Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Causal machine learning for predicting treatment outcomes

Stefan Feuerriegel[1,2] ✉, Dennis Frauen[1,2], Valentyn Melnychuk[1,2], Jonas Schweisthal[1,2], Konstantin Hess[1,2], Alicia Curth[3], Stefan Bauer[4,5], Niki Kilbertus[2,4,5], Isaac S. Kohane[6] & Mihaela van der Schaar[7,8]

Tanawat Wuttiyakorn
M.Sc. Student
Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol university

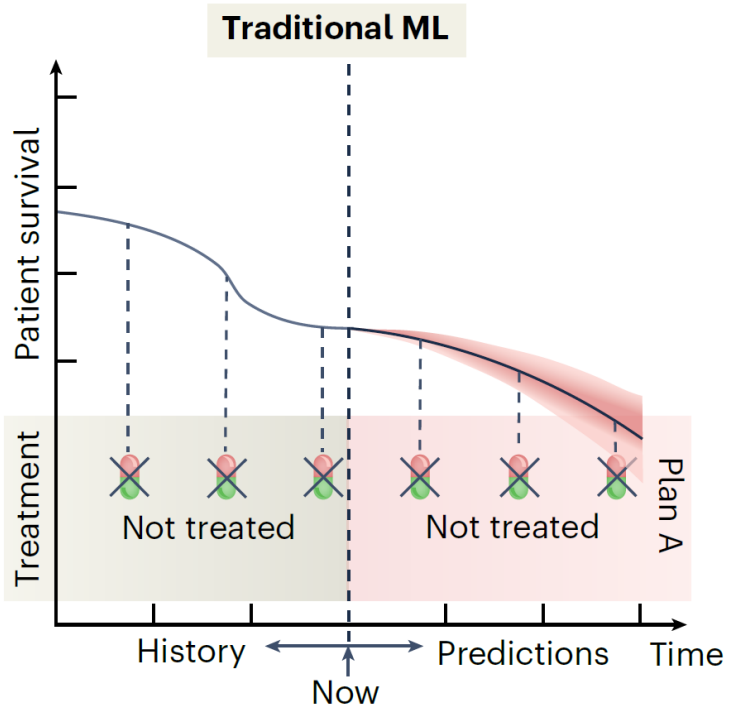
Traditional predictive ML

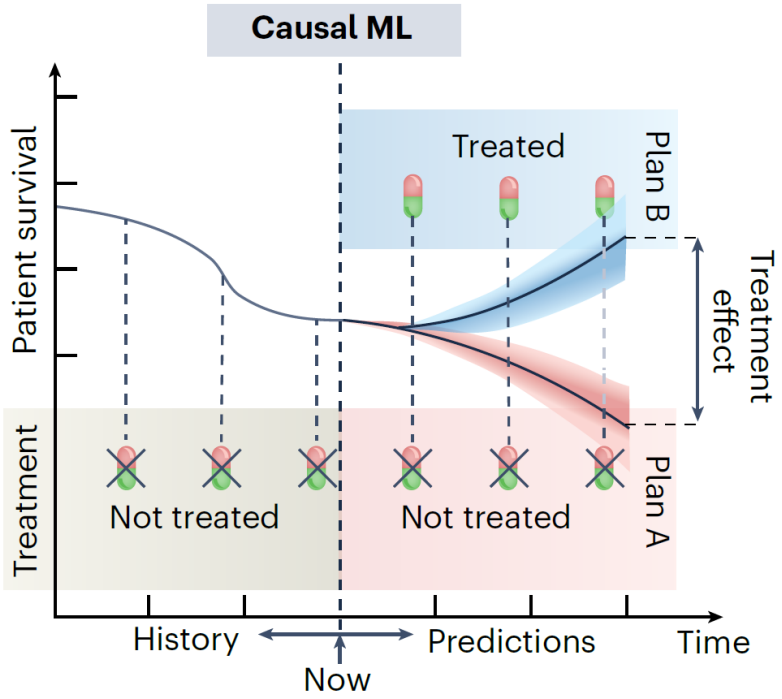
Aim at predicting outcome

Vs.

Causal ML

Quantify changes in outcome due to treatment
Aim to answer ‘what if’ questions


Traditional ML
Patient survival
Treatment
Not treated
Not treated
Plan A
History
Predictions
Time
Now


Causal ML
Patient survival
Treated
Plan B
Treatment effect
Treatment
Not treated
Not treated
Plan A
History
Predictions
Time
Now

Causal ML vs. Traditional statistics

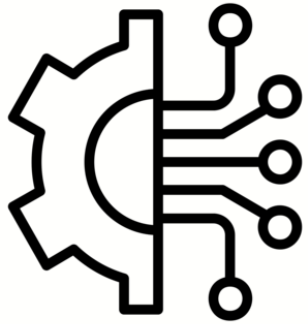
Traditional statistics

- Often assume knowledge about the parametric form of the association between treatment and the outcome
- Often use the simple model, such as linear regression
- Often preferred for small sample size
- However, such knowledge is often not available or unrealistic, especially for high dimensional datasets

Causal ML vs. Traditional statistics

Causal ML

- Allow for less rigid models, non-parametric models
- Can capture complex disease dynamic
- Non-linear model can be used to capture heterogeneity in treatment effect
- Require larger sample sizes



Fundamental problem of causal inference with RWD

- Can only observe the factual outcome,
  but never observes the counterfactual outcome

b

Traditional ML

Data

Patient	Covariates	Treatment	Patient outcome
1	Age, sex, etc.	0	–1.0
2	↓	1	2.3
3	↓	1	0.3

Task

Patient	Covariates	Treatment	Patient outcome
1	Age, sex, etc.	1	?
2	↓	0	?

Causal ML

Data

Patient	Covariates	Treatment	Patient outcome	
			If not treated	If treated
1	Age, sex, etc.	0	–1.0	
2	↓	1		2.3
3	↓	1		0.3

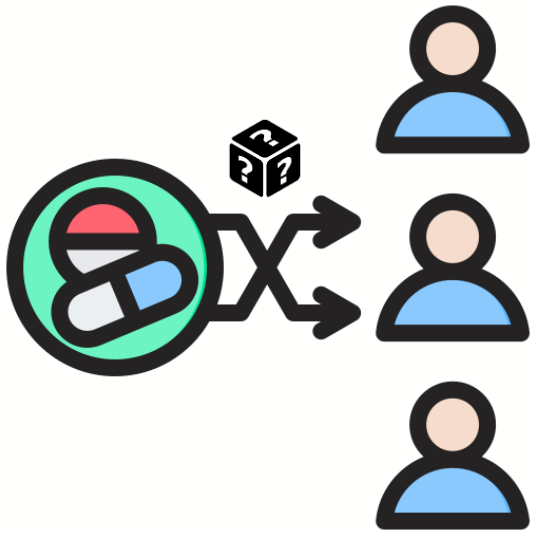
Task

Patient	Covariates	Potential outcomes		Treatment effect
		If not treated	If treated	If treated → If not treated
1	Age, sex, etc.	?	?	?
2	↓	?	?	?

☐ Missing observations ? Prediction targets

Fundamental problem of causal inference with RWD

- Treatment assignment is not fully randomized
- Treatment assignment depends on covariates
- The Assumptions MUST be made



Assumptions

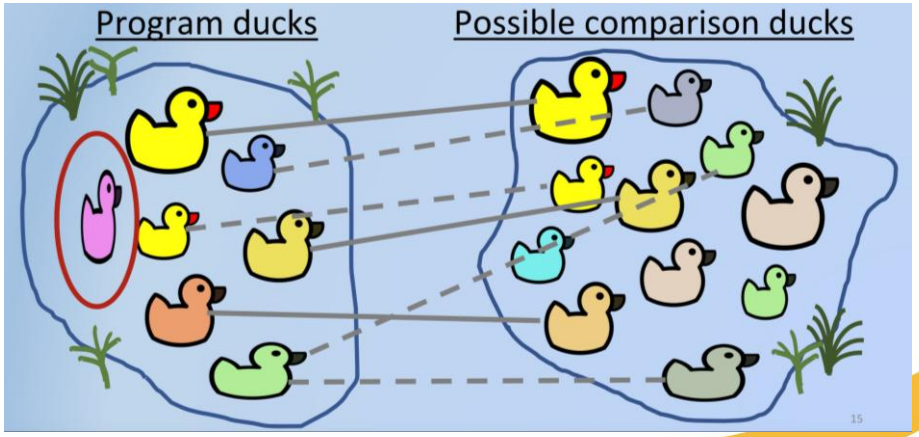
1. Stable unit treatment value assumption (SUTVA)
2. Positivity (Overlap)
3. Ignorability (Unconfoundness)

Causal inference methods

Propensity Score Matching

- Matches treated and untreated units based on their propensity score (the probability of receiving the treatment given covariates)
- Reduces bias by ensuring comparable treatment and control groups

- Matching can introduce bias, especially in high-dimensional data, where units are more likely to be far apart
- Unit without matches will be excluded, reducing the data size

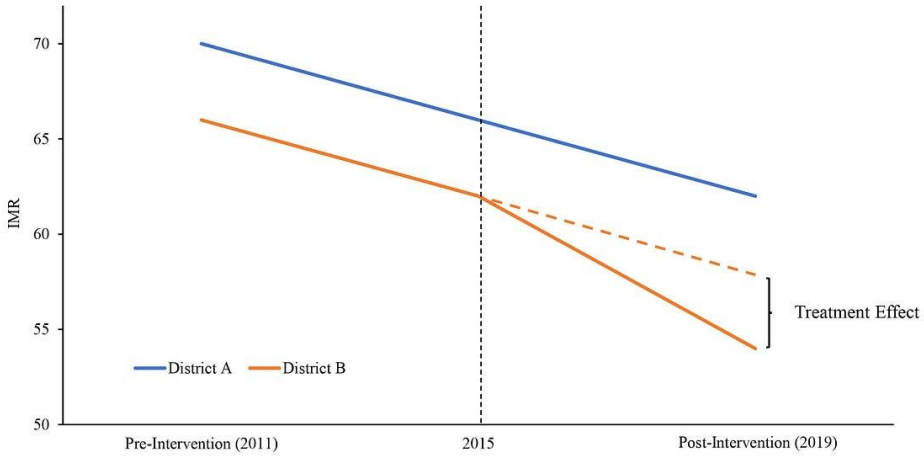


Causal inference methods

Difference-in-Difference

- Compares the change in outcomes over time between a treatment group and a control group.
- Controls for time-invariant confounding

- Assumes parallel trends between groups before the intervention



Causal inference methods

Causal Forest

- Adapt the decision tree and random forest to estimate heterogeneous treatment effects
- Captures heterogeneity in treatment effects

- Computationally intensive and requires sufficient sample size

Causal inference methods

G-computation

- Modeling the relationship between covariates, treatment, and the outcome

$$f(T, X; \theta) = \beta_0 + \beta_1 T + \beta_2 X_1 + \beta_3 X_2$$

- Once the outcome model is trained, it is used to predict the counterfactual

$$\hat{Y}(T = 1) = f(T = 1, X; \hat{\theta})$$

$$\hat{Y}(T = 0) = f(T = 0, X; \hat{\theta})$$

- then compute the average treatment effect by taking the difference between the average predicted outcomes under treatment and control

$$\text{ATE} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{Y}_i(T = 1) - \hat{Y}_i(T = 0) \right]$$

Causal inference methods

Double Machine Learning (DML)

- Aims to obtain an unbiased estimate by using flexible machine learning to adjust for observed confounders
- Originates from the Frisch-Waugh-Lovell theorem, to isolate causal effects by controlling for covariates

Key Principles of DML

- Orthogonalization

(making the causal parameter orthogonal (or uncorrelated) to the nuisance function)

- Cross prediction
- Flexible ML Models

Frisch-Waugh-Lovell theorem

$$y_i = \beta_0 + \kappa T_i + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i$$

Frisch-Waugh-Lovell theorem

$$y_i = \beta_0 + \kappa T_i + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i$$

X_1 Treatment

X_2 Covariates

Frisch-Waugh-Lovell theorem

$$y_i = \beta_0 + \kappa T_i + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i$$

X_1 Treatment

X_2 Covariates

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Source: Facure M (2023) Causal Inference in Python. O'Reilly Media

Frisch-Waugh-Lovell theorem

$$\hat{y^*} = \hat{\gamma}_1 X_2$$

$$\hat{X}_1 = \hat{\gamma}_2 X_2$$

Frisch-Waugh-Lovell theorem

$$\hat{y}^* = \hat{\gamma}_1 X_2$$

$$\tilde{y}_1 = y - \hat{y}^*$$

Outcome Residual

$$\hat{X}_1 = \hat{\gamma}_2 X_2$$

$$\tilde{X}_1 = X_1 - \hat{X}_1$$

Treatment Residual

Orthogonalization !!

Frisch-Waugh-Lovell theorem

$$\hat{y}^* = \hat{\gamma}_1 X_2$$

$$\hat{X}_1 = \hat{\gamma}_2 X_2$$

Outcome Residual

$$\tilde{y}_1 = y - \hat{y}^*$$

Treatment Residual

$$\tilde{X}_1 = X_1 - \hat{X}_1$$

$$\tilde{y} = \hat{\beta}_1 \tilde{X}_1$$

Source: Facure M (2023) Causal Inference in Python. O'Reilly Media

Frisch-Waugh-Lovell theorem

Outcome Residual Treatment Residual

$$(Y - (Y \sim X)) \sim (T - (T \sim X))$$

$$Y_i - E[Y_i|X_i] = \tau \cdot (T_i - E[T_i|X_i]) + \epsilon$$

Frisch-Waugh-Lovell theorem

Outcome Residual Treatment Residual

$$(Y - (Y \sim X)) \sim (T - (T \sim X))$$

Outcome Residual ATE Treatment Residual

$$Y_i - E[Y_i|X_i] = \tau \cdot (T_i - E[T_i|X_i]) + \epsilon$$

- Frisch-Waugh-Lovell theorem

$$Y_i - E[Y_i|X_i] = \tau \cdot (T_i - E[T_i|X_i]) + \epsilon$$

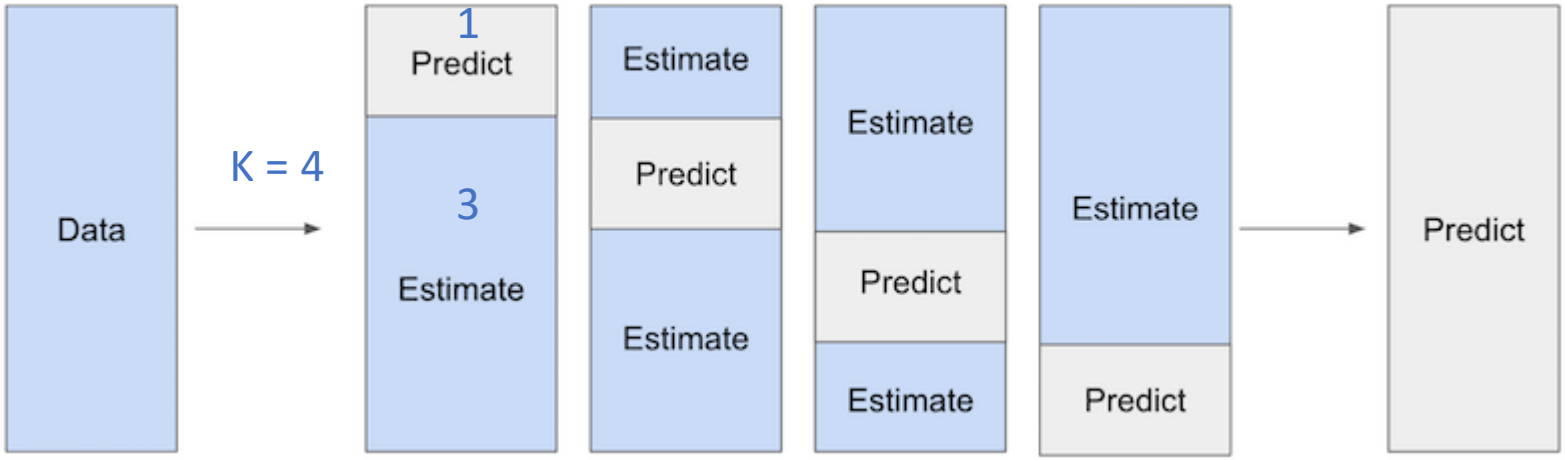
- Double Machine Learning

$$Y_i - \hat{M}_y(X_i) = \tau \cdot (T_i - \hat{M}_t(X_i)) + \epsilon$$

Outcome model

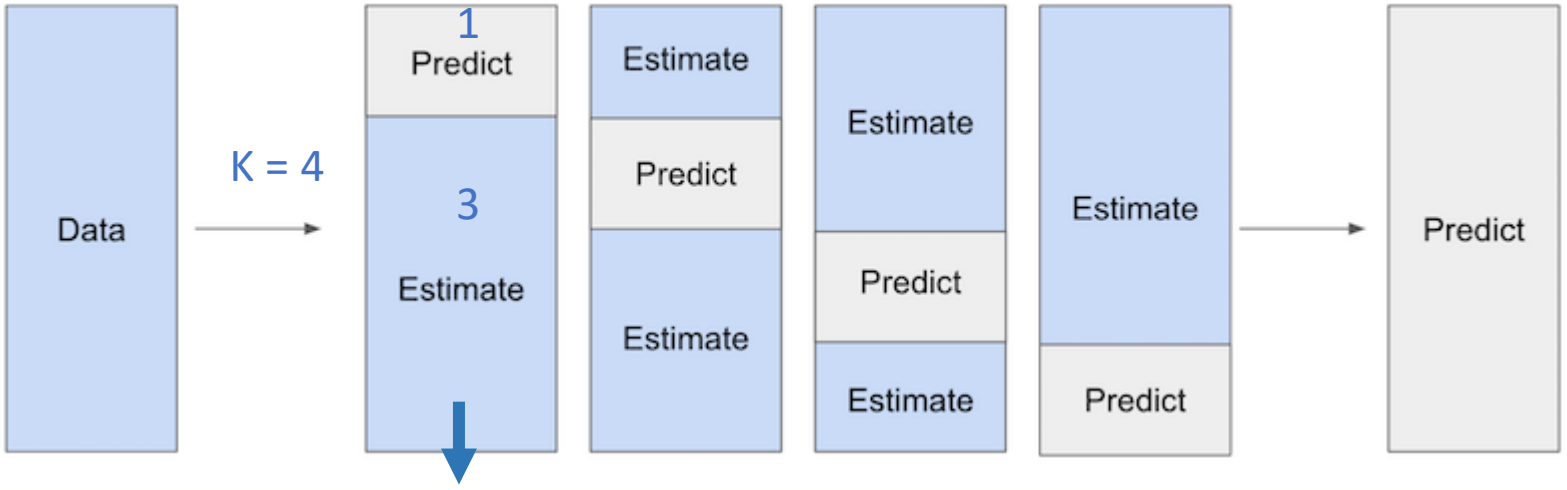
Treatment model

Cross prediction and out-of-fold residuals



Source: Facure M (2023) Causal Inference in Python. O'Reilly Media

Cross prediction and out-of-fold residuals

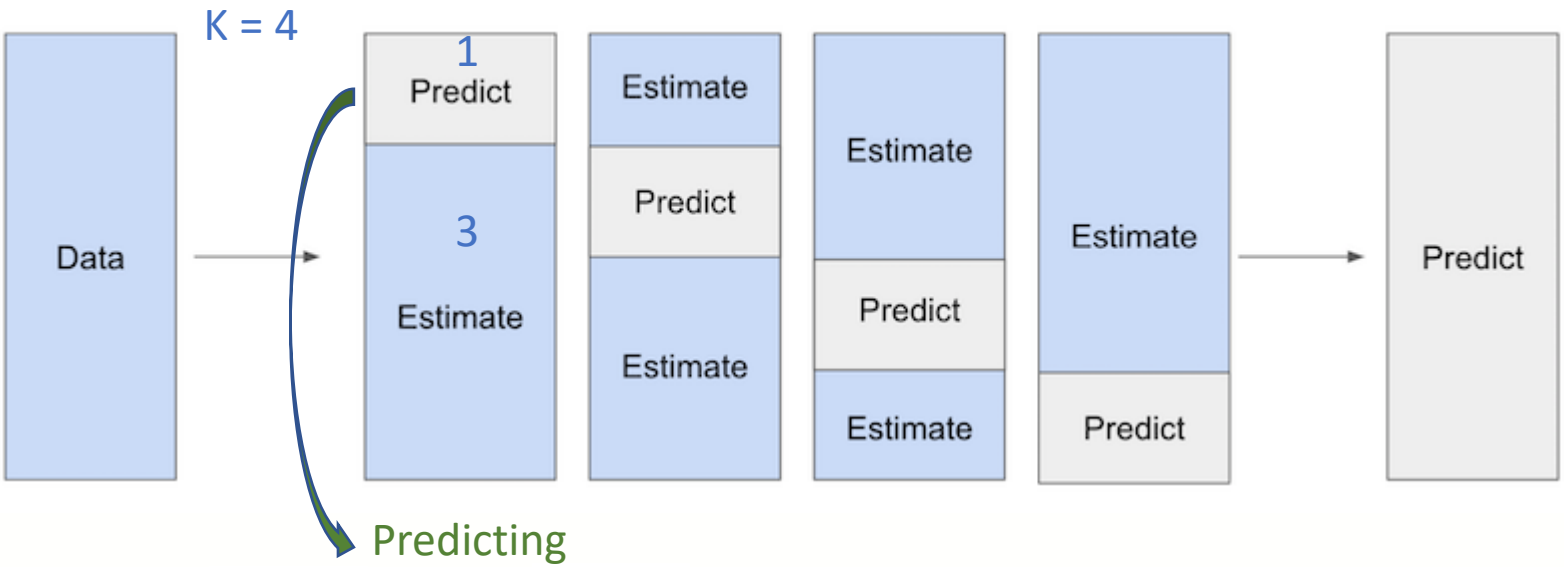


$$\hat{M}_y(X_i) \qquad \hat{M}_t(X_i)$$

$$\hat{y^*} = \hat{\gamma}_1 X_2 \qquad \hat{X}_1 = \hat{\gamma}_2 X_2$$

Source: Facure M (2023) Causal Inference in Python. O'Reilly Media

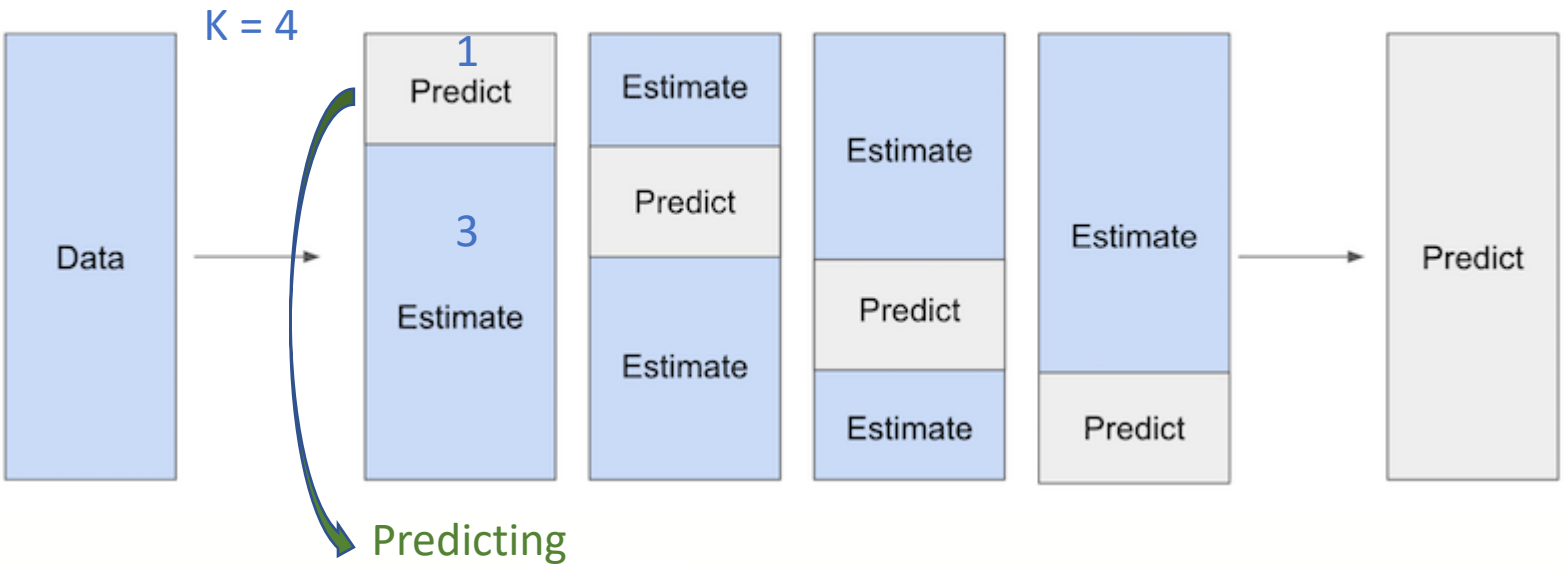
Cross prediction and out-of-fold residuals



Wisdom of the Land

Source: Facure M (2023) Causal Inference in Python. O'Reilly Media

Cross prediction and out-of-fold residuals



$$\hat{M}_y(X_i) \qquad \hat{M}_t(X_i)$$

$$\hat{y}^* = \hat{\gamma}_1 X_2 \qquad \hat{X}_1 = \hat{\gamma}_2 X_2$$

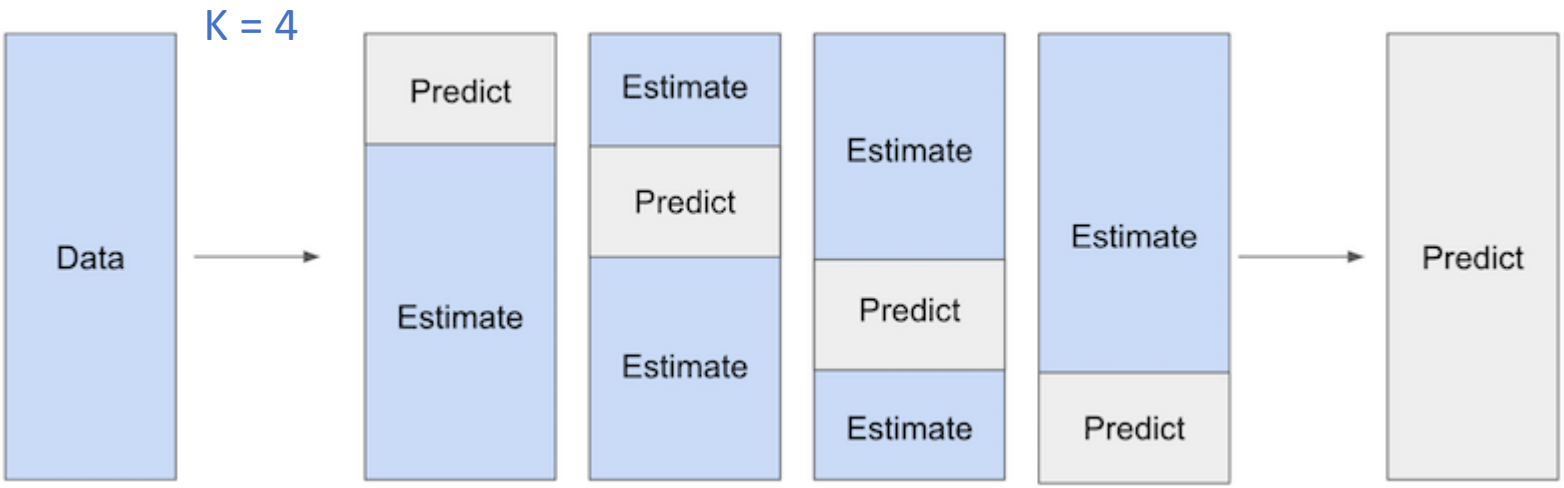
$$\tilde{y}_1 = y - \hat{y}^* \qquad \tilde{X}_1 = X_1 - \hat{X}_1$$

Outcome Residual

Treatment Residual

Source: Facure M (2023) Causal Inference in Python. O'Reilly Media

Cross prediction and out-of-fold residuals



$$\hat{M}_y(X_i) \qquad \hat{M}_t(X_i)$$

$$\hat{y}^* = \hat{\gamma}_1 X_2 \qquad \hat{X}_1 = \hat{\gamma}_2 X_2$$

$$\tilde{y}_1 = y - \hat{y}^* \qquad \tilde{X}_1 = X_1 - \hat{X}_1$$

Outcome Residual

Treatment Residual

K Folds

Outcome Residuals

Treatment Residuals

Wisdom of the Land

Source: Facure M (2023) Causal Inference in Python. O'Reilly Media

Cross prediction and out-of-fold residuals

Outcome Residual

Treatment Residual

$$\tilde{y} = \hat{\beta}_1 \tilde{X}_1$$

ATE

Source: Facure M (2023) Causal Inference in Python. O'Reilly Media

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

RESEARCH ARTICLE Open Access

External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning

Nicolas Loiseau*†, Paul Trichelair†, Maxime He, Mathieu Andreux, Mikhail Zaslavskiy, Gilles Wainrib and Michael G. B. Blum

- External control arm analyses
- Objective is to compare different statistical approaches for estimating the average treatment effect
- Dataset: Synthetic Simulations and Internal Replication Study
- Methods: Four statistical methods were compared
  - Propensity Score Matching (PSM)
  - Inverse Probability of Treatment Weighting (IPTW)
  - G-Computation
  - Doubly Debiased Machine Learning (DDML)

Synthetic Simulations

Synthetic data includes:

- Binary Exposure (T): treatment (T=1) or control (T=0)
- Covariates (X): 20 covariates X
- Number of Patients: n=250, 500, and 1000

Two scenarios:

1. Homogeneous Treatment Effect
   - treatment effect is the same for all individuals
   - outcome models are linear functions of covariates

$$y = f(X, \Omega) + \theta T + \epsilon$$

2. Heterogeneous Treatment Effect
   - treatment effect varies based on interactions between covariates and the treatment.

$$y = (1 - T)f(X, \Omega_0) + Tf(X, \Omega_1) + \theta T + \epsilon$$

Internal Replication Study

To evaluate the methods using real-world clinical data by mimicking observational settings while having access to the true causal effect for validation

Data Source:

- 5 randomized clinical trials (RCTs) evaluating the efficacy of Canagliflozin in patients with Type 2 DM, primary endpoint of change in HbA1c from baseline

Trial	Nb. patients	Inclusion criteria	Arms	Background therapy
NCT01106625 [39]	469		Canagliflozin 300 Sitaglipin 100	Metformin and Sulphonylurea
NCT01137812 [40]	755		Canagliflozin 300 Canaglifozin 100 Placebo	Metformin and Sulphonylurea
NTC01106651 [41]	659	Age: 55 to 80 y.o.	Canagliflozin 300 Canaglifozin 100 Placebo	Metformin and Sulphonylurea (357 patients) Metformin (302 patients)
NCT01106677 [42]	1284		Canagliflozin 300 Canaglifozin 100 Sitaglipin 100 Placebo	Metformin
NCT00968812 [43]	1450	45≥BMI≥22	Canagliflozin 300 Canaglifozin 100 Glimepiride 100	Metformin

Creating Observational Experiment by

- Replacing the control arm of one trial with the treated arm from another trial

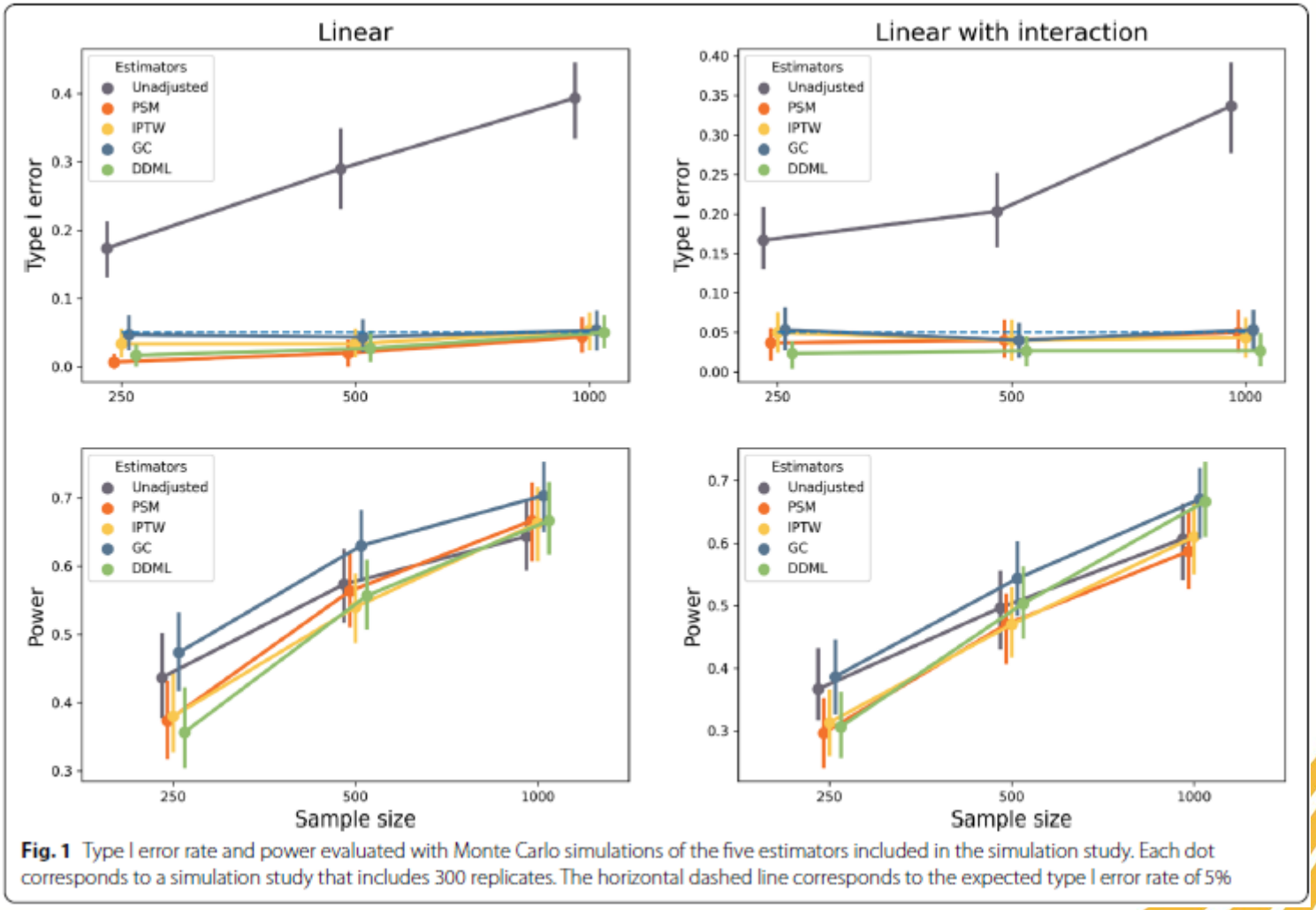


Fig. 1 Type I error rate and power evaluated with Monte Carlo simulations of the five estimators included in the simulation study. Each dot corresponds to a simulation study that includes 300 replicates. The horizontal dashed line corresponds to the expected type I error rate of 5%

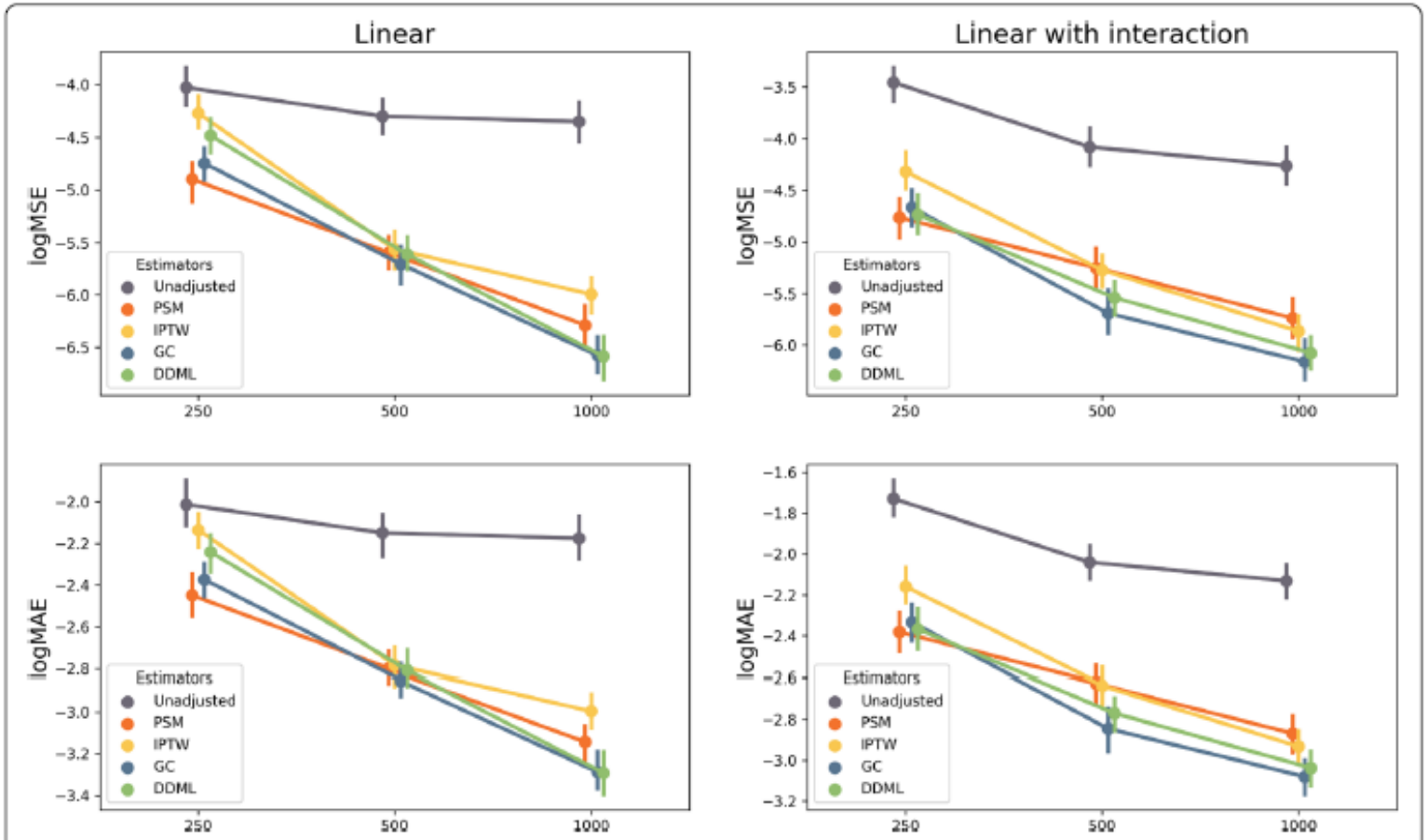


Fig. 2 Logarithm of the Mean Absolute Error (MAE) and Mean Squared Error (MSE) of the five estimators included in the simulation study. Each dot corresponds to a simulation study that includes 300 replicates

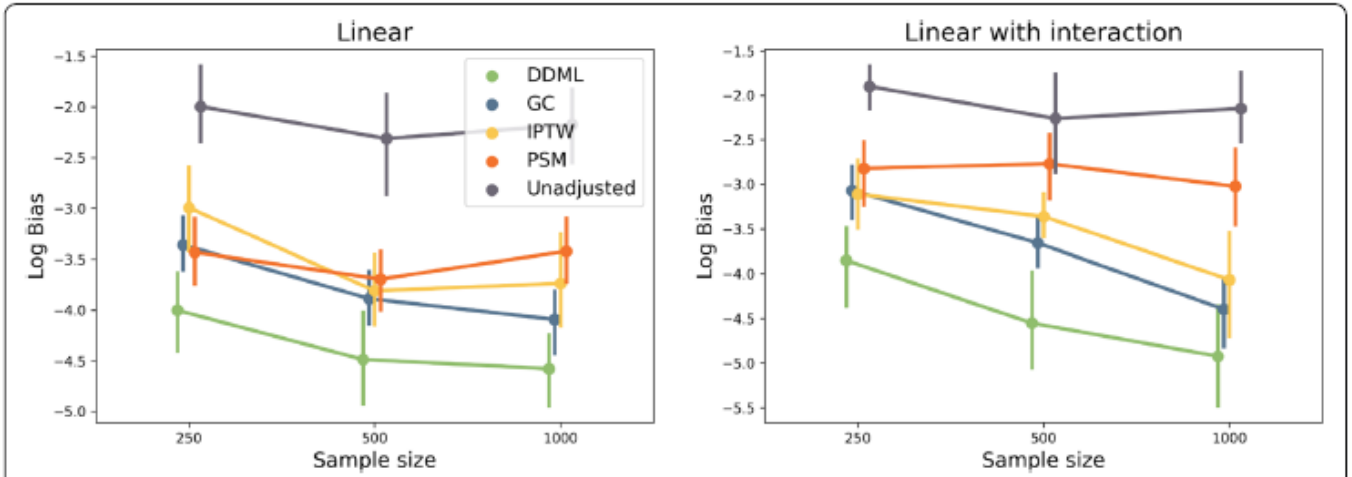


Fig. 3 Logarithm of the bias of the five estimators included in the simulation study. Each dot corresponds to a simulation study that includes 300 replicates

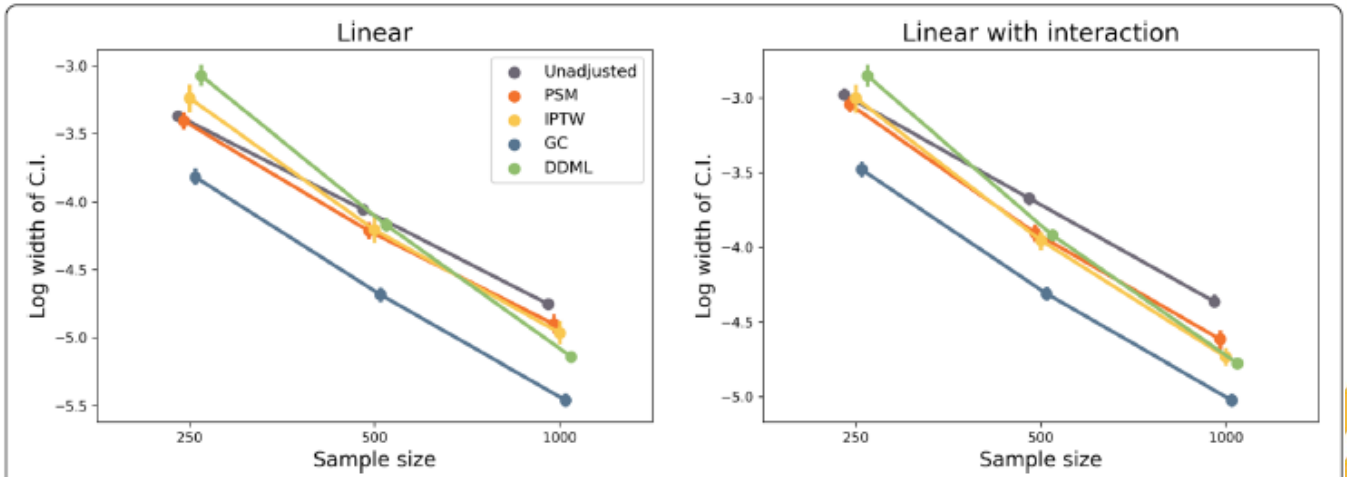


Fig. 4 Log width of the 95% Confidence Intervals (C.I) for the different methods. To measure the log width, we compute the logarithm of the variance of a Gaussian distribution which 95% C.I. would match the observed C.I. Each dot corresponds to a simulation study that includes 300 replicates

Table 2 Results of the negative control experiments when the experimental and control arms are the same. MSE and MAE are respectively the mean squared error and the mean average error between the ATT estimation and the ground truth, which is null. Coverage is the percentage of confidence intervals that contain zero

	MSE(x1000)	MAE(x100)	C.I. width (x1000)	Coverage(%)
Unadjusted	8.73	7.62	24.9	78% (7/9)
PSM	6.46	6.79	28.3	100% (9/9)
IPTW	4.79	5.91	27.9	100% (9/9)
G-computation	**3.53**	**4.79**	22.0	88% (8/9)
DDML	4.72	5.70	28.9	100% (9/9)

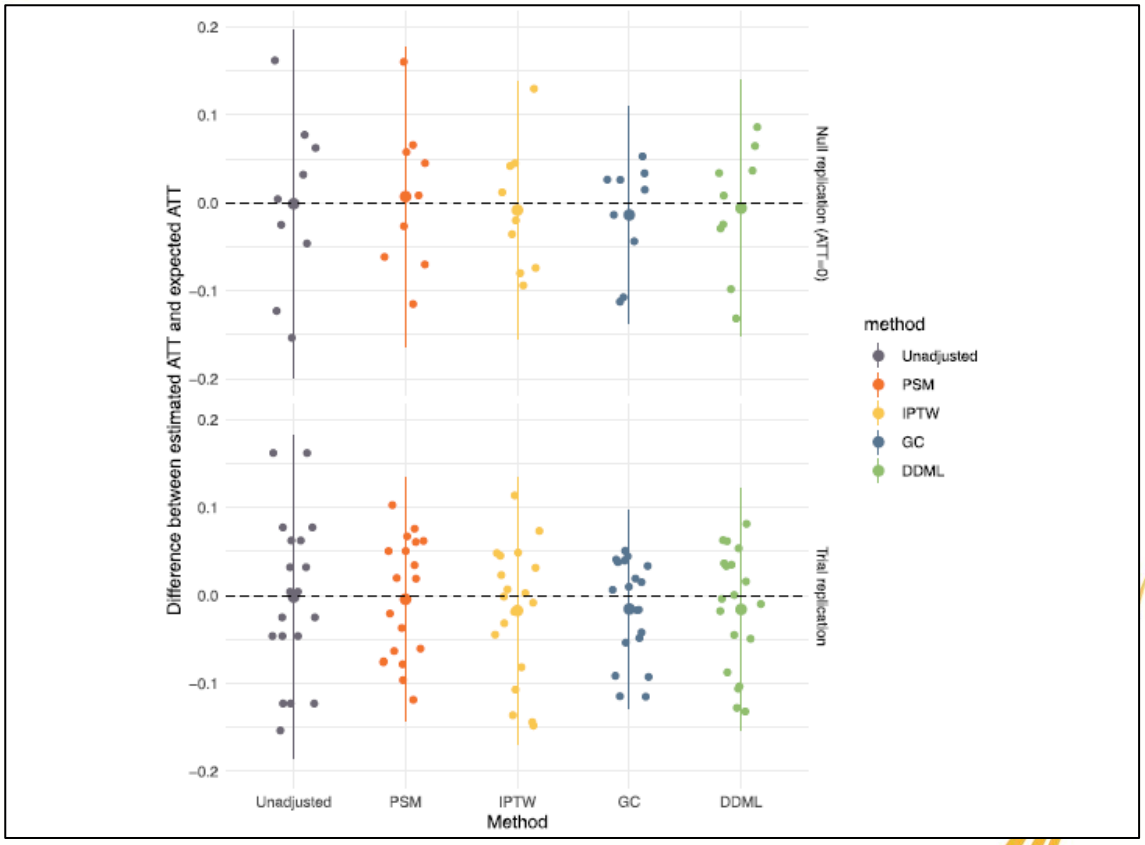


Table 3 Results of the RCT replication experiments. Pseudo MSE and MAE are respectively the pseudo mean squared error and the pseudo mean average error obtained by replacing the unknown ground truth with the RCT estimate. Estimate agreement is the percentage of RCT 95% confidence intervals that contain ATT estimation. Regulatory agreement is the percentage of time the cutoff $P < 0.05$ obtained from the non-randomized experiments agrees with the RCT result about $P < 0.05$

	Pseudo MSE(x1000)	Pseudo MAE(x100)	C.I. Width(x100)	Estimate Agreement	Regulatory Agreement
Unadjusted	7.94	7.30	25.1	84.2% (16/19)	73.7% (14/19)
PSM	4.51	6.15	29.0	89.5% (17/19)	73.7% (14/19)
IPTW	5.75	5.86	28.5	89.5% (17/19)	78.9% (15/19)
G-computation	**3.26**	**4.68**	25.9	100% (19/19)	78.9% (15/19)
DDML	4.70	5.60	31.3	100% (19/19)	84.2% (16/19)

Conclusions

- **G-Computation:** preferred when precision and power are critical, particularly in small-to-moderate sample sizes.
- **DDML**: excellent choice for high-dimensional data and large sample sizes, offering robust and unbiased estimates.
- **Propensity score**: less reliable in the tested scenarios

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Thank You!