Causal Machine Learning Methods for Heterogeneous Treatment Effect Detection

Séminaire STATQAM

Phil Boileau — November 2025

Collaborators

Hani Zaki

Mireille Schnitzer

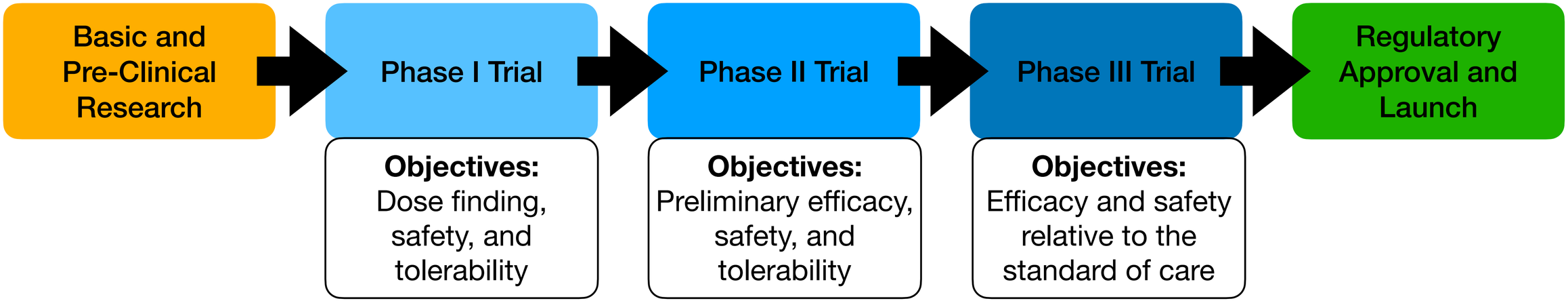
Gabriele Lileikyte

Patrick Lawler

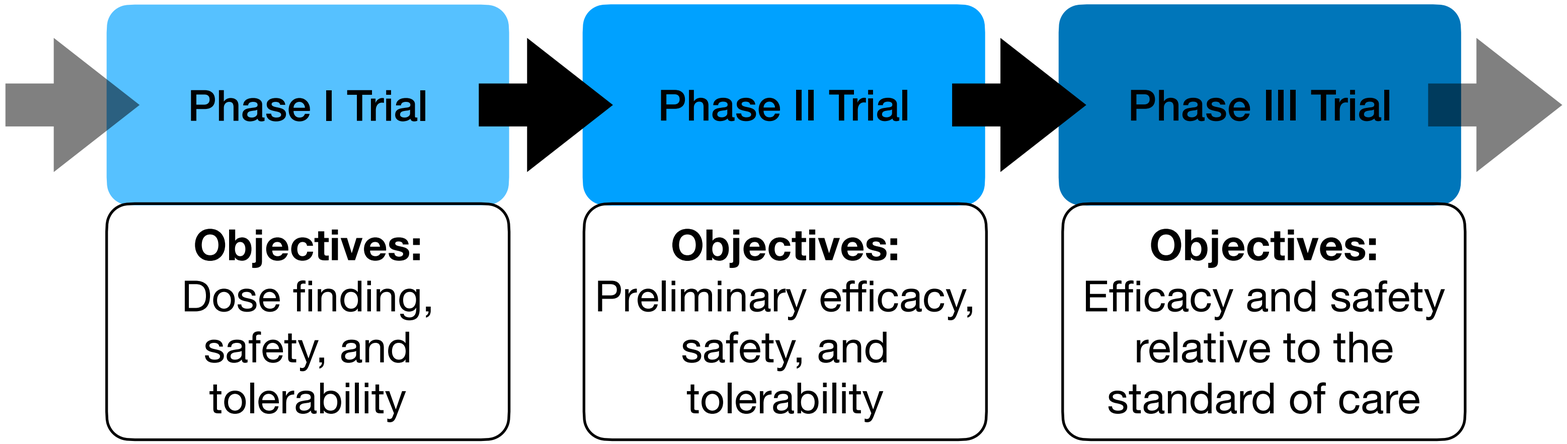
Niklas Nielsen

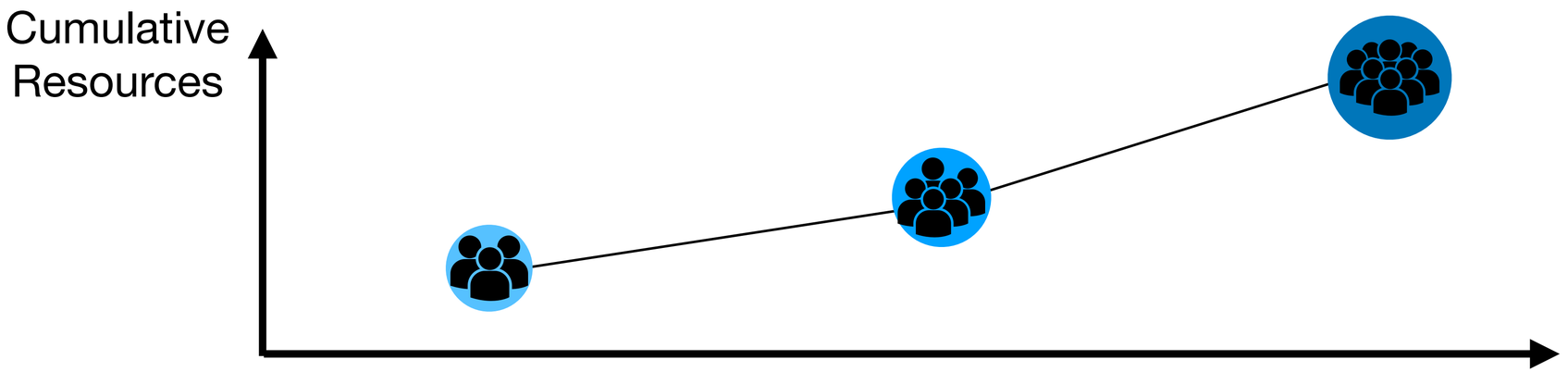
Motivation

Drug Development Pipeline

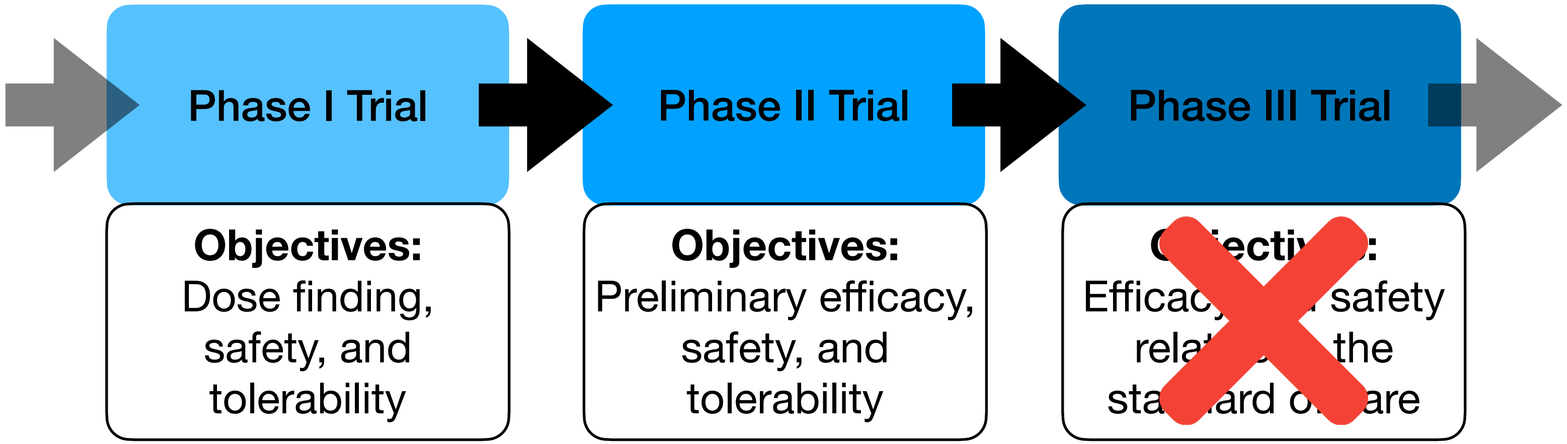


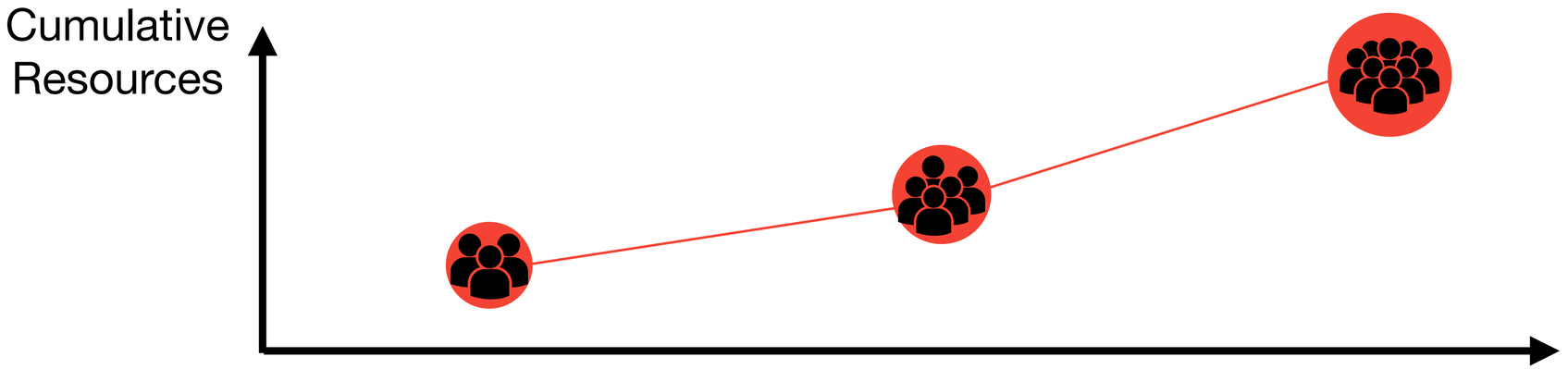
Drug Development Pipeline





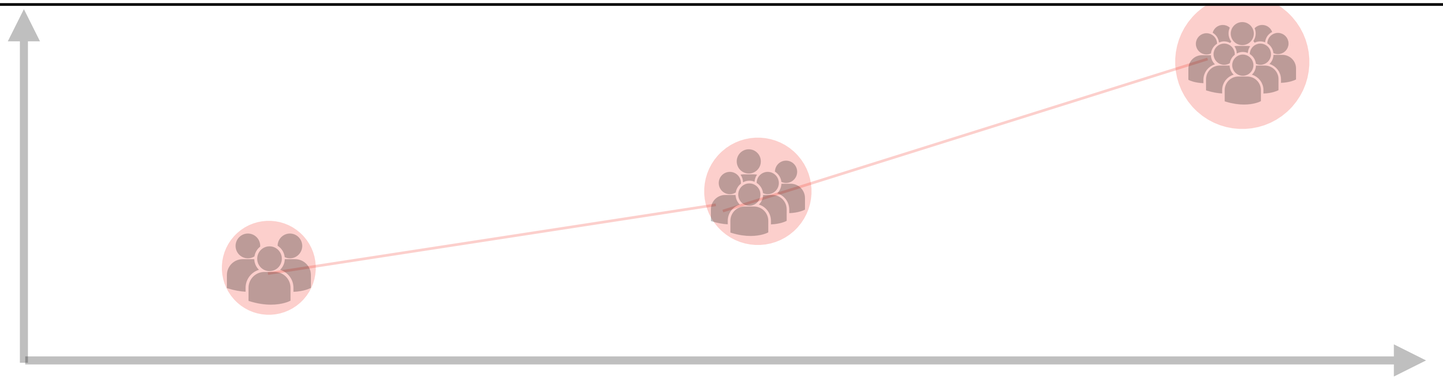
Drug Development Pipeline





Drug Development Pipeline

54% of late-stage trials failed among the 634 identified in public and commercial databases between 1998 and 2008.



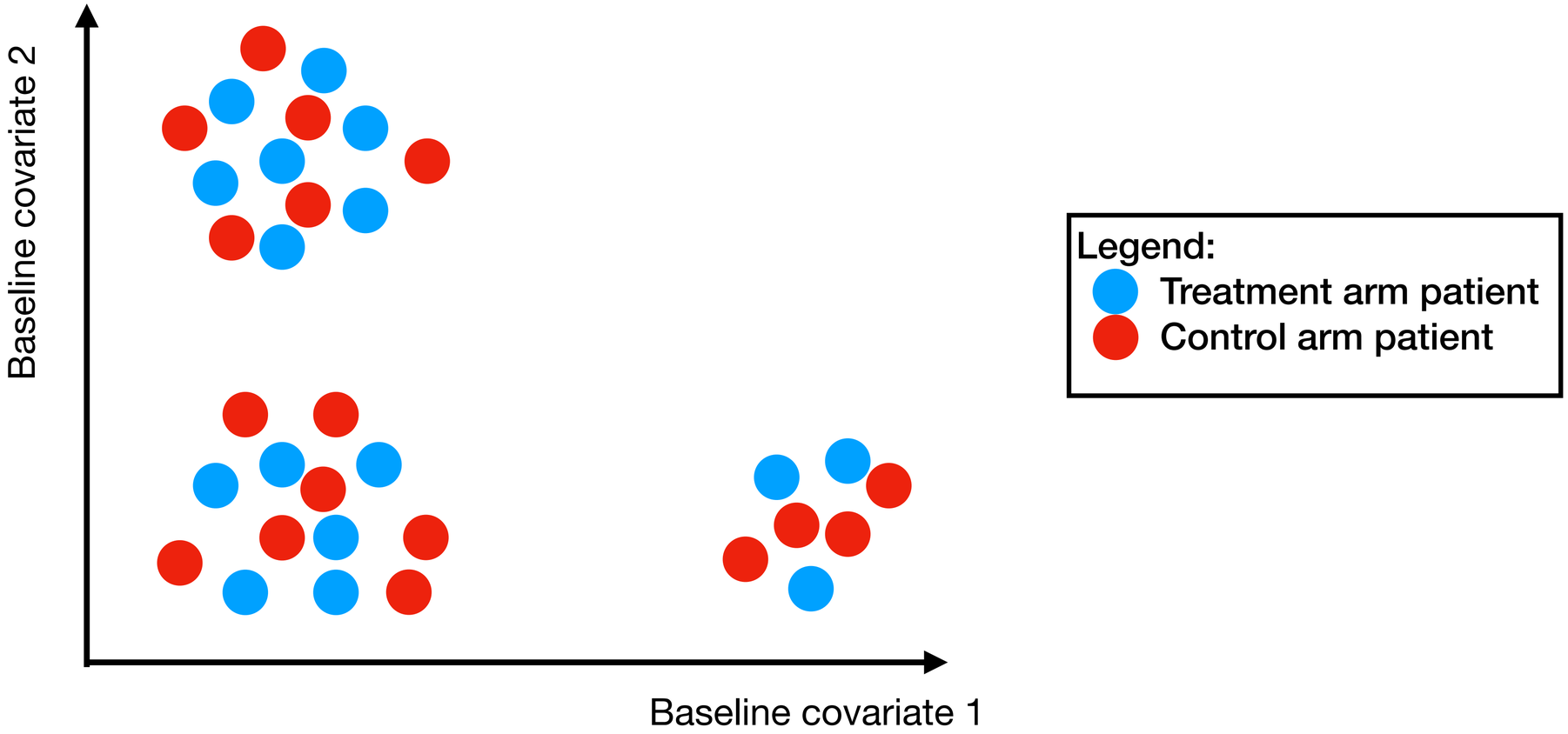
Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Intern Med.* 2016;176(12):1826–1833. doi:10.1001/jamainternmed.2016.6008

Avoiding Failures in Late-Stage Trials

- Late-stage trials fail frequently:
  - Sponsors waste resources that would have been better allocated elsewhere
  - Patients do not participate in other trials of potentially efficacious therapy
- Understanding why late-stage trials fail is needed to:
  - **Mitigate risk** of failure in future trials
  - **Salvage failed trial data** to inform future developments

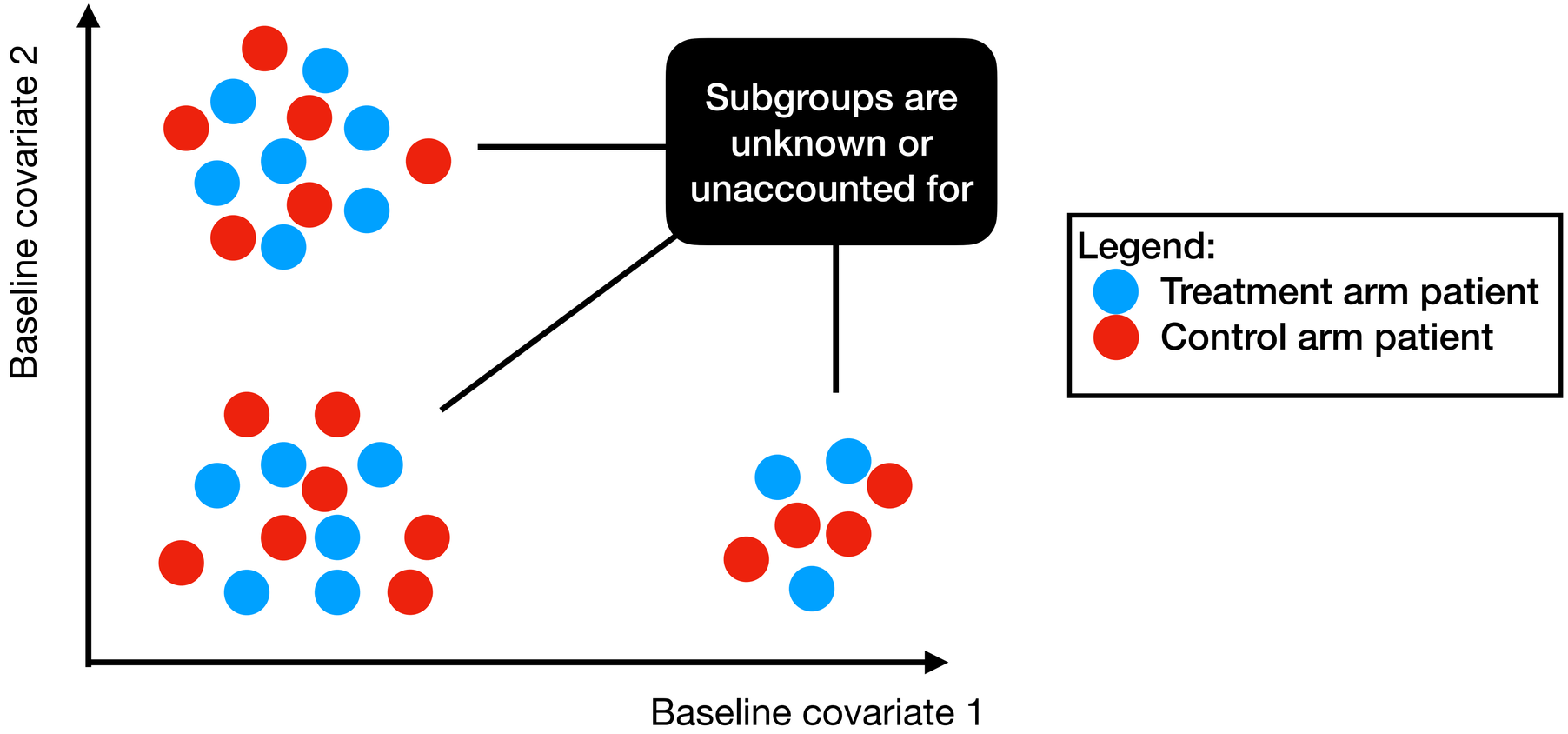
Heterogeneous Subgroups Drive Results

Sketch of a Phase II Trial



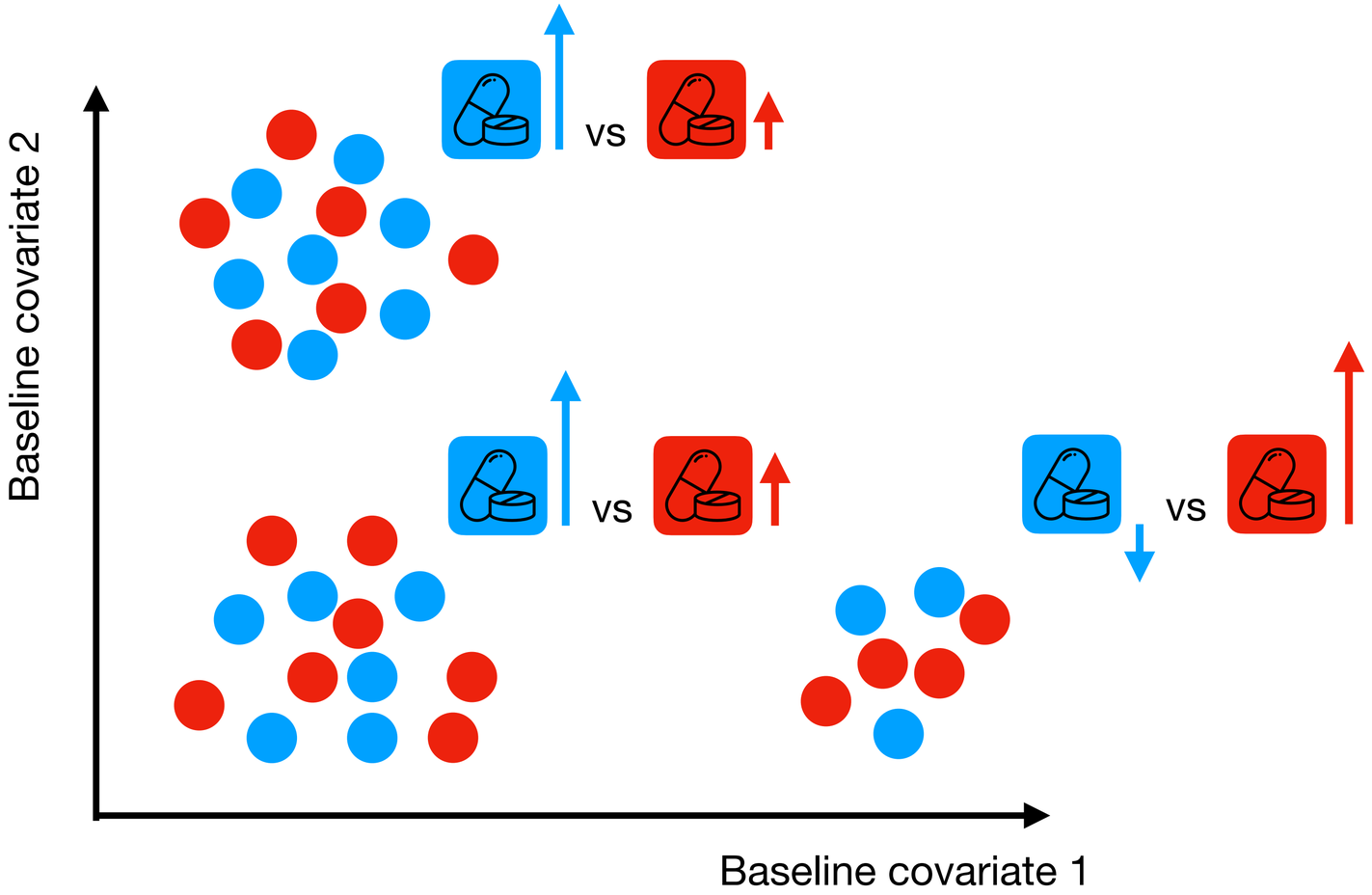
Heterogeneous Subgroups Drive Results

Sketch of a Phase II Trial



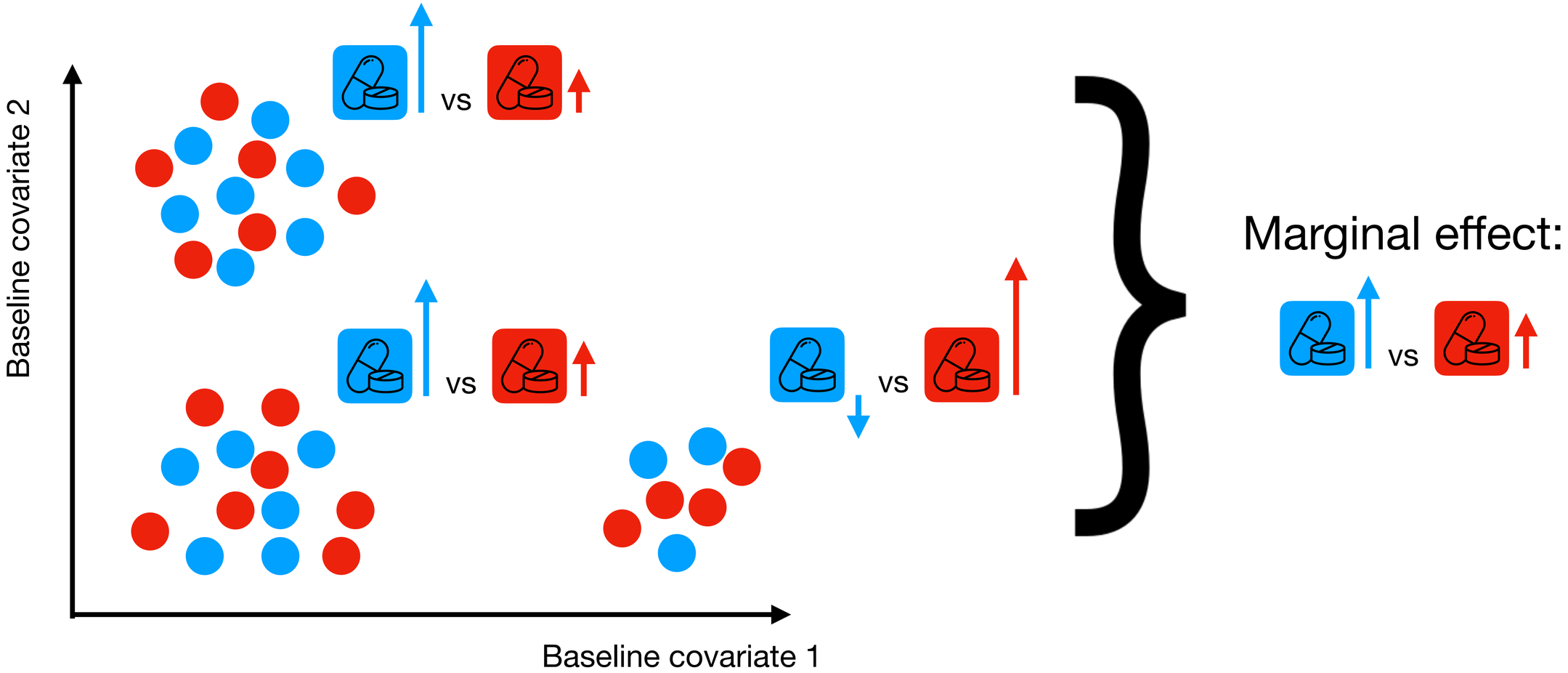
Heterogeneous Subgroups Drive Results

Sketch of a Phase II Trial



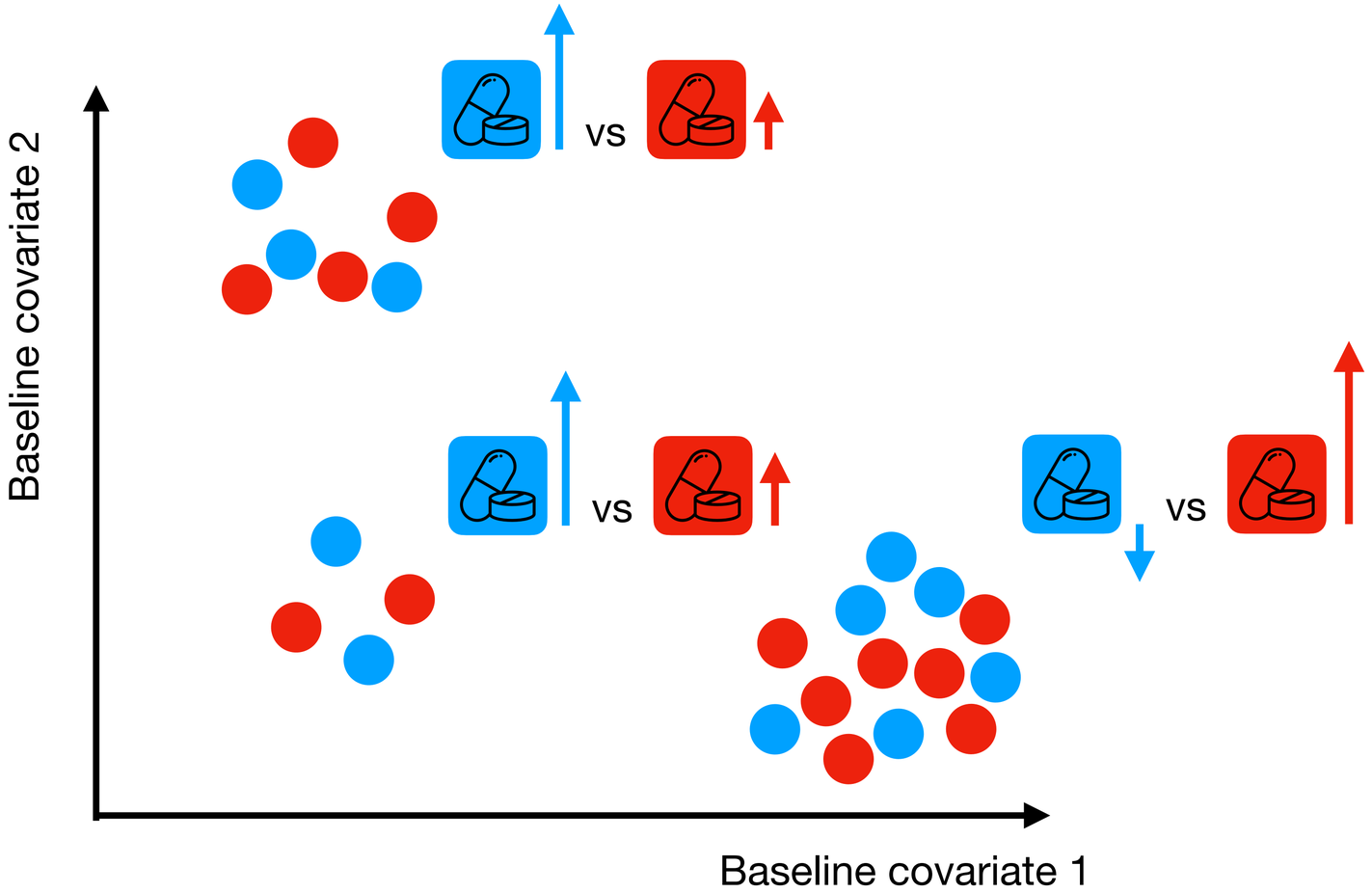
Heterogeneous Subgroups Drive Results

Sketch of a Phase II Trial



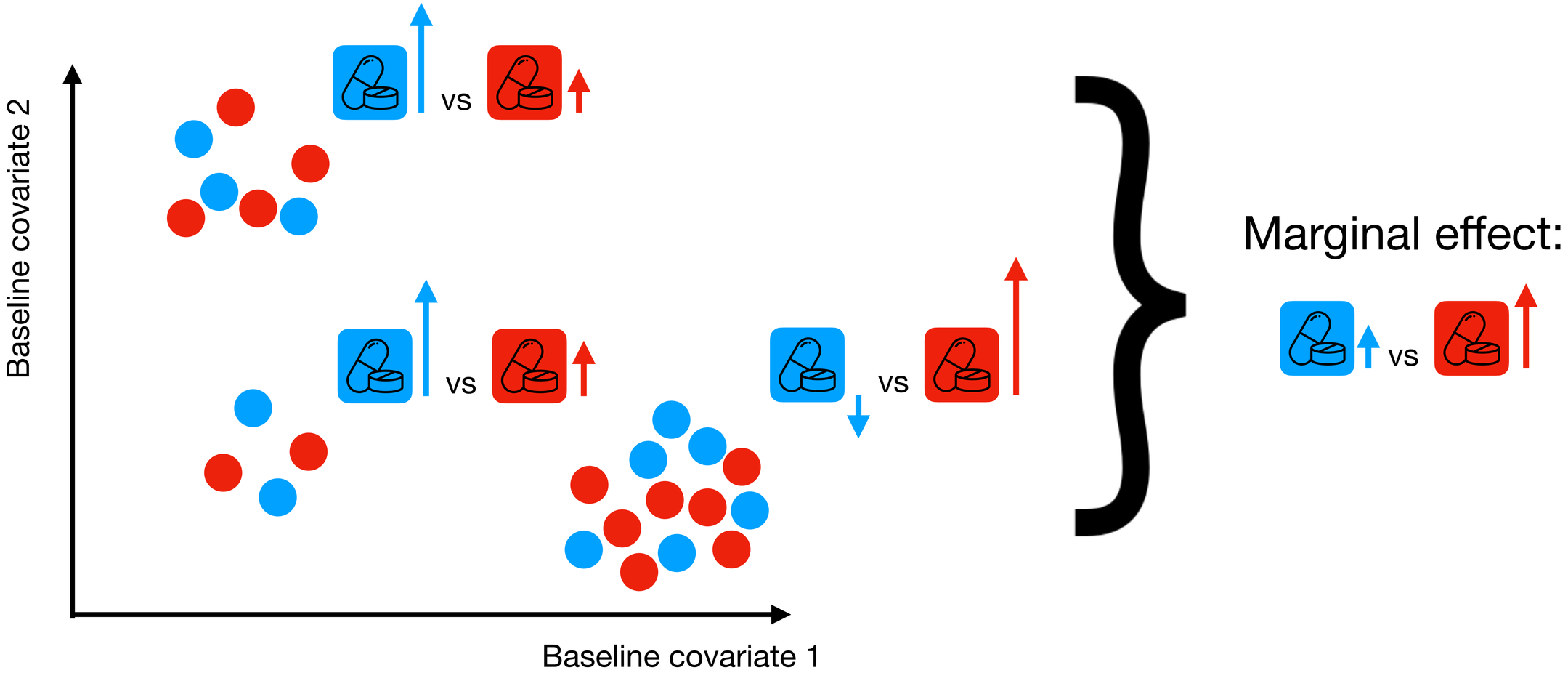
Heterogeneous Subgroups Drive Results

Sketch of *Subsequent Phase III Trial*

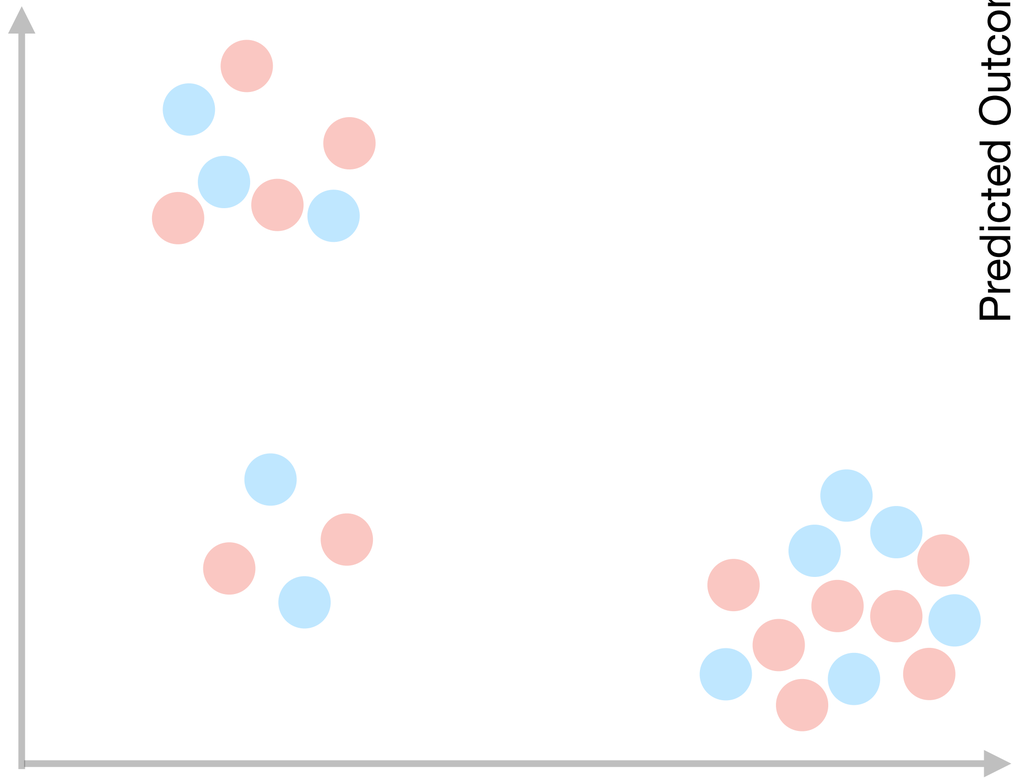


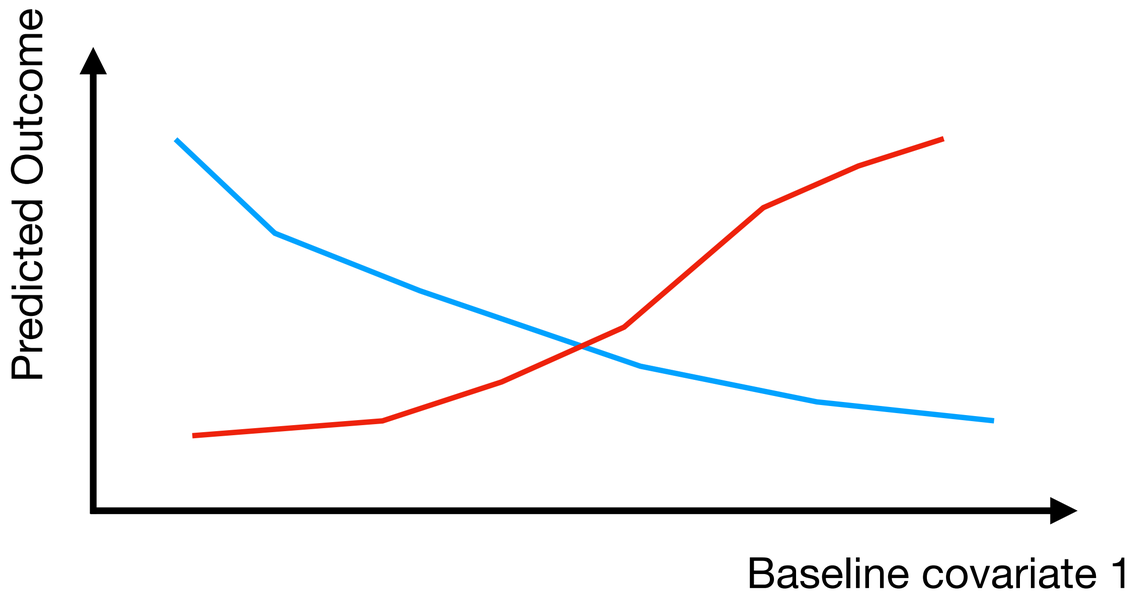
Heterogeneous Subgroups Drive Results

Sketch of *Subsequent Phase III Trial*



Treatment Effect Modifiers Define Subgroups





Baseline covariate 1 is a
treatment effect modifier,
a pre-treatment covariate that interacts
with the treatment

Why Uncover Heterogeneous Treatment Effects?

- Uncovering heterogeneous treatment effects can:
  - Mitigate therapy development risks
  - Inform treatment guidelines
  - Explain biological mechanisms of the therapy

How to Uncover Heterogeneous Treatment Effects?

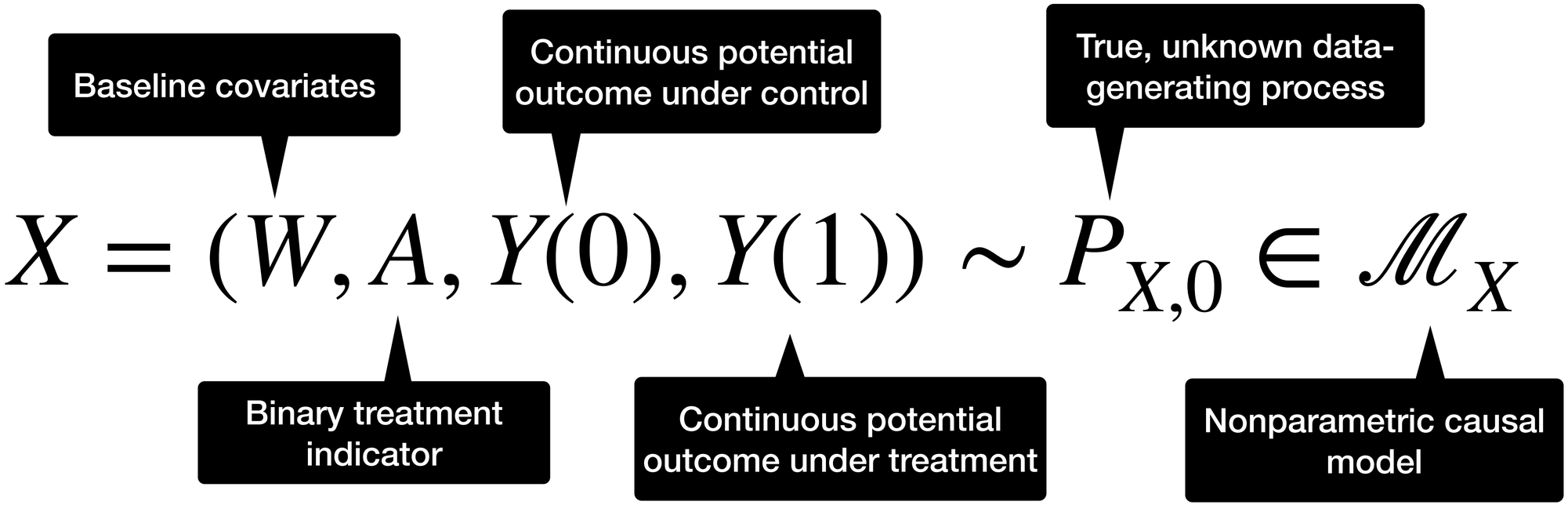
- **Uncovering heterogeneous treatment effects is challenging**
- Standard procedures rely on treatment effect modifiers being collected
- The treatment effect modifiers driving the heterogeneity might:
  - Not be known in advanced,
  - Not collected due to resource constraints, or
  - Mismeasured due to error or technological limitations
- **Can we still uncover heterogeneous treatment effects without them?**

Problem Formulation

The Complete Data-Generating Process

$$X = (W, A, Y(0), Y(1)) \sim P_{X,0} \in \mathscr{M}_X$$

The Complete Data-Generating Process



Potential Outcomes

Observation	W_1	...	W_p	A	$Y(0)$	$Y(1)$
1	0	…	12.6	1	2.5	5.1
2	0	…	40.9	0	3.7	4.2
3	1	…	1.1	0	10.3	8.7
...	…	…	…	…	…	…
n	1	…	-4.2	1	5.0	5.0

Potential Outcomes

Observation	W_1	...	W_p	A	$Y(0)$	$Y(1)$
1	0	…	12.6	1	2.5	**5.1**
2	0	…	40.9	0	**3.7**	4.2
3	1	…	1.1	0	**10.3**	8.7
...	…	…	…	…	…	…
n	1	…	-4.2	1	5.0	**5.0**

The Individual Treatment Effect Distribution and the ATE

Observation	$Y(0)$	$Y(1)$	$Y(1) - Y(0)$
1	2.5	5.1	**2.6**
2	3.7	4.2	**1.5**
3	10.3	8.7	**-1.6**
...	...	...	**...**
n	5.0	5.0	**0.0**

- We can contrast the potential outcomes to define the individual treatment effect for each observation
- The individual treatment effects are summarized to quantify the treatment effect
- E.g., average treatment effect (ATE)

$$\mathbb{E}_{P_{X,0}}[Y(1) - Y(0)]$$

The CATE

Observation	$Y(0)$	$Y(1)$	$Y(1) - Y(0)$
1	2.5	5.1	**2.6**
2	3.7	4.2	**1.5**
3	10.3	8.7	**-1.6**
...	...	...	**...**
n	5.0	5.0	**0.0**

- The ATE summarizes the average treatment effect in the population
- The conditional average treatment effect (CATE) is a causal parameter that can quantify heterogeneity as a function of treatment effect modifiers

$$\mathbb{E}_{P_{X,0}}[Y(1) - Y(0) \mid W]$$

Limitations of the CATE

- The CATE is a wildly popular parameter in the causal inference literature and is increasingly used in applications because it provides a mechanistic understanding of heterogeneity
- It has limitations, however:

  1. Treatment effect modifiers must be contained in W
  2. It is generally difficult to estimate with high-dimensional W

E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics, 17(2):3008–3049, Jan. 2023. ISSN 1935-7524, 1935-7524.
P. Boileau, N. Leng, and S. Dudoit. Guidance on individualized treatment rule estimation in high dimensions. The International Journal of Biostatistics, May 2025a. ISSN 1557-4679.
P. Boileau, N. T. Qi, M. J. van der Laan, S. Dudoit, and N. Leng. A flexible approach for predictive biomarker discovery. Biostatistics, 24(4):1085–1105, 2023. ISSN 1465-4644.
P. Boileau, N. Leng, N. S. Hejazi, M. van der Laan, and S. Dudoit. A nonparametric framework for treatment effect modifier discovery in high dimensions. Journal of the Royal Statistical Society Series B: Statistical Methodology, 87 (1):157–185, 2025b. ISSN 1369-7412.

Limitations of the CATE

- The CATE is a wildly popular parameter in the causal inference literature and is increasingly used in applications because it provides a mechanistic understanding of heterogeneity
- It has limitations, however:
    1. Treatment effect modifiers must be contained in W
    2. It is generally difficult to estimate with high-dimensional W
- **Can we define a causal estimand that avoids these limitations?**

E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics, 17(2):3008–3049, Jan. 2023. ISSN 1935-7524, 1935-7524.
P. Boileau, N. Leng, and S. Dudoit. Guidance on individualized treatment rule estimation in high dimensions. The International Journal of Biostatistics, May 2025a. ISSN 1557-4679.
P. Boileau, N. T. Qi, M. J. van der Laan, S. Dudoit, and N. Leng. A flexible approach for predictive biomarker discovery. Biostatistics, 24(4):1085–1105, 2023. ISSN 1465-4644.
P. Boileau, N. Leng, N. S. Hejazi, M. van der Laan, and S. Dudoit. A nonparametric framework for treatment effect modifier discovery in high dimensions. Journal of the Royal Statistical Society Series B: Statistical Methodology, 87 (1):157–185, 2025b. ISSN 1369-7412.

Defining a Homogeneous Treatment Effect

Observation	$Y(0)$	$Y(1)$	$Y(1) - Y(0)$
1	2.5	4.0	**1.5**
2	3.7	5.2	**1.5**
3	10.3	11.8	**1.5**
...	...	...	**...**
n	5.0	6.5	**1.5**

- The ATE only provides a complete summary of the treatment effect if the individual treatment effect is constant
- The individual treatment effect is constant when the treatment is homogeneous:

$$Y(0) + \gamma = Y(1)$$

where $\gamma \in \mathbb{R}$ is the ATE

Testing the Homogeneous Treatment Effect Hypothesis

- Can we evaluate homogeneity of the treatment effect using the constant treatment effect definition?

Testing the Homogeneous Treatment Effect Hypothesis

- Can we evaluate homogeneity of the treatment effect using the constant treatment effect definition? **Yes!**
  - Consider that $\mathbb{V}_{P_{X,0}}[Y(0) + \gamma] = \mathbb{V}_{P_{X,0}}[Y(0)] = \mathbb{V}_{P_{X,0}}[Y(1)]$
  - Then $\mathbb{V}_{P_{X,0}}[Y(1)] - \mathbb{V}_{P_{X,0}}[Y(0)] \neq 0 \implies Y(0) + \gamma \neq Y(1)$

Takeaway: We can uncover treatment effect heterogeneity through simple contrasts of potential outcome variances without needing to condition on treatment effect modifiers.

Causal Estimand: Differential Variance

Absolute Differential Variance

$$\Psi_C(P_{X,0}) = \sqrt{\mathbb{V}_{P_{X,0}}[Y(1)]} - \sqrt{\mathbb{V}_{P_{X,0}}[Y(0)]}$$

Relative Differential Variance

$$\Gamma_C(P_{X,0}) = \frac{\mathbb{V}_{P_{X,0}}[Y(1)]}{\mathbb{V}_{P_{X,0}}[Y(0)]}$$

Causal Estimand: Differential Variance

Absolute Differential Variance

$$\Psi_C(P_{X,0}) = \sqrt{\mathbb{V}_{P_{X,0}}[Y(1)]} - \sqrt{\mathbb{V}_{P_{X,0}}[Y(0)]}$$

Relative Differential Variance

$$\Gamma_C(P_{X,0}) = \frac{\mathbb{V}_{P_{X,0}}[Y(1)]}{\mathbb{V}_{P_{X,0}}[Y(0)]}$$

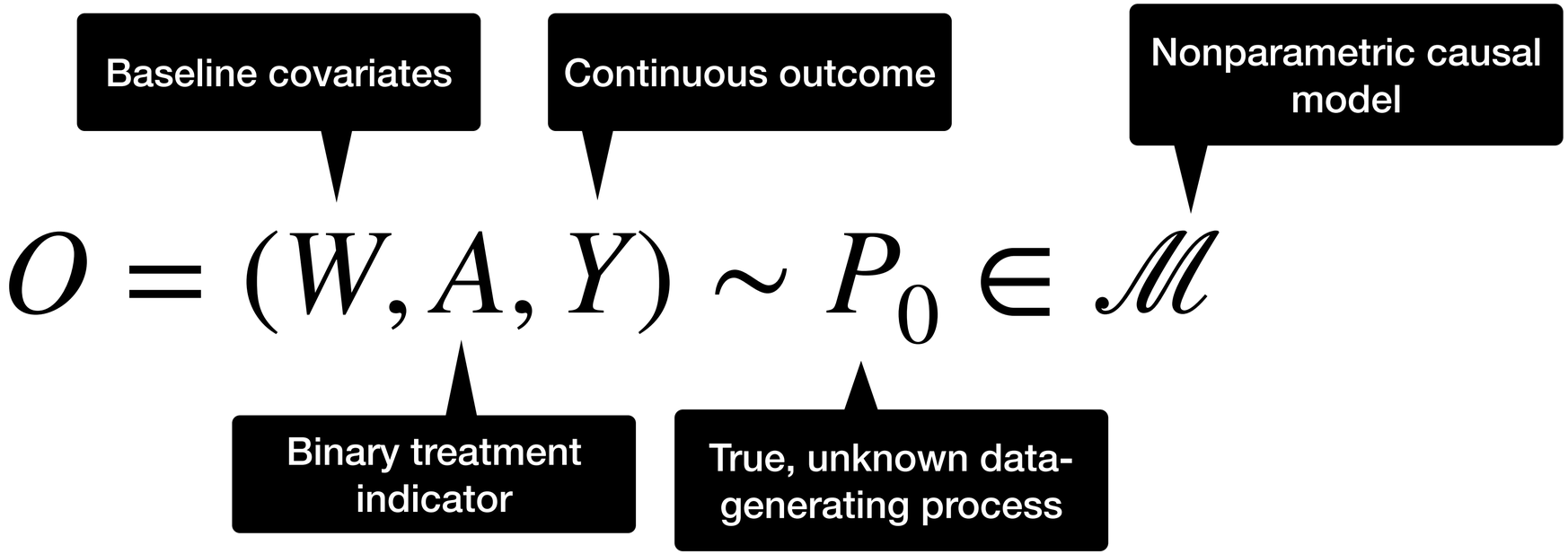
Formal testing about treatment effect heterogeneity:

$$H_0\colon \Psi_C(P_{X,0}) = 0 \qquad H_A\colon \Psi_C(P_{X,0}) \neq 0$$

or

$$H_0\colon \Gamma_C(P_{X,0}) = 1 \qquad H_A\colon \Gamma_C(P_{X,0}) \neq 1$$

The Observed Data-Generating Process



The Observed Data-Generating Process

Observation	W_1	...	W_p	A	$Y(0)$	$Y(1)$	Y
1	0	…	12.6	1	2.5	**5.1**	**5.1**
2	0	…	40.9	0	**3.7**	4.2	3.7
3	1	…	1.1	0	**10.3**	8.7	10.3
...	…	…	…	…	…	…	…
n	1	…	-4.2	1	5.0	**5.0**	**5.0**

Causal Identifiability Conditions

$\Psi(P_{X,0})$ and $\Gamma(P_{X,0})$ can be estimated from the observed data under

1. *Full exchangeability*: $A \perp\!\!\!\perp (Y(1), Y(0)) \,|\, W$
2. *Consistency*: $Y = AY(1) + (1-A)Y(0)$
3. *Positivity*: $0 < \mathbb{P}_{P_{X,0}}[A = 1 \,|\, W] < 1$ almost surely

Then

$$\begin{aligned}\Psi_C(P_{X,0}) &= \Psi(P_0) \\ &= \sqrt{\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y^2 \,|\, A=1, W]] - \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y \,|\, A=1, W]]^2} - \\ &\quad \sqrt{\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y^2 \,|\, A=0, W]] - \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y \,|\, A=0, W]]^2}\end{aligned}$$

$$\begin{aligned}\Gamma_C(P_{X,0}) &= \Gamma(P_0) \\ &= \frac{\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y^2 \,|\, A=1, W]] - \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y \,|\, A=1, W]]^2}{\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y^2 \,|\, A=0, W]] - \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y \,|\, A=0, W]]^2}\end{aligned}$$

Causal Identifiability Conditions

$\Psi(P_{X,0})$ and $\Gamma(P_{X,0})$ can be estimated from the observed data under

1. *Full exchangeability*: $A \perp\!\!\!\perp (Y(1), Y(0)) \,|\, W$
2. *Consistency*: $Y = AY(1) + (1-A)Y(0)$
3. *Positivity*: $0 < \mathbb{P}_{P_{X,0}}[A = 1 \,|\, W] < 1$ almost surely

Important! These conditions are satisfied in clinical trials

Then

$$\Psi_C(P_{X,0}) = \Psi(P_0)$$
$$= \sqrt{\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y^2 \,|\, A=1, W]] - \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y \,|\, A=1, W]]^2} -$$
$$\sqrt{\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y^2 \,|\, A=0, W]] - \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y \,|\, A=0, W]]^2}$$

$$\Gamma_C(P_{X,0}) = \Gamma(P_0)$$
$$= \frac{\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y^2 \,|\, A=1, W]] - \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y \,|\, A=1, W]]^2}{\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y^2 \,|\, A=0, W]] - \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[Y \,|\, A=0, W]]^2}$$

Statistical Inference

Inference Based on the Efficient-Influence Function

Define target parameter

$$\Theta(P_0)$$

→

Compute efficient influence function in $\mathscr{M}$

$$D^{\Theta}(P_0)(O)$$

→

Derive nonparametric estimators: one-step, D/DML, TMLE

$$\hat{\Theta}_j(\hat{P}_n)$$

→

Establish asymptotic properties

$$\sqrt{n}\left(\hat{\Theta}(\hat{P}_n) - \Theta(P_0)\right) = \sqrt{n}\mathbb{E}_{P_n}\left[D^{\Theta}(P_0)(O)\right] + o_P(1)$$

M. J. van der Laan and S. Rose. Targeted Learning: Causal Inference for Observational and Experimental Data. S pringer Series in Statistics. Springer, New York, NY, 2011. ISBN 978-1-4419-9781-4 978-1-4419-9782-1.
M. J. van der Laan and S. Rose. Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies. Springer Series in Statistics. Springer International Publishing, Cham, 2018. ISBN 978-3-319-65303-7 978-3-319-65304-4.
O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. The American Statistician, 76(3):292–304, July 2022. ISSN 0003-1305.
E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review. In Handbook of Statistical Methods for Precision Medicine. Chapman and Hall/CRC, 2024. ISBN 978-1-003-21622-3.

Inference Based on the Efficient-Influence Function

Define target parameter

$\Theta(P_0)$

→

Compute efficient influence function in $\mathscr{M}$

$D^{\Theta}(P_0)(O)$

→

Derive nonparametric estimators: one-step, D/DML, TMLE

$\hat{\Theta}_j(\hat{P}_n)$

Flexible estimation of nuisance parameters using machine learning methods

→

Establish asymptotic properties

$$\sqrt{n}\left(\hat{\Theta}(\hat{P}_n) - \Theta(P_0)\right) = \sqrt{n}\mathbb{E}_{P_n}\left[D^{\Theta}(P_0)(O)\right] + o_P(1)$$

M. J. van der Laan and S. Rose. Targeted Learning: Causal Inference for Observational and Experimental Data. S pringer Series in Statistics. Springer, New York, NY, 2011. ISBN 978-1-4419-9781-4 978-1-4419-9782-1.
M. J. van der Laan and S. Rose. Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies. Springer Series in Statistics. Springer International Publishing, Cham, 2018. ISBN 978-3-319-65303-7 978-3-319-65304-4.
O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. The American Statistician, 76(3):292–304, July 2022. ISSN 0003-1305.
E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review. In Handbook of Statistical Methods for Precision Medicine. Chapman and Hall/CRC, 2024. ISBN 978-1-003-21622-3.

Causal Machine Learning Estimators

We derived four estimators for each of $\Psi(P_0)$ and $\Gamma(P_0)$:

- Targeted maximum likelihood estimator (TMLE)
- One-step estimator
- Cross-fitted TMLE
- Cross-fitted one-step estimator

Nuisance Parameter Estimation

- These causal machine learning estimators generally require estimation of the following nuisance parameters

    1. *Propensity score*: $\mathbb{P}_{P_0}[A = 1 \mid W]$
    2. *Outcome regression*: $\mathbb{E}_{P_0}[Y \mid W, A]$
    3. *Squared outcome regression*: $\mathbb{E}_{P_0}[Y^2 \mid W, A]$

- They can be estimated with parametric or flexible machine learning estimators
- We encourage using Super Learner ensembles

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super Learner. Statistical Applications in Genetics and Molecular Biology, 6(1), Sept. 2007. ISSN 1544-6115.

Asymptotic Properties: Double Robustness

These causal machine learning estimators are consistent if:

1. The propensity score estimator is consistent, or
2. The outcome and squared outcome regressions are consistent

Takeaway: The propensity score is known in clinical trials. These estimators are consistent even when the outcome and squared outcome regressions are misspecified.

Asymptotic Properties: Asymptotic Linearity

These causal machine learning estimators are asymptotically linear if*:

1. The propensity score estimator converges at $o_P(n^{-1/4})$,
2. The outcome regression estimator converges at $o_P(n^{-1/4})$, and
3. The squared outcome regression estimator converges at $o_P(n^{-1/4})$,

*or *the propensity score is known.*

Takeaway: The propensity score is known in clinical trials. These estimators are asymptotically normal with known variance even when the outcome and squared outcome regressions are misspecified. We can construct Wald-type confidence intervals and perform formal hypothesis tests.

Simulation Studies

Simulation Study Overview

- We performed three simulation studies:
    1. Assessment of double robustness
    2. Assessment of asymptotic linearity
    3. Assessment of measurement error and comparison to a CATE-based hypothesis test for detecting treatment effect heterogeneity
- Only the third simulation study is presented here.

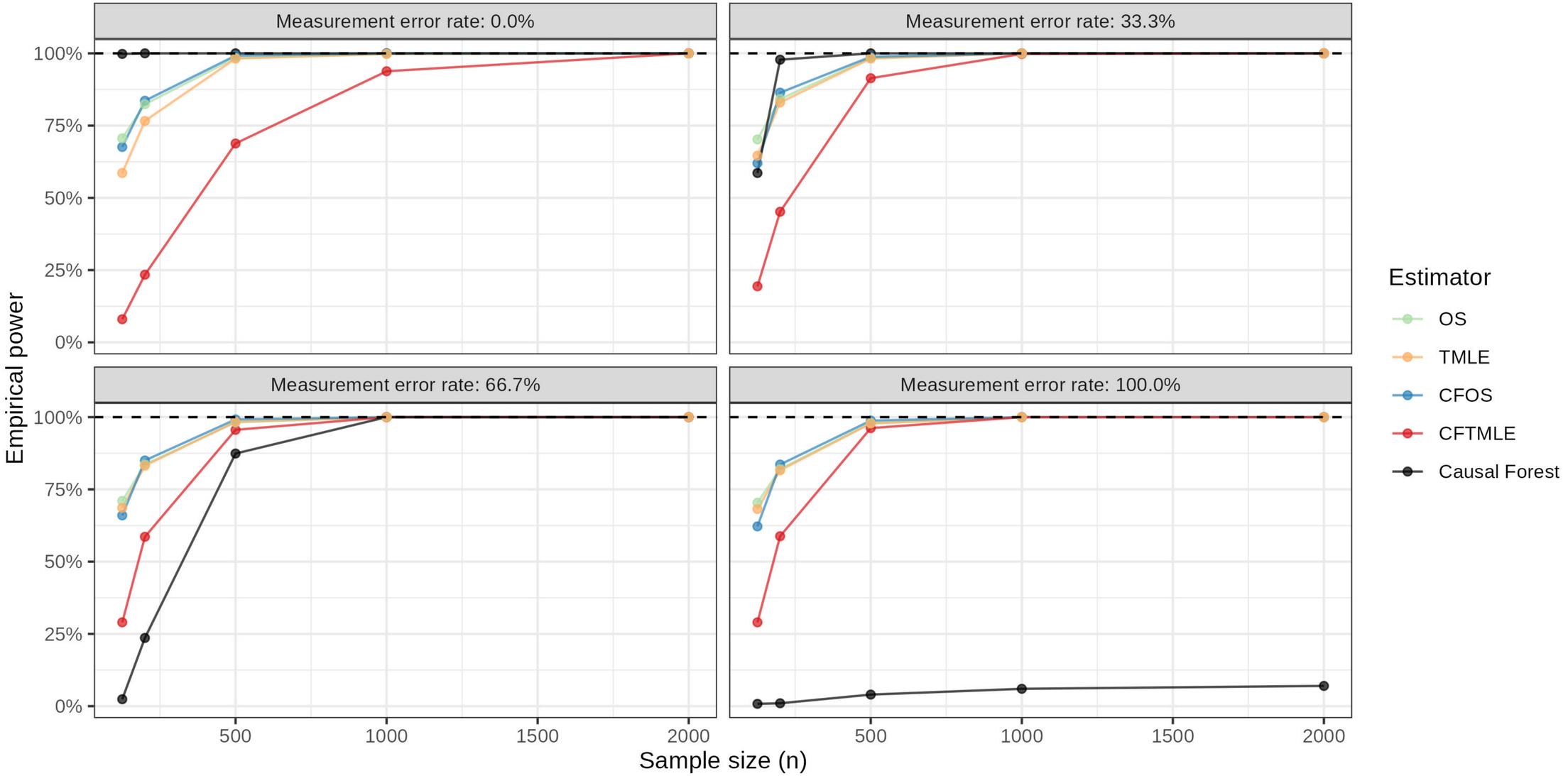
S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using Random Forests. Journal of the American Statistical Association, 113(523):1228–1242, July 2018. ISSN 0162-1459.
V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernández-Val. Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. Technical Report w24678, National Bureau of Economic Research, Cambridge, MA, June 2018.

Setup

- We simulate randomized controlled trial datasets with one treatment effect modifier
- The treatment effect modifier is mismeasured, and only the mismeasured effect modifier is observed
- Can our causal machine learning estimators detect treatment effect heterogeneity?

$$W_1 \sim N(0,1)$$
$$W_2 \sim N(0,1)$$
$$V \sim \text{Bern}(0.5)$$
$$M \sim \text{Bern}(m)$$
$$U \sim \text{Bern}(0.2)$$
$$V^{\text{obs}} = \begin{cases} U \text{ if } M = 1 \\ V \text{ if } M = 0 \end{cases}$$
$$A \mid W, V = A \sim \text{Bern}(0.5)$$
$$Y \mid W, V, A \sim N(1 - 2A + W_1^2 + 2I(W_2 < 0) + 4AV, 1)$$

Absolute Differential Variance Results



Application

Targeted Temperature Management Trials

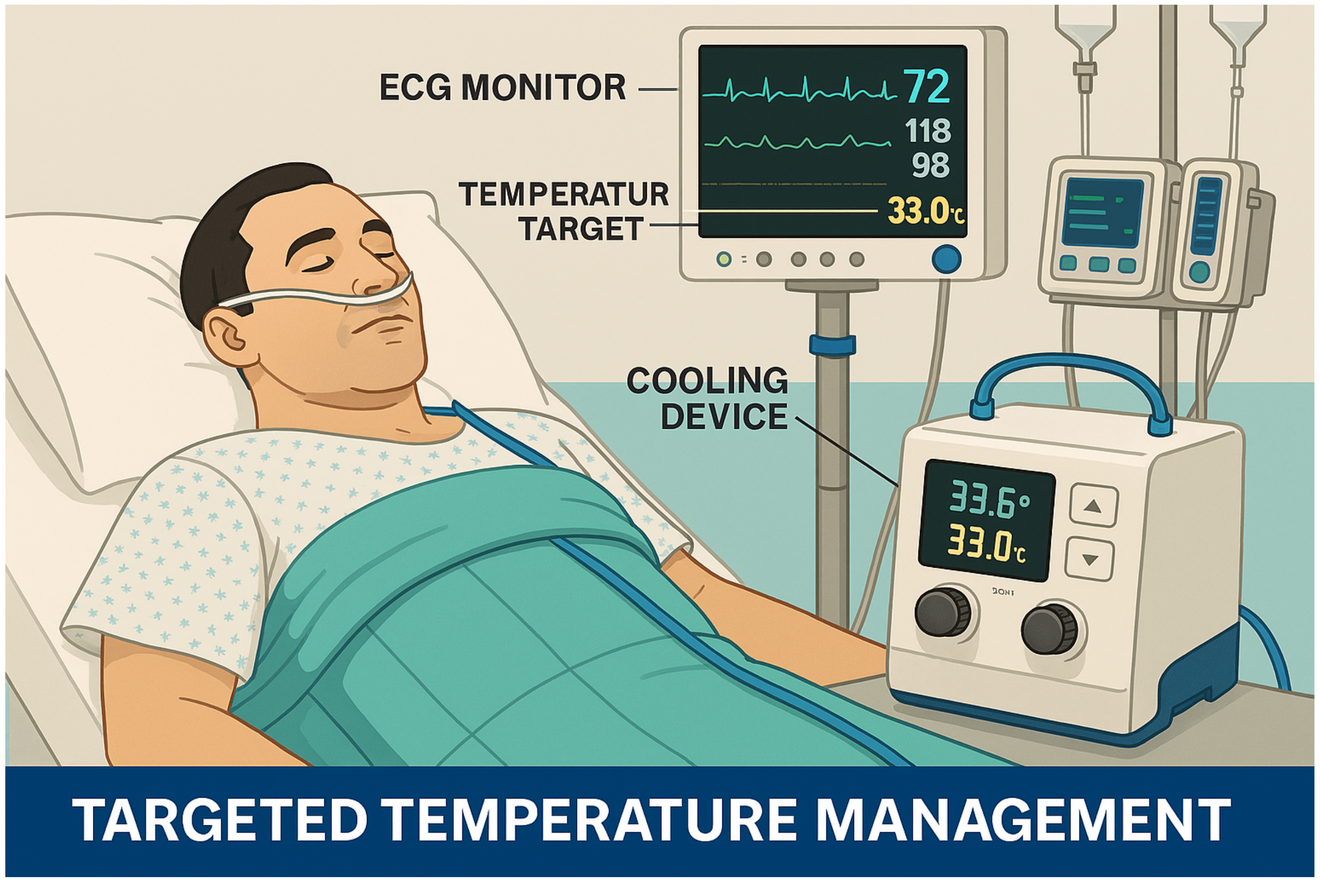


Image generated by co-pilot, please excuse typos.

- Targeted temperature management induces mild, controlled hypothermia
- Cooling generally reduces neurological damage after a loss and return of blood flow to the brain
- There’s uncertainty regarding the optimal target temperature in out-of-hospital cardiac arrest patients

Targeted Temperature Management Trials

- Two trials were recently conducted to investigate the choice of temperature
  1. The Target Temperature Management at 33°C versus 36°C after Cardiac Arrest (TTM; NCT01020916) (n = 950)
  2. Hypothermia versus Normothermia after Out-of-hospital Cardiac Arrest (TTM2; NCT02908308) (n = 1900)
- TTM and TTM2 trials' primary outcome was six-month overall survival
- Neither trial found a statistically significant difference in the targeted temperature management interventions

N. Nielsen, et al. Targeted Temperature Management at 33○C versus 36○C after Cardiac Arrest. New England Journal of Medicine, 369(23):2197–2206, Dec. 2013. ISSN 0028-4793.
J. Dankiewicz, et al. Hypothermia versus Normothermia after Out-of-Hospital Cardiac Arrest. New England Journal of Medicine, 384(24):2283–2294, June 2021. ISSN 0028-4793.

Attempts at Uncovering Heterogeneous Treatment Effects

- Some critical care physicians hypothesize that low-temperature hypothermia provides enhanced benefits to patients with signs of severe neurological injury or prolonged loss of blood flow
- Recent, in-progress work attempting to detect this hypothesized treatment effect heterogeneity using several CATE-based approaches has been unsuccessful
- This is possibly attributed to treatment effect modifiers being approximated (e.g., time to advanced life support) or missing from the data

J. D. Mark, J. L. Lopez, W. Wahood, R. A. Colombo, M. Danckers, A. A. Damluji, J. N. Katz, and C. L. Alviar. Personalizing Temperature Targets After Cardiac Arrest. JACC: Advances, 4(10):102042, Aug. 2025. ISSN 2772-963X.
G. Lileikyte,et al. Detecting Heterogeneous Treatment Effects in the TTM and TTM2 Trials. In preparation.

Absolute Differential Variance Inference

- We treated the primary endpoint, days of survival from intervention initiation to six months, as a continuous outcome since there was virtually no right-censoring.
- We inferred the absolute differential variance using the one-step estimator and TMLE to test the homogeneous treatment effect hypothesis
- We adjusted for several baseline clinical covariates by estimating the outcome and squared outcome regressions using linear models, and used the known propensity scores

Absolute Differential Variance Inference

Trial	Estimator	Estimate (days)	Standard Error	P-Value
TTM	One-step	3.9	7.1	0.58
TTM	TMLE	3.6	7.1	0.61
TTM2	One-step	0.9	0.3	<0.01
TTM2	TMLE	0.9	0.3	<0.01

- We fail to reject the null of a homogeneous treatment effect in TTM
- We reject the the null of homogeneous treatment effect in TTM2
  - The absolute differential variance estimate is *less than a day*, suggesting that a *heterogeneous treatment effect*, if present, is not *clinically meaningful*

Discussion

Causal Machine Learning for Differential Variance Inference

- We developed new causal parameters for detecting heterogeneous treatment effects that do not rely on access to treatment effect modifiers
- We provided general causal identifiability conditions for higher-order moments of potential outcomes
- We derived causal machine learning estimators to perform inference about these parameters in randomized experiments and observational studies
- We established conditions for these estimators' desirable asymptotic properties: double robustness and asymptotic linearity
- We showed that inference about these parameters complements standard CATE-based approaches for uncovering heterogeneous treatment effects

Next Steps and Extensions

- **Methodology**: Develop a sequential heterogeneous treatment effect inference workflow for applying differential variance followed by CATE-based estimators while accounting for post-selection inference
- **Application**: Build power analysis tools to guide clinical trial design
- **Application**: Apply differential variance procedures to observational study data
- **Methodology**: Extend differential variance parameters for other study designs (e.g., longitudinal) and outcome types (e.g., time-to-event)

McGill

Centre universitaire
de santé McGill
Institut de recherche

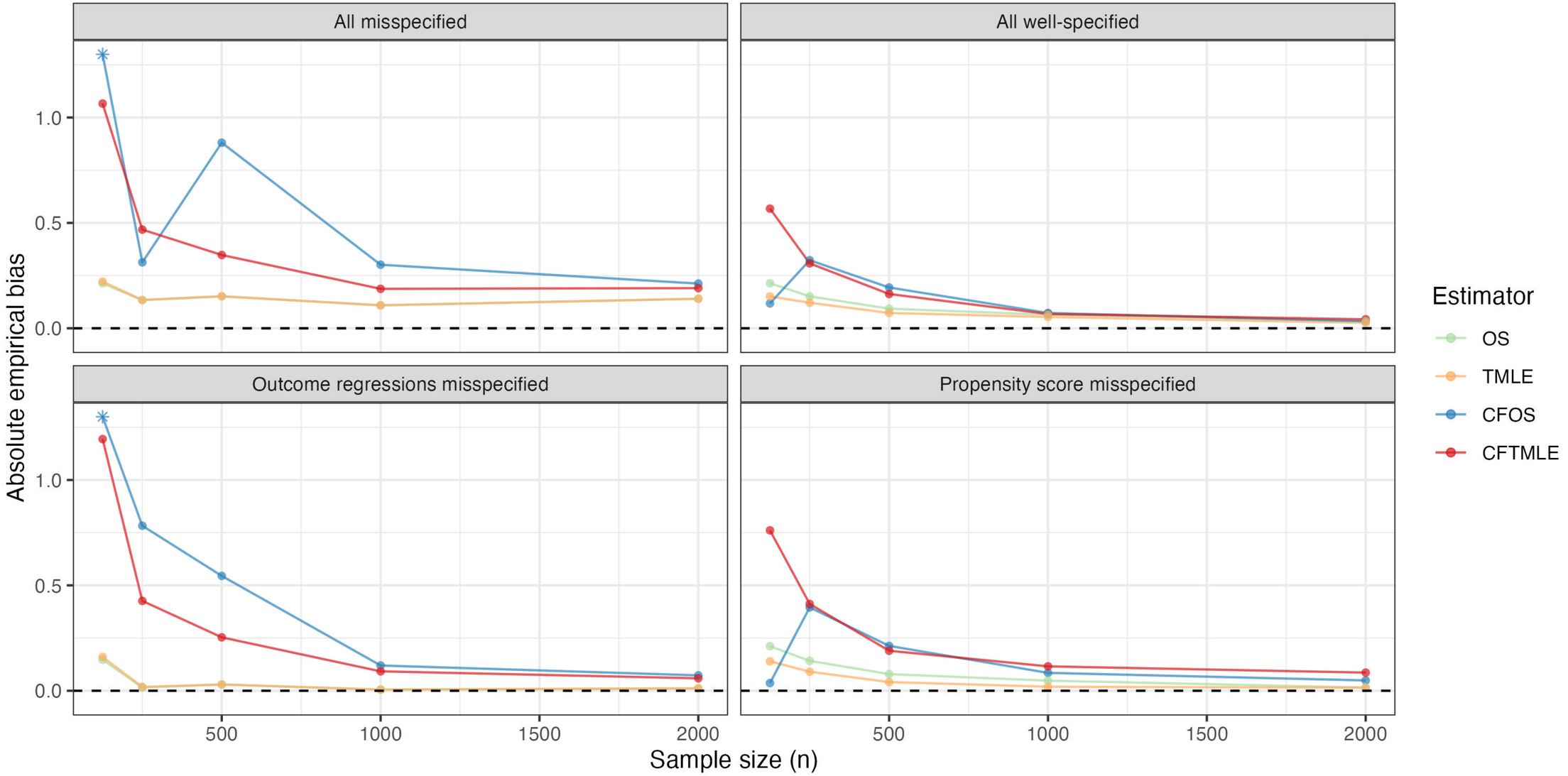
McGill University
Health Centre
Research Institute

Merci!

Contact info: philippe.boileau@mcgill.ca
Publications and software: pboileau.ca

Appendix

Simulation Results: Double Robustness



Simulation Results: Asymptotic Linearity