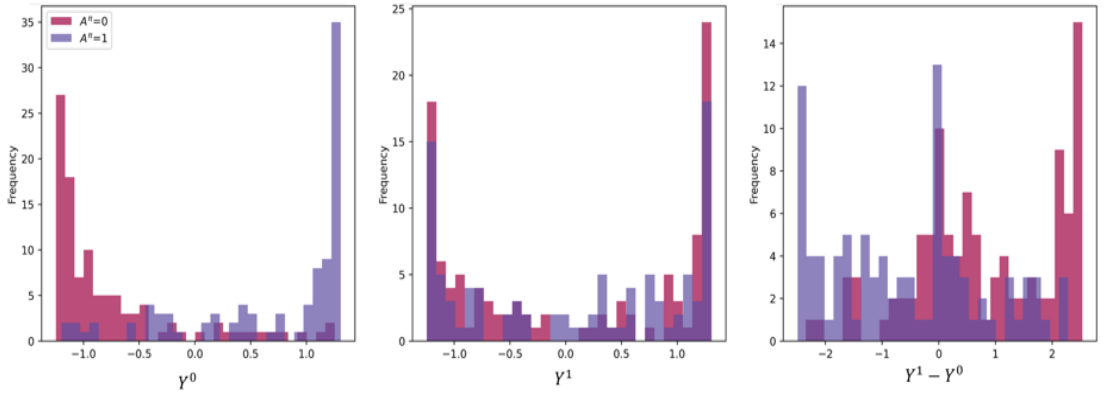
DepMap, Broad. Project score chronos. https://doi.org/10.6084/m9.figshare.14461980.v2, 2021. Dataset.

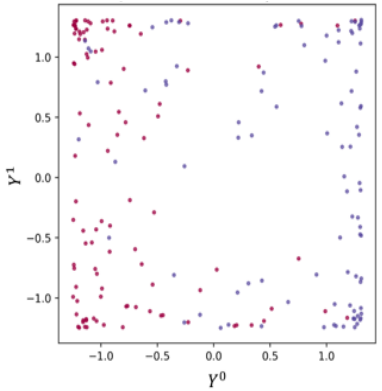
Toy Example

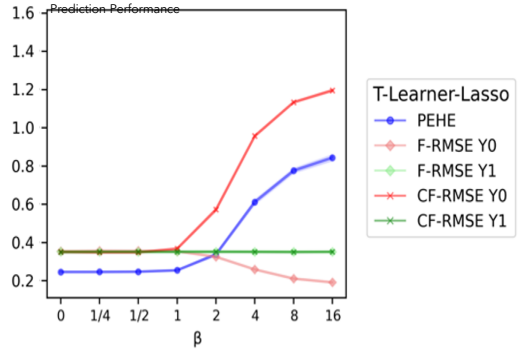










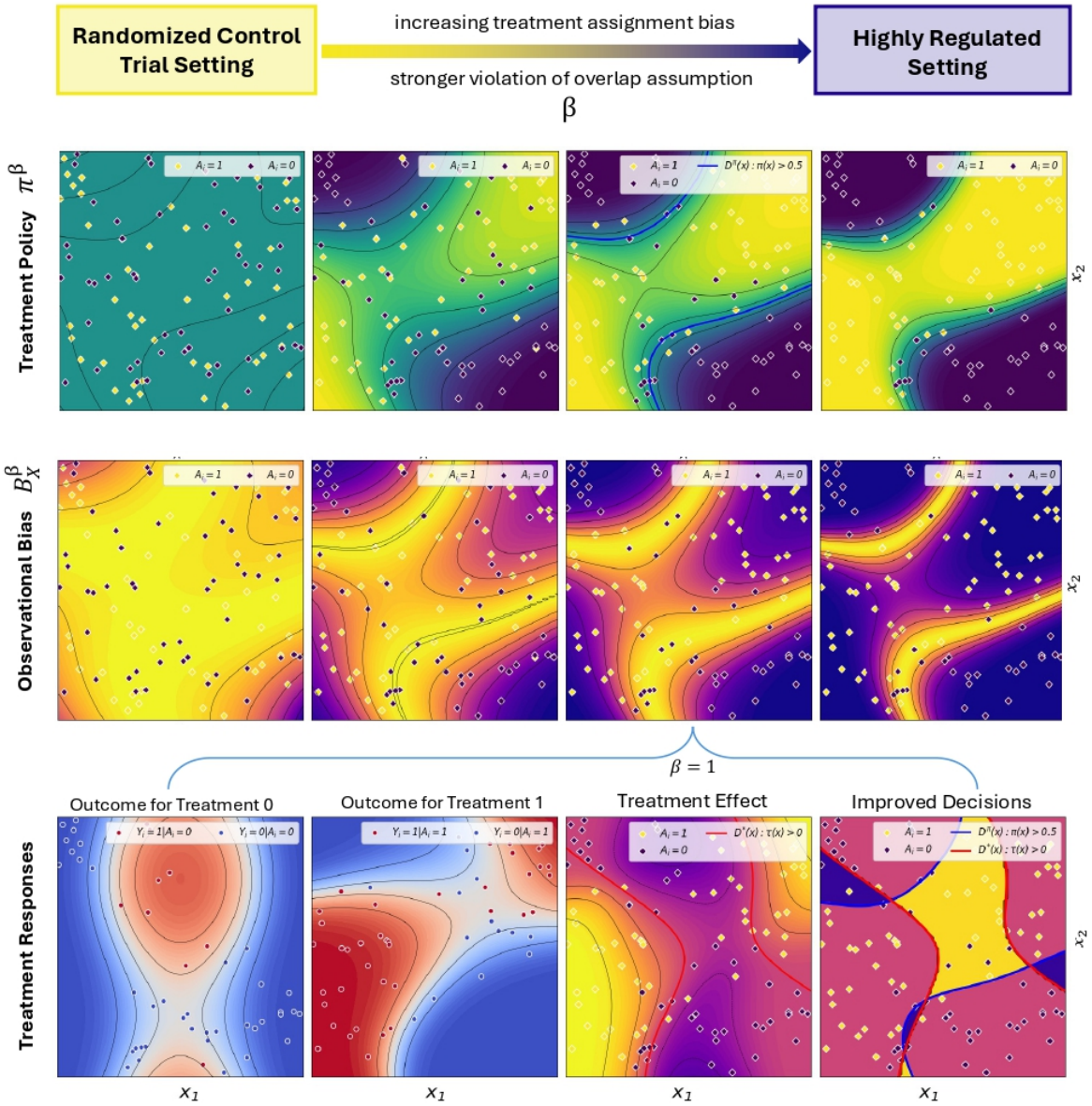


$$Y^a = f_{nl}(f_i^a(x_0, x_1))$$

$$f_4^0(x_0, x_1) := x_0$$
$$f_4^1(x_0, x_1) := x_1$$
$$\pi_4^\beta(x_0, x_1) \sim x_0$$

$$f_{nl}(\cdot) = \frac{1}{1+\exp(-10*(\cdot-0.5))}$$

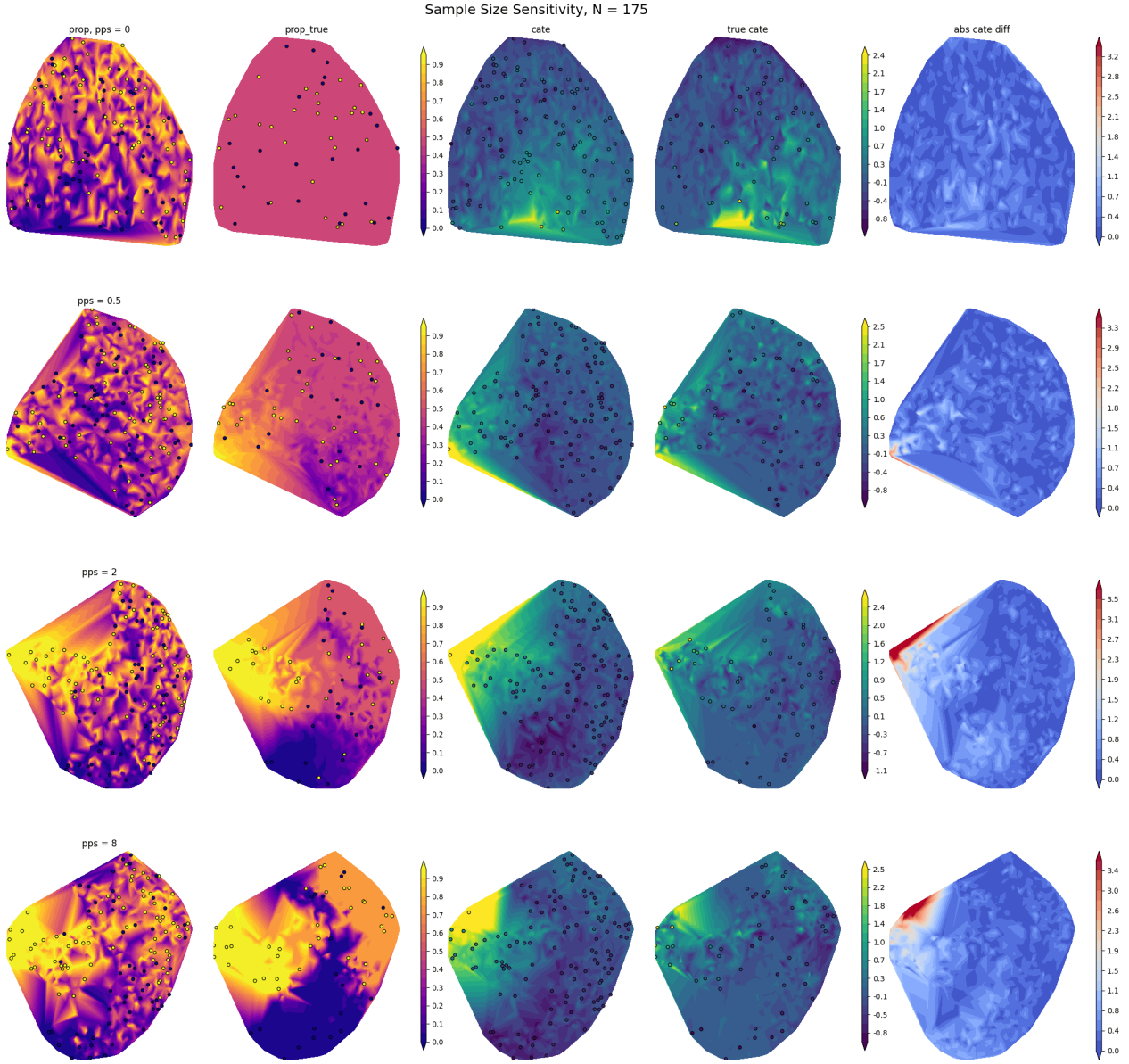
Toy Example


Randomized Control Trial Setting
increasing treatment assignment bias
stronger violation of overlap assumption
β
Highly Regulated Setting
Treatment Policy π^{β}
Observational Bias B_X^{β}
$\beta = 1$
Outcome for Treatment 0
Outcome for Treatment 1
Treatment Effect
Improved Decisions
Treatment Responses
x_1
x_2

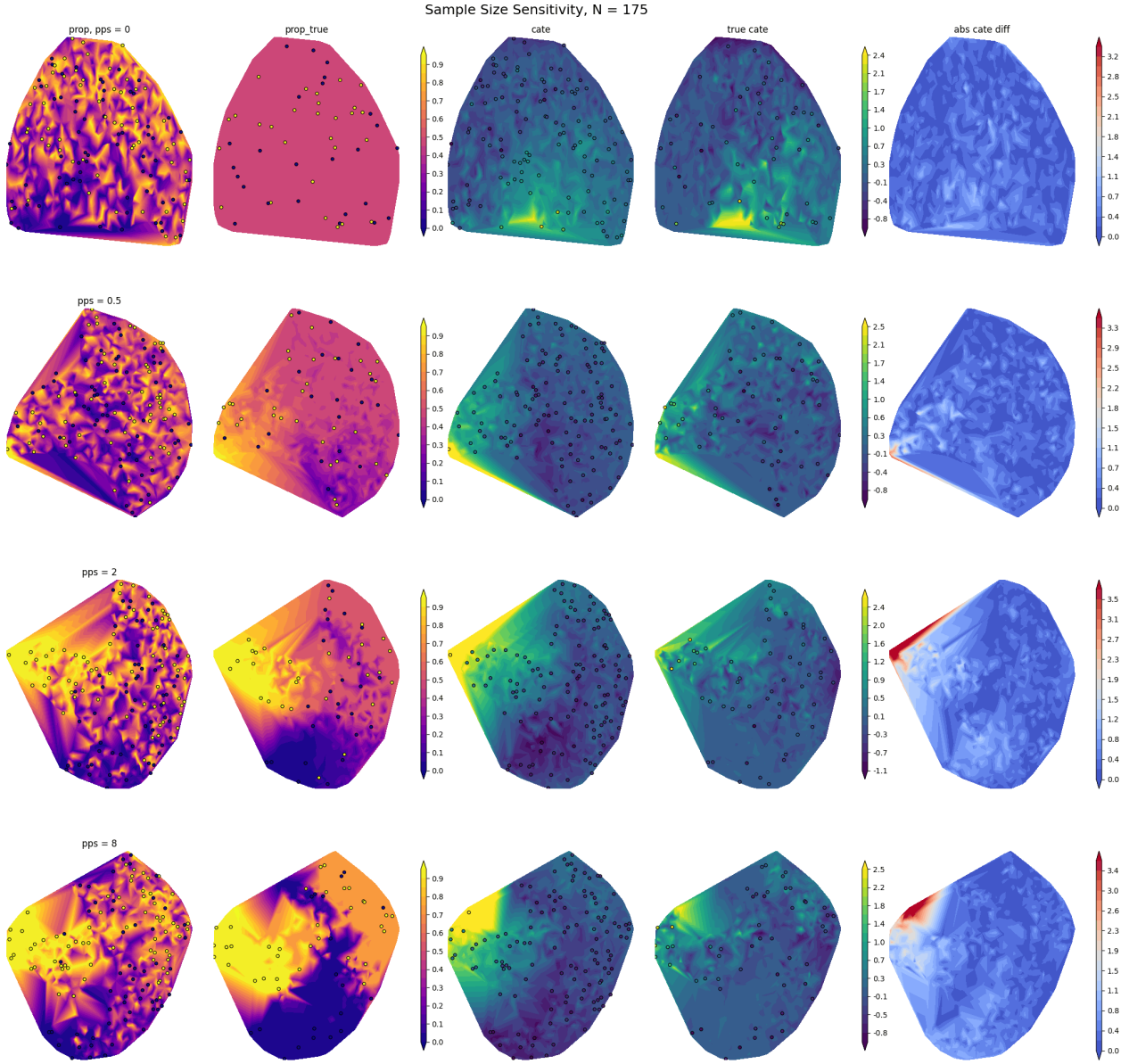
Simulation Settings

Setting	Overall Usefulness based on Mike's Intuition (0-10)	Overall Closeness to Reality (0-10)	Real Patient Data	Real Outcome Mechanism	Real Counterfactual Outcomes	Counterfactual Outcome and PEHE Evaluation	Selective Feature Importance Evaluation	Prognostic/ Predictive Feature Importance Evaluation	Sample Size Effect	Propensity Scale Effect, T Imbalance	Y Imbalance, #Num Dims #Treatments, #Imp. Feat.
Toy, TY-simulated	2 (debugging)	0	✗	✗	✗	✓	✓	✓	✓	✓	✓
In-vivo	3 (realistic, but no counterfactual evaluation)	10	✓	✓	✗	✗	✗	✗	✗	✗	✗
Real X, TY-simulated (linear)	4 (all evaluation types, but too simple)	3	✓	✗	✗	✓	✓	✓	✗	✓	✓
Synthetic X, TY-simulated (linear)	4 (all evaluation types, but very unrealistic)	2	✗	✗	✗	✓	✓	✓	✓	✓	✓
Real X, TY-simulated (Y learned via ML)	7 (more realistic, but too few samples)	6	✓	~	✗	✓	✓	✓	✗	✓	✓
Synthetic X, TY-simulated (Y learned via ML)	8 (more samples, all evaluation, but synthetic data)	5	✗	~	✗	✓	✓	✓	✓	✓	✓
Real X, TY-simulated (biological sim.)	8 (more realistic, domain knowledge based, but too few samples)	7	✓	~	~	✓	✓	✓	✗	✓	✓
Synthetic X, TY-simulated (biological sim.)	8 (more samples, all evaluation, domain knowledge based, but synthetic data)	5	✗	~	~	✓	✓	✓	✓	✓	✓
In-vitro	9 (realistic, counterfactual, but no propensity analysis)	10	✓	✓	✓	✓	✗	✗	✗	✗	✗
In-vitro, T-simulated	9 (realistic, counterfactual, propensity analysis, but few samples)	9	✓	✓	✓	✓	✓	✗	✗	✓	✗
In-vitro, T-simulated, Large	10 (realistic, counterfactual, propensity analysis, sample size analysis)	9	✓	✓	✓	✓	✓	✗	✓	✓	✗

Simulation Settings



Simulation Settings



A-Drug Results

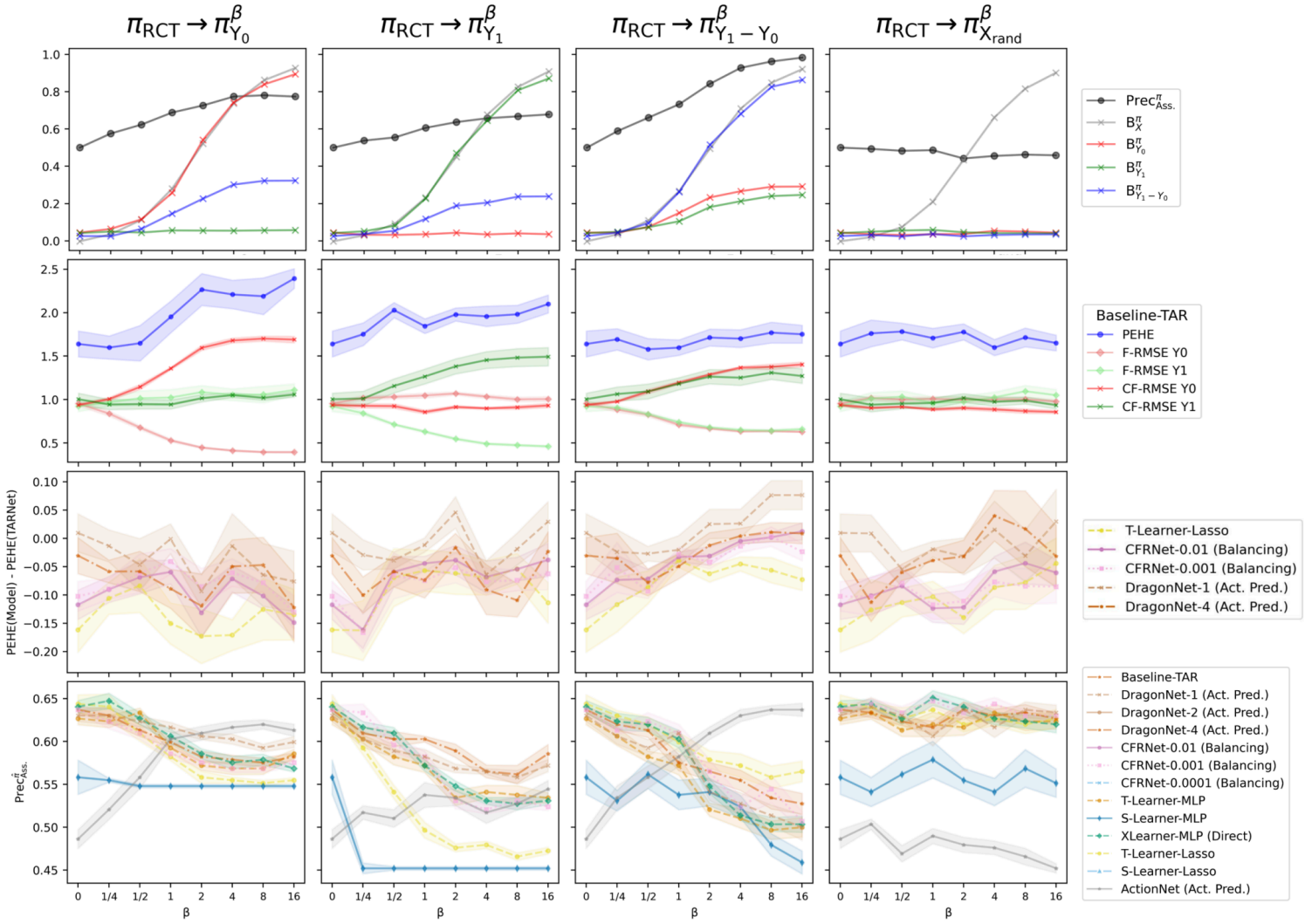
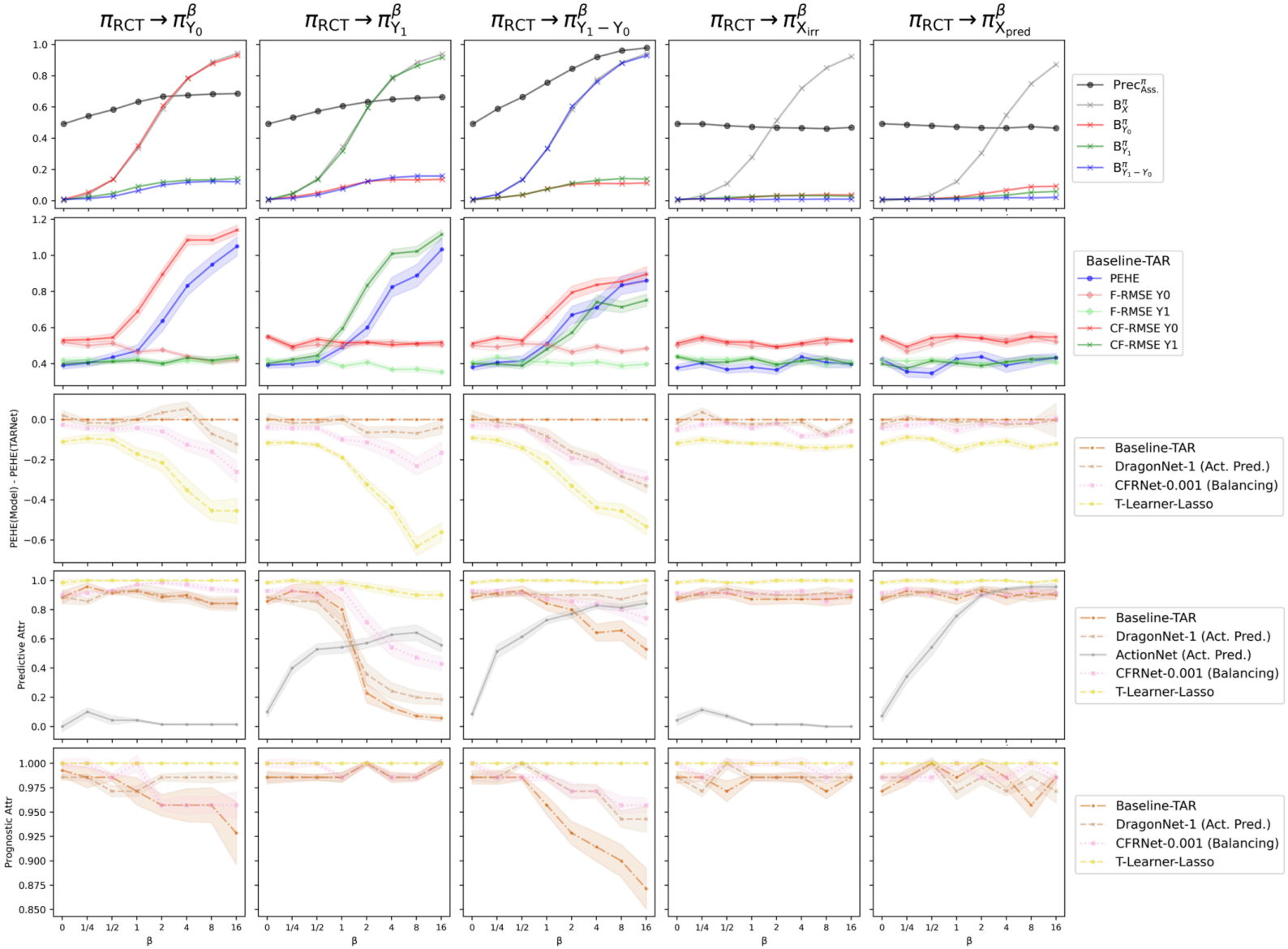
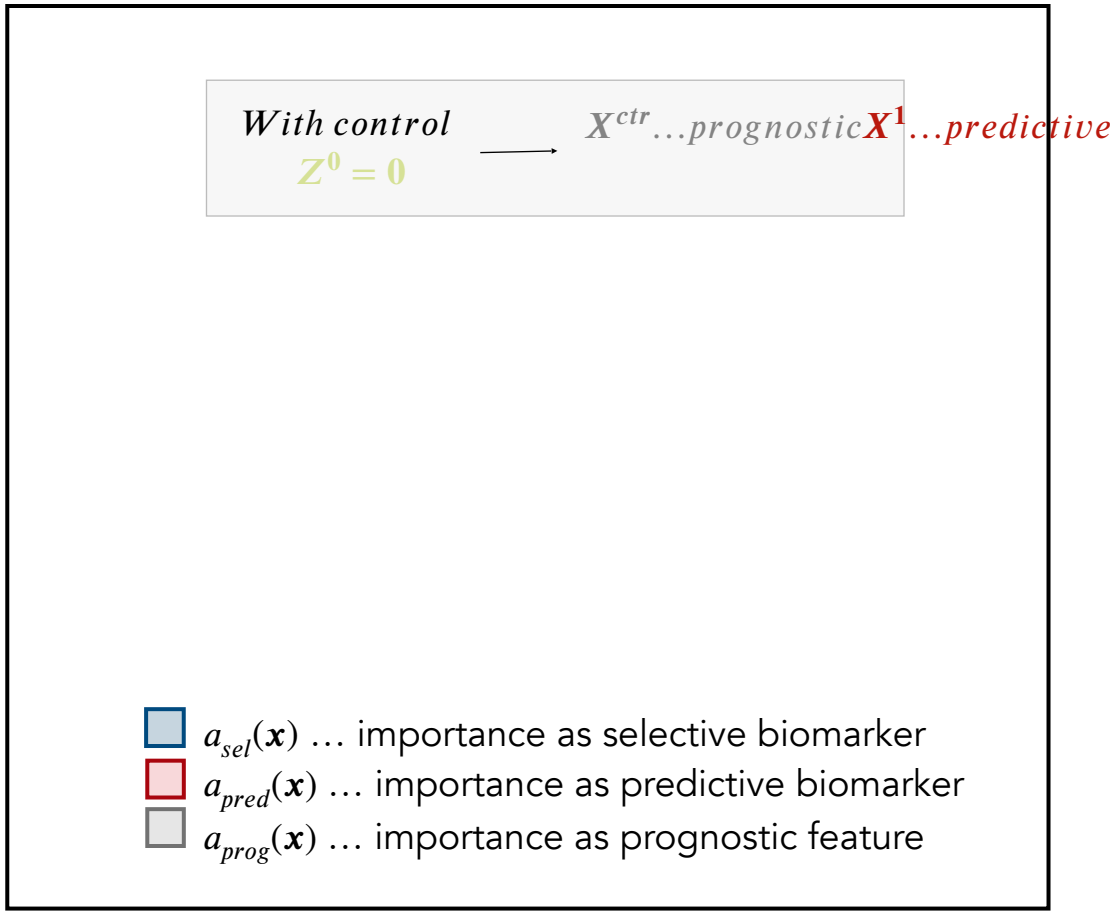


Figure 9: Results for A-Drug.

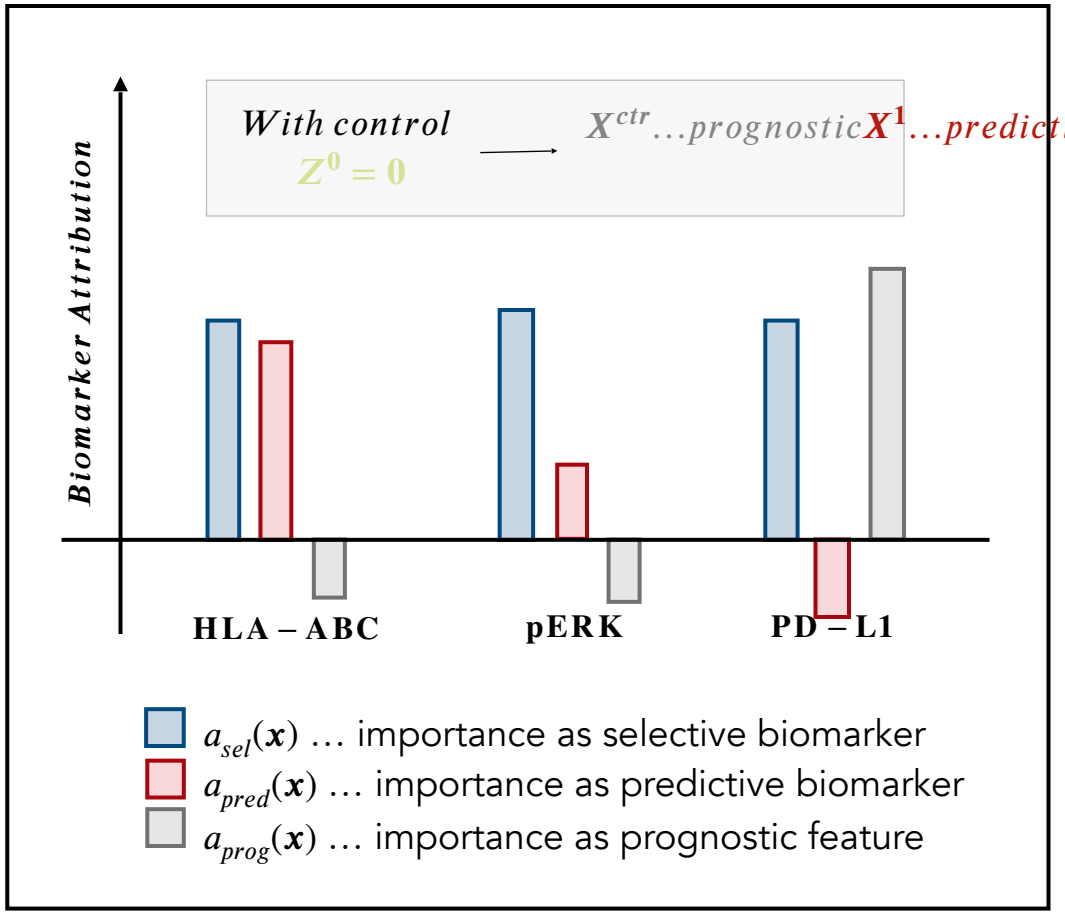
AY-TCGA Results



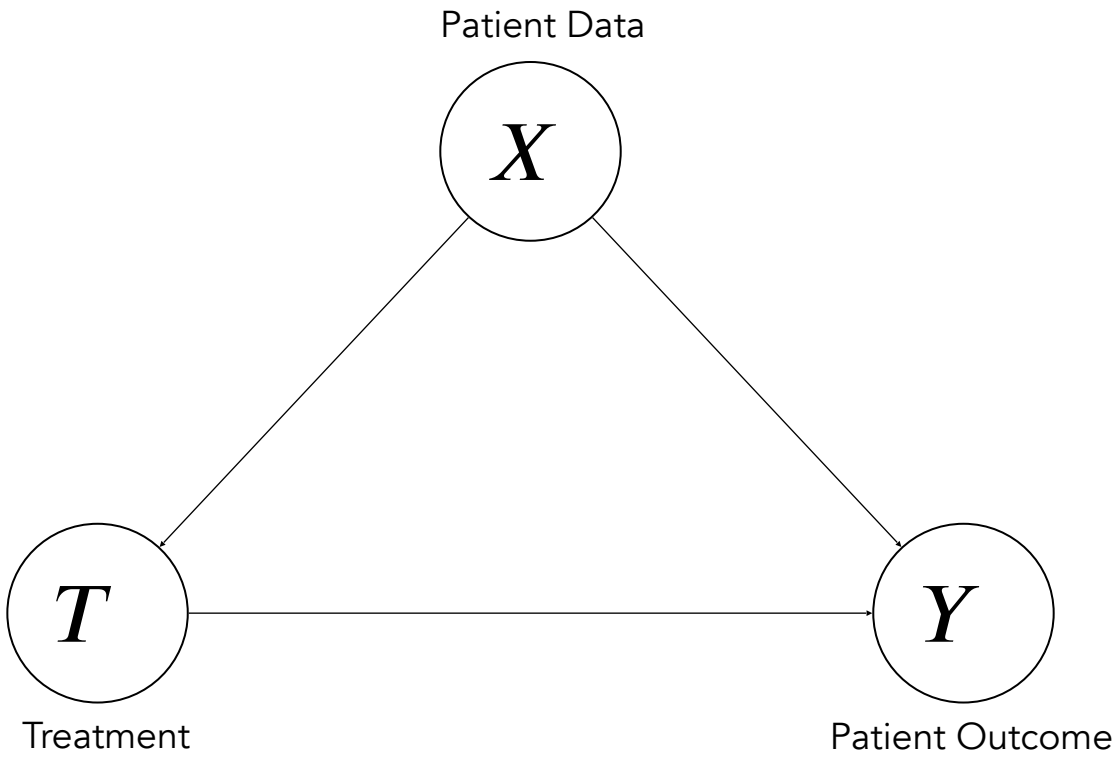
Simulation Settings



Simulation Settings

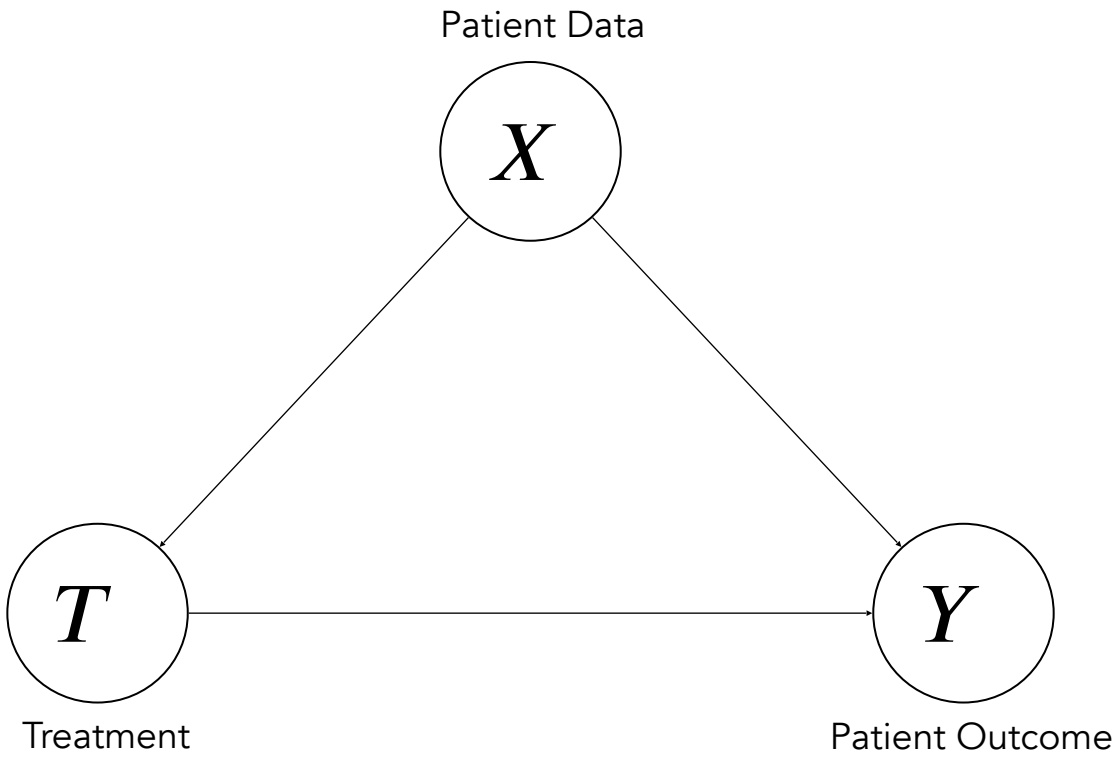


Counterfactuals



Counterfactuals

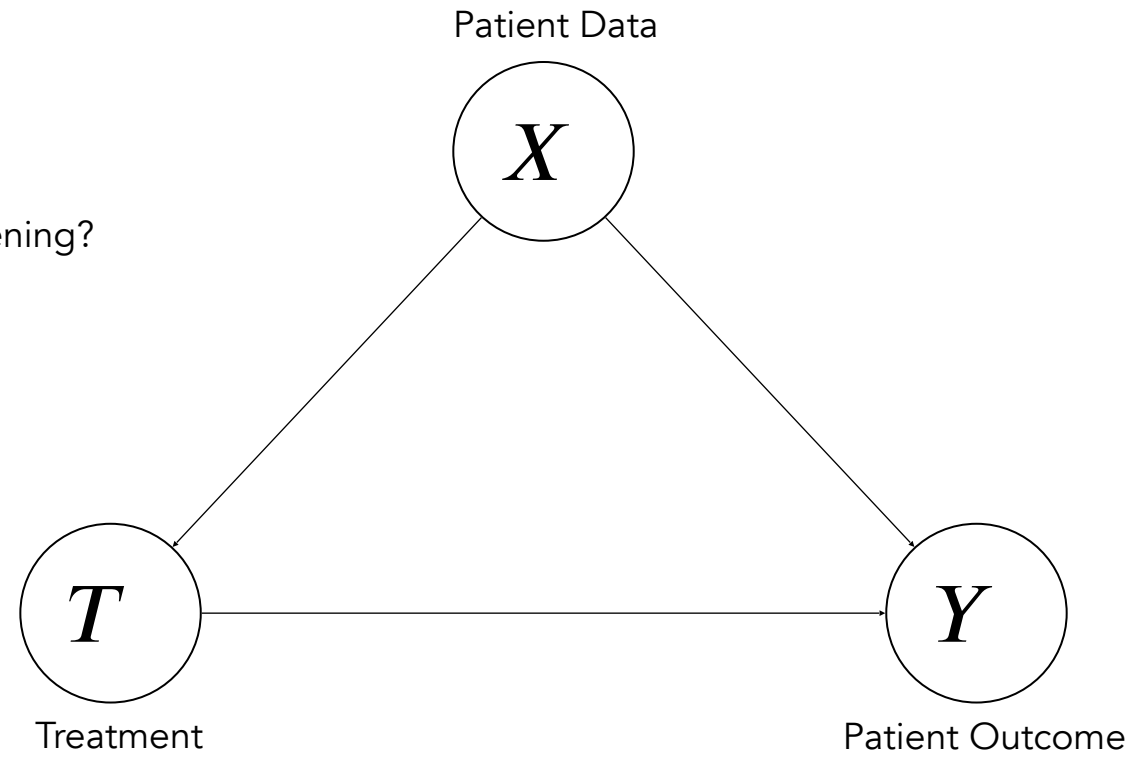
$P(Y \mid X, T)$



Counterfactuals

$P(Y \mid X, T)$

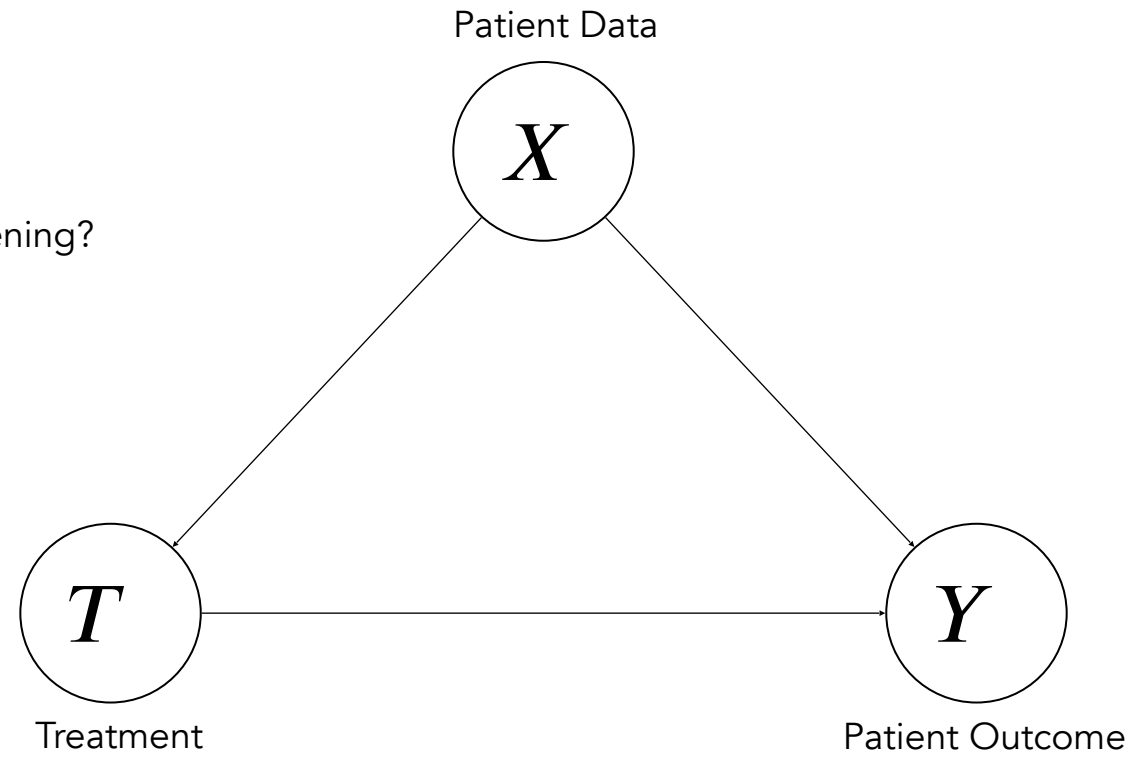
What is happening?



Counterfactuals

$P(Y \mid X, T)$

What is happening?

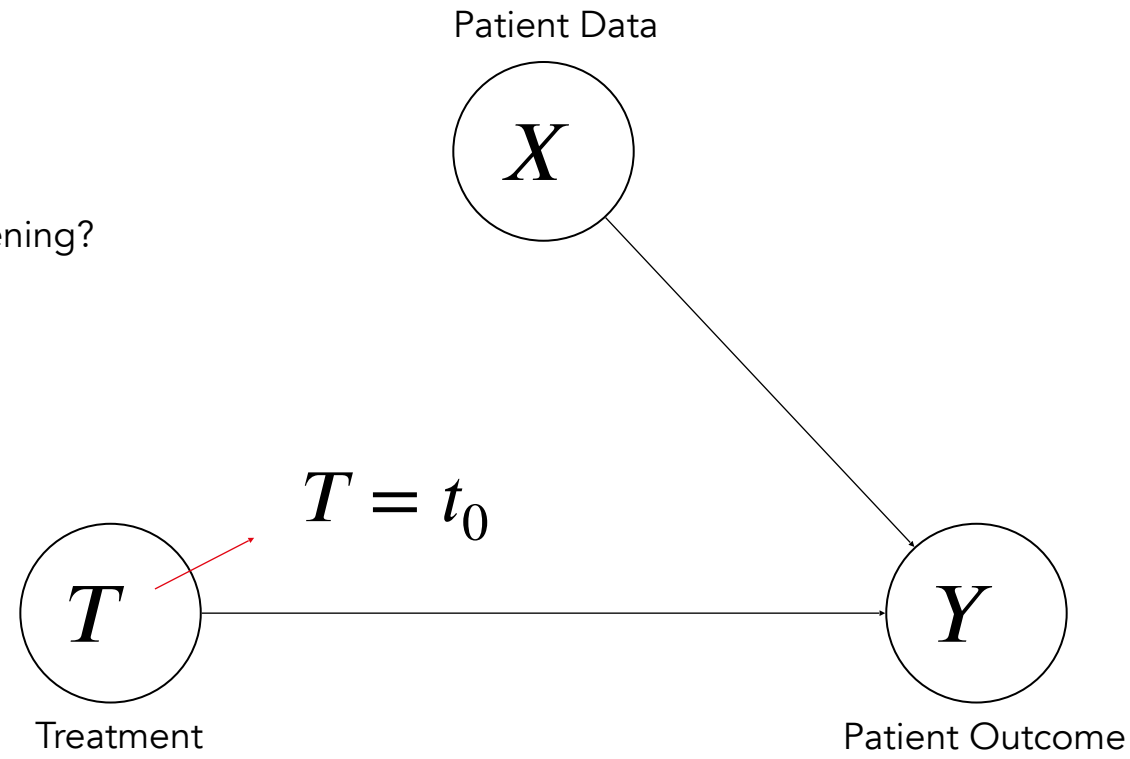


Counterfactuals

$P(Y \mid X, T)$

What is happening?

What happens if?

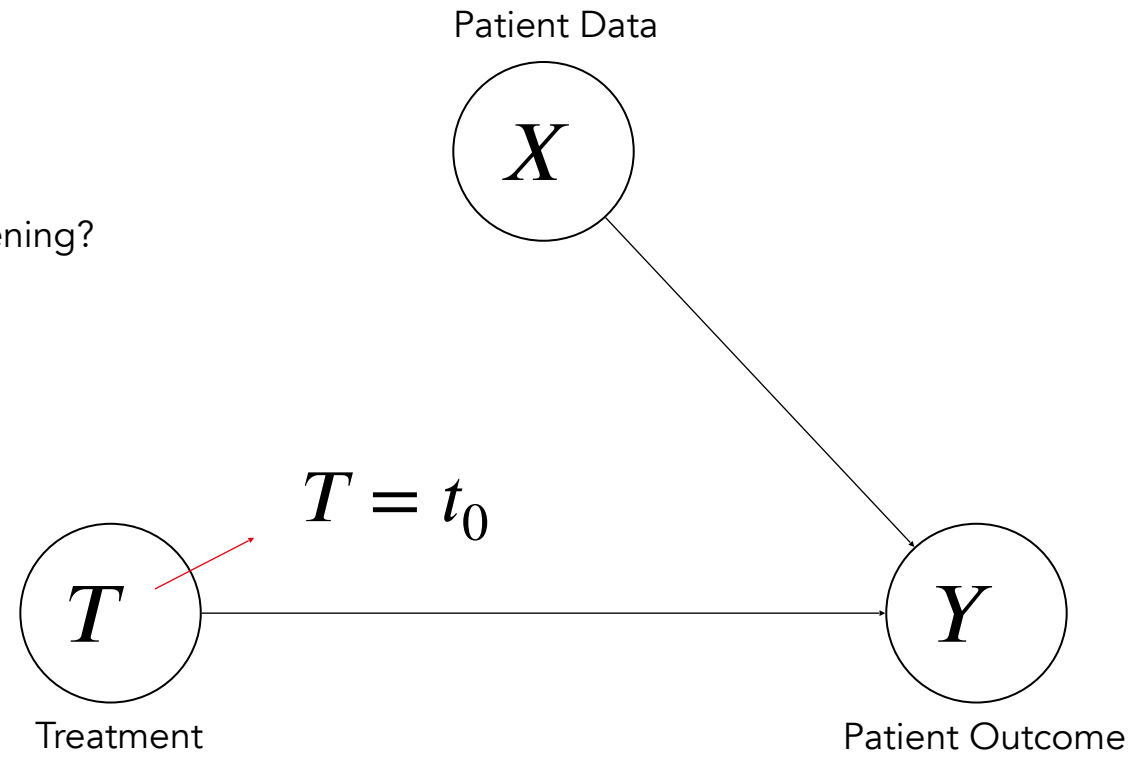


Counterfactuals

$P(Y \mid X, T)$

What is happening?

What happens if?



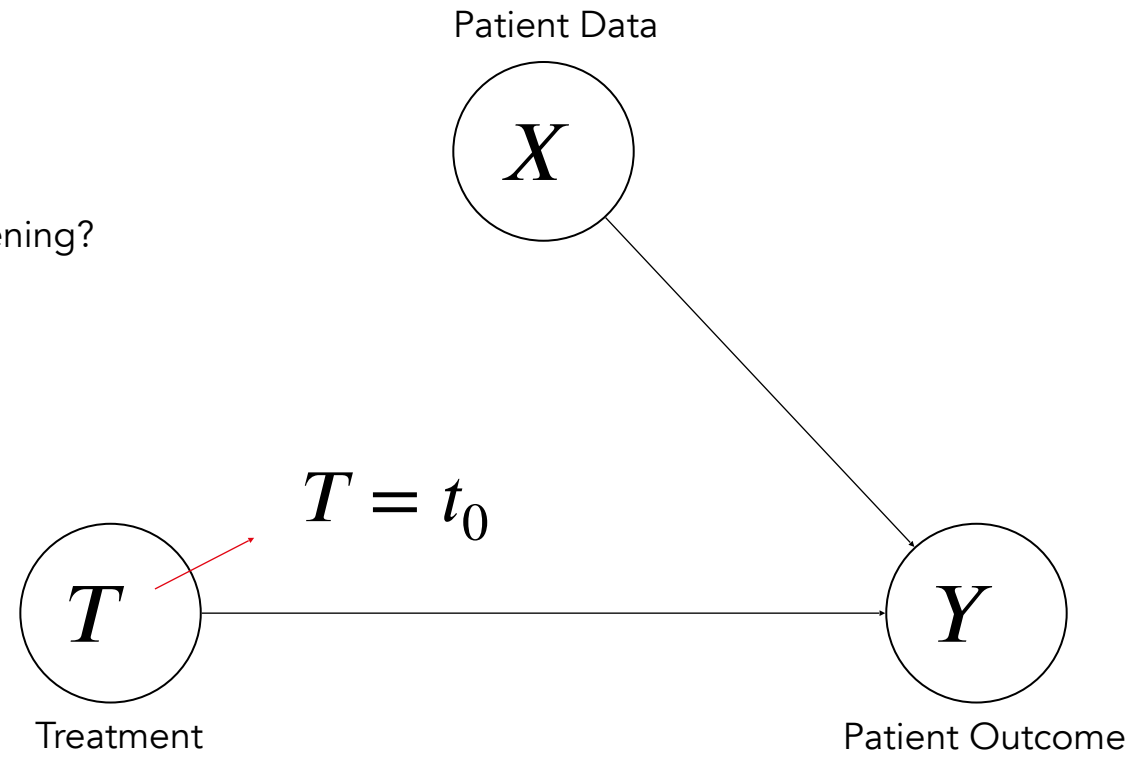
Counterfactuals

$P(Y \mid X, T)$

What is happening?

What happens if?

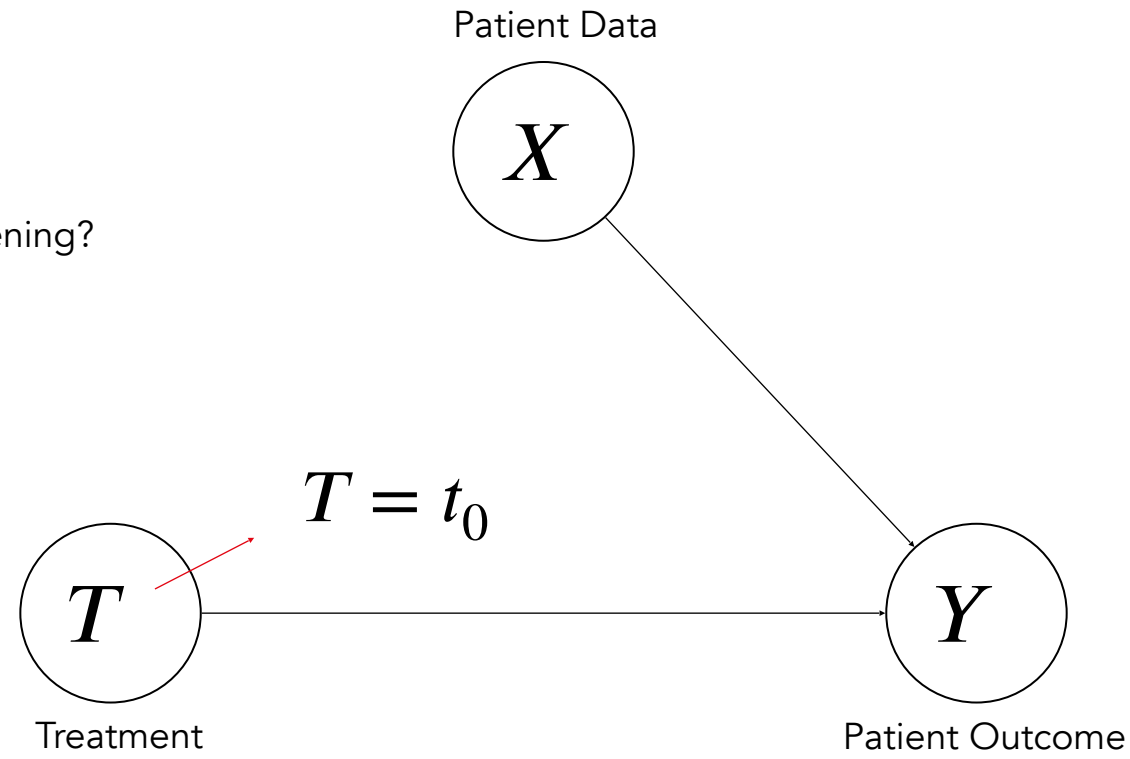
What would have happened if?



Counterfactuals

$P(Y \mid X, T)$ → What is happening?

What happens if? → $T = t_0$



What would have happened if? → This is what we want to learn from data!

Clinical Biomarkers

Predictive

A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

Prognostic

A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

Glossary - BEST (Biomarkers, EndpointS, and other Tools) Resource - NCBI Bookshelf (nih.gov)

Clinical Biomarkers

Predictive

A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

Prognostic

A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

Glossary - BEST (Biomarkers, EndpointS, and other Tools) Resource - NCBI Bookshelf (nih.gov)

Clinical Biomarkers

Susceptibility/Risk
A biomarker that indicates the potential for developing a disease or medical condition in an individual who does not currently have clinically apparent disease or the medical condition.

Diagnostic
A biomarker used to detect or confirm presence of a disease or condition of interest or to identify individuals with a subtype of the disease.

Monitoring
A biomarker measured repeatedly for assessing status of a disease or medical condition or for evidence of exposure to (or effect of) a medical product or an environmental agent.

Safety
A biomarker measured before or after an exposure to a medical product or an environmental agent to indicate the likelihood, presence, or extent of toxicity as an adverse effect.

Prognostic
A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

Predictive
A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

Response
A biomarker used to show that a biological response, potentially beneficial or harmful, has occurred in an individual who has been exposed to a medical product or an environmental agent.

Glossary - BEST (Biomarkers, EndpointS, and other Tools) Resource - NCBI Bookshelf (nih.gov)

Clinical Biomarkers

Prognostic

A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

Predictive

A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

Glossary - BEST (Biomarkers, EndpointS, and other Tools) Resource - NCBI Bookshelf (nih.gov)

Clinical Biomarkers

Prognostic

A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

Predictive

A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

Glossary - BEST (Biomarkers, EndpointS, and other Tools) Resource - NCBI Bookshelf (nih.gov)

True vs. Clinical

True vs. Clinical

Clinical Predictive Biomarker → "Selective Feature"

A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

→ How should the patient be treated?

True vs. Clinical

Clinical Predictive Biomarker → "Selective Feature"

A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

→ How should the patient be treated?

Clinical Prognostic Biomarker → "Selective Feature"

A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

→ Should the patient be treated?

True vs. Clinical

Clinical Predictive Biomarker → "Selective Feature"

A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

→ How should the patient be treated?

Clinical Prognostic Biomarker → "Selective Feature"

A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

→ Should the patient be treated?

True Predictive Biomarker → "Predictive Feature"

A biomarker ~~used to identify~~ determining individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

→ What happens when the patient is treated?

True vs. Clinical

Clinical Predictive Biomarker → "Selective Feature"

A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

⟶ How should the patient be treated?

Clinical Prognostic Biomarker → "Selective Feature"

A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

⟶ Should the patient be treated?

True Predictive Biomarker → "Predictive Feature"

A biomarker ~~used to identify~~ determining individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

⟶ What happens when the patient is treated?

True Prognostic Biomarker → "Prognostic Feature"

A biomarker ~~used to identify~~ determining the likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

⟶ What happens when the patient remains untreated?