Universität St.Gallen

Whom to prioritize? – Causal effects of treatment target non-adherence on mortality in intensive care medicine

EUHEA Conference 2026

Dr. Justus Vogel, Andrés Gallardo
Chair of Health Economics, Policy, and Management // School of Medicine, School of Management // University of St. Gallen

Dr. med. Julian Klug
University Hospital Lausanne

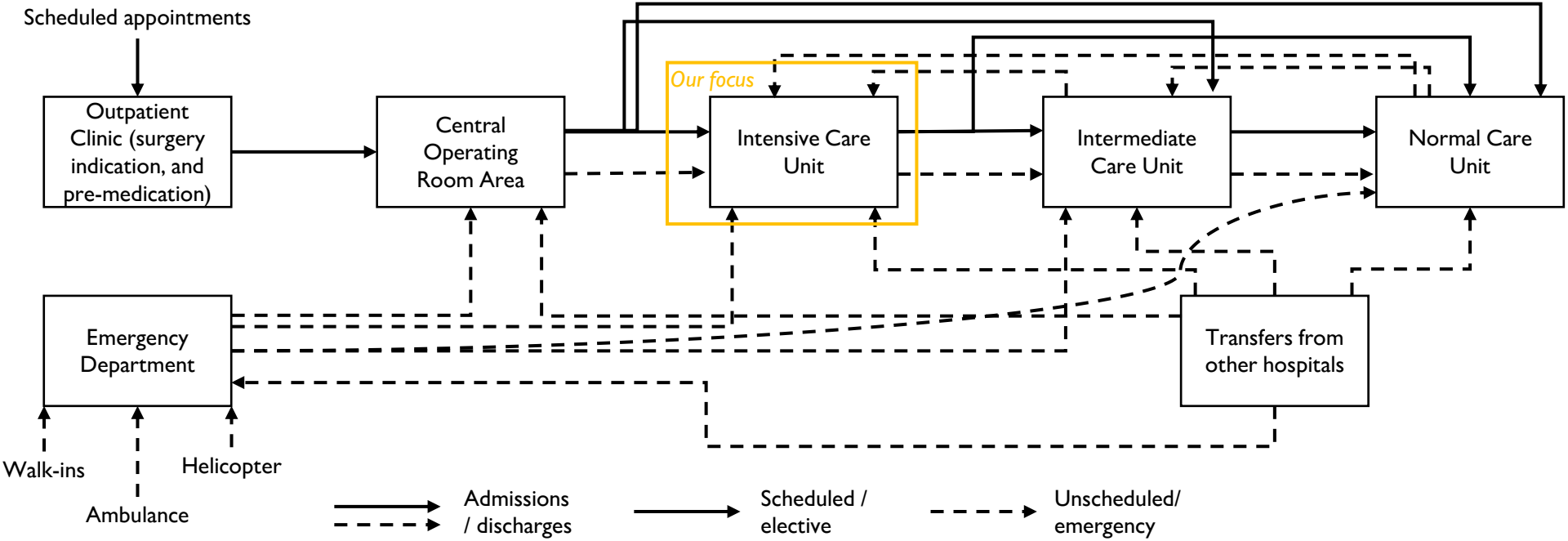
Prof. Dr. med. Miodrag Filipovic & Prof. Dr. med. Urs Pietsch
Cantonal Hospital of St. Gallen

Rotterdam, July 17, 2026

From insight to impact.

1. Introduction and Background

Capacities in Intensive Care Units are under constant pressure and strongly constraint – who profits the most from monitoring?



Sources: Keskinocak, P., & Savva, N. (2019); Bai et al. (2018); Chan et al. (2012); Kramer et al. (2016); Niven et al. (2014); Vogel et al. (2025)

Our research objective is to estimate and leverage **individualized average treatment effects** (IATEs) for learning a policy that helps physicians prioritize the right patients

2. Methods

Prediction and ATEs alone are not sufficient for (this kind of) decision support!

Perspective | Published: 19 April 2024

Causal machine learning for predicting treatment outcomes

Stefan Feuerriegel ✉, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane & Mihaela van der Schaar

Nature Medicine **30**, 958–968 (2024) | Cite this article

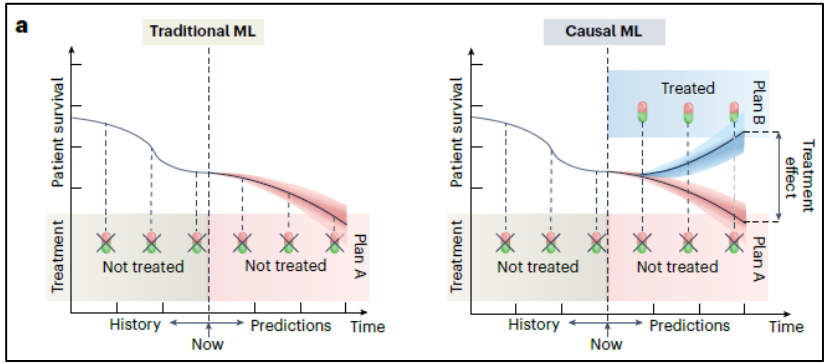
31k Accesses | **34** Citations | **177** Altmetric | Metrics

Abstract

Causal machine learning (ML) offers flexible, data-driven methods for predicting treatment outcomes including efficacy and toxicity, thereby supporting the assessment and safety of drugs. A key benefit of causal ML is that it allows for estimating individualized treatment effects, so that clinical decision-making can be personalized to individual patient profiles. Causal ML can be used in combination with both clinical trial data and real-world data, such as clinical registries and electronic health records, but caution is needed to avoid biased or incorrect predictions. In this Perspective, we discuss the benefits of causal ML (relative to traditional statistical or ML approaches) and outline the key components and steps. Finally, we provide recommendations for the reliable use of causal ML and effective translation into the clinic.

Sources: Feuerriegel et al. (2024)

Causal ML | Predict a patient's outcome when adherent vs. non-adherent – delta is the effect



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

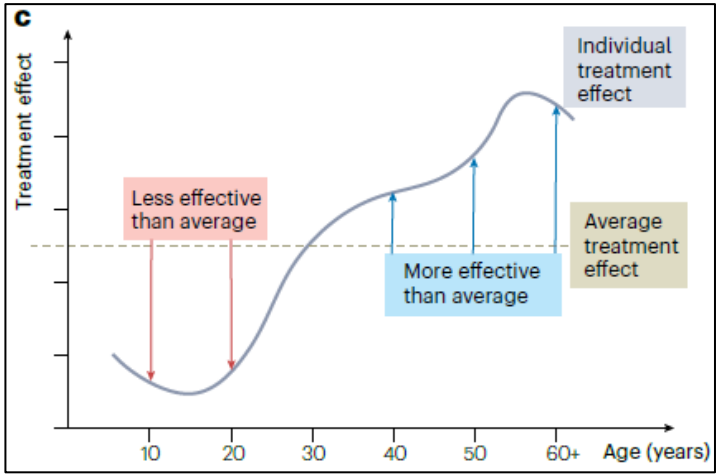
Data

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



Sources: Figure 1, p. 959, from Feuerriegel et al. (2024)

To estimate causal effects, we rely on the potential outcomes framework and doubly-robust CML…

To estimate causal effects, we rely on the potential outcomes framework (Rubin 1974). In an i.i.d. sample, there is a set of patients $i = 1, 2, \ldots, n$, each described by a feature vector $X_i \in \mathbb{R}$, a treatment assignment W_i which, in a binary setting, equals 1 if i was treated and 0 otherwise, and an outcome $Y_i \in \mathbb{R}$. The (potential) outcome according to the treatment status then is given by $Y_i(W_i)$ and the causal effect of a treatment for patient i is given by $Y_i(1) - Y_i(0)$.The central challenge in causal inference is that each example is either treated or not, and thus only $Y_i(1)$ or $Y_i(0)$ can be observed.

Unconfoundedness

$$[\{Y_i(0), Y_i(1)\} \perp W_i] \mid X_i$$

Common support / overlap

$$0 < p(W_i = 1 | X_i = x) < 1 \; \forall \; x \in X$$

Stable-unit-treatment-value assumption

Exogeneity

AIPW / double-robustness yields an acceptable rate of convergence ($1/\sqrt[4]{n}$) and limits bias

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{W_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}_1(X_i)) - \frac{(1 - W_i)}{1 - \hat{e}(X_i)}(Y_i - \hat{\mu}_0(X_i))\right)$$

Sources: Rubin (1974); Rosenbaum and Rubin (1983); Angrist et al. (1996); Imbens (2000); Lechner (2001); Wooldridge (2010); Chernozhukov et al. (2018); Wager and Athey (2018); Athey et al. (2019)

… and use causal forests to learn how strongly effects of treatment target non-adherence differ between patients

Individualized Average Treatment Effect (IATE)

$$IATE(W;x) = E[Y(1) - Y(0) | X = x]$$

Logic of causal trees and forests (Athey and Imbens, 2016; Wager and Athey, 2018)

$$\hat{\tau}_b(x) = \frac{1}{N_\ell^1} \sum_{i:X_i \epsilon \ell(x;b)} Y_i^1 - \frac{1}{N_\ell^0} \sum_{i:X_i \epsilon \ell(x;b)} Y_i^0$$

$$\hat{\tau}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b(x)$$

"Honest" Trees for valid inference and smaller bias

Generalized Random Forests (Athey et al., 2019)

$$\hat{\tau}(x) = \sum_{i=1}^{n} \alpha(x,i) \tilde{Y}_i \qquad \tilde{Y}_i = \frac{\big(Y_i - \hat{m}(X_i)\big)\big(W_i - \hat{e}(X_i)\big)}{\hat{e}(X_i)\big(1 - \hat{e}(X_i)\big)} \qquad \alpha(x,i) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbf{1}\{X_i \in l_b(x)\}}{|l_b(x)|}$$

Sources: Lechner and Mareckova (2025); Athey and Imbens (2016); Wager and Athey (2018); Athey et al. (2019)

Treatment & Outcome | We investigate the effect of non-adherence to mean arterial pressure targets on 28-day mortality

Treatment (W)

- Non-adherence to treatment target for mean arterial pressure (most common target)
- Lower threshold >90% of patients at 65 mmHg
- Adherence quartiles: Q1 most adherent, Q4 least adherent
- Measure confounders prior to adherence to avoid bias
- 27 sliding windows; 1st window considers 2 hours confounders and the following 8 hours for adherence; 1 hour added to confounders and 8-hour adherence shifted 1 hour for each new window

Outcome (Y)

- 28-day mortality
- Average 28-day mortality (first window): Overall 6.72%, Q1 3.81%, Q4 9.58%

We can extract and engineer hundreds of features to use them as confounders (and potentially IVs) – sample size of 10,314 ICU stays

Feature category	Features	Numeric type	Considered measurements	Description / examples
Medication & drugs	5,478	continuous	total dose over the covariate window	every substance × unit × administration form; e.g. fentanyl (microgram, IV), noradrenalin (microgram)
Clinical scores	54	52 continuous, 1 dummy, 1 factor	SAPS 3 first hour; SAPS II windowed (long windows); GCS & pupillometry over the covariate window	SAPS 3 (+ sub-scores), SAPS II, GCS, pupillometry (NPi)
Vital signs	36	continuous	summaries over the covariate window	MAP, systolic BP, heart rate, SpO2, temperature (mean, SD, min, max, p05, p95)
Blood gas analysis	28	continuous	summaries over the covariate window	signed base-excess, lactate, bicarbonate, pO2, etc.
Ventilation & respiratory support	14	12 continuous, 2 dummy	over the covariate window	ventilation mode, FiO2 (%), PEEP
Laboratory tests	9	continuous	long covariate windows (≥25h) only	priority labs, e.g. haemoglobin, platelets, sodium
Basic data	8	4 continuous, 2 dummy, 2 factor	time invariant	age, sex, BMI, height, weight, admission type
Fluid output	2	continuous	over the covariate window	urine output
Therapy	2	dummy	admission time	MDSi admission-time flags
Infusions	1	continuous	over the covariate window	total infusion volume (ml)
Total	**5,623**[a]			

[a] Short-window total; long windows (≥25h) reach 5,646 features (laboratory tests +9, SAPS II physiology sub-scores +14).
[b] Diagnoses, infections, mobilisation, and blood products are deliberately excluded as they are not valid pre-treatment confounders (e.g. diagnoses are partly recorded post-admission).
[c] Most of the 5,478 medication columns are all-zero in a given window (drug not given) and are dropped before estimation; e.g. ∼688 are non-zero in the 2h window.

Data years 2018 to 2024

Roughly 1,900 ICU stays per year, 18-20 operatable beds

Sample size: 10,314 ICU stays for first and 4,326 for last window

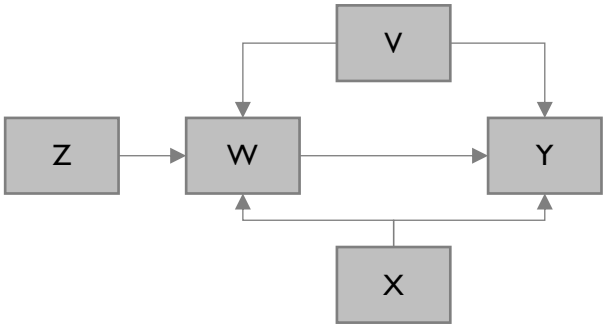
We are not done yet!
Data cleaning and feature engineering is a never-ending story with real-world data…

More confounders may be engineered from clinical scores, infusions, ventilation mode, therapies, basic data (e.g., number of prior ICU stays), etc.

One strength of our study is that we can control for hundreds of confounders – typical overlap for real-world data settings

Unconfoundedness

$$[\{Y_i(0), Y_i(1)\} \perp W_i] \mid X_i$$



Y: outcome

W: treatment (status)

X: confounders

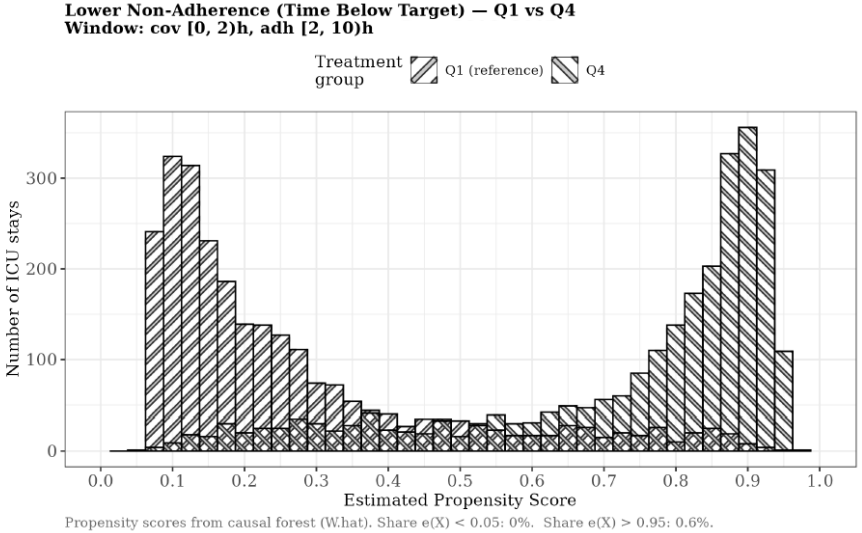
V: Unobserved confounder

Z: Instrumental variable (IV)[1]

Common support / overlap

$$0 < p(W_i = 1 | X_i = x) < 1 \; \forall \; x \in X$$

Example: First treatment window



1) Possibly, certain medication that only influences mean arterial pressure and does not influence mortality risk could serve as IV.
Sources: Angrist et al. (1996); Imbens (2000); Lechner (2001)

3./4. (First) Results and Discussion

Strong non-adherence increases mortality risk across all treatment windows

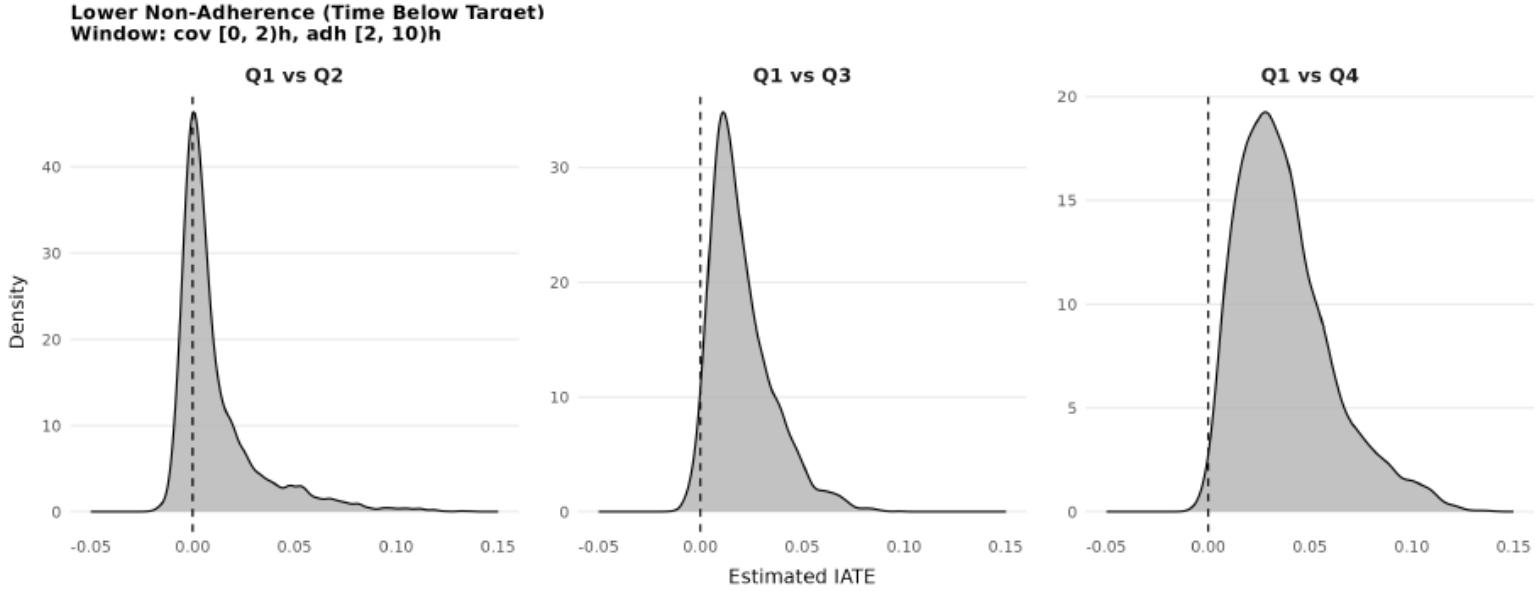
All treatment windows

Direction	Contrast	Significant	Mean ATE	Sign
Lower	Q1Q2	2 / 27	−0.58	mixed (null)
Lower	Q1Q3	2 / 27	+1.19	25/27 positive
Lower	**Q1Q4**	**17 / 27**	**+3.52**	**all positive**
Binary	any-vs-none	**13 / 27**	**+1.84**	**all positive**

Primary window

Contrast	*n*	ATE [95% CI]	Sig.
Q1 vs Q2	5,112	+1.18 [0.03, 2.33]	Yes (barely)
Q1 vs Q3	5,139	+**2.11** [0.59, 3.64]	**Yes**
Q1 vs Q4	5,157	+**3.89** [2.03, 5.75]	**Yes**

Primary window IATE distribution | The stronger the non-adherence, the larger the share of patients with positive effects



IATE = individual-level average treatment effect from causal forest.
Vertical dashed line at zero.

Generally, strong non-adherence to MAP targets seems to increase mortality risk

Individualized effects are positive for most patients, yet the magnitude of the effect differs – this enables ranking patients according to their effect and can be used for decision-making

Whom to prioritize?

Open to-dos remain until we can be fully confident of our results…

1. Include more confounders and improve feature engineering
2. Explore IV strategy to manage remaining confounding
3. Hyperparameter tuning (more trees? Minimum node sizes?) and evaluation of standard errors for IATEs
4. Develop decision policy for prioritization
5. Explore other treatment targets and add more data (from other centers…)

APPENDIX

Variable Importance Measure Score – Top 10

Rank	Variable	Importance
1	Age	0.029
2	SAPS 3 score	0.028
3	FiO2 (p95)	0.028
4	FiO2 (mean)	0.023
5	SAPS 3 age sub-score	0.022
6	SpO2 score (mean)	0.020
7	FiO2 (max)	0.019
8	FiO2 (sd)	0.017
9	Systolic BP (p95)	0.017
10	Admission hour	0.016

Identifying assumptions | SUTVA and Exogeneity should both be fulfilled

Stable Unit Treatment Value Assumption

The treatment state of one unit only influences its own outcome

Treatment variation between units is minimal, meaning that the observed outcome of one patient in the treatment state corresponds to the potential outcome for all patients in that state

"No spillover"

Discharging one patient does not directly influence the readmission risk of another patient

What if $B_a \geq 0$ holds? And no additional discharges occur even if they were possible/ meaningful from a medical perspective? → This might influence the treatment state of another patient but not their outcome

Exogeneity

Patient characteristics used as confounders (X) are not influenced by the discharge at point in time t (W, the treatment)

We observe X before the discharge occurs

Sources: Angrist et al. (1996); Imbens (2000); Lechner (2001)

With unconfoundedness, we can restate the ATE with conditional response functions $\mu_w(x)$

With unconfoundedness, the ATE can be expressed as:

$$\tau = E\left[\mu_1(X_i) - \mu_0(X_i)\right]$$

where

$$\mu_w(x) = E[Y_i \mid X_i = x, W_i = w]$$

To estimate $\hat{\tau}$, we can learn $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ from our data and calculate:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right)$$

Biased if non-parametric distribution – also biased when "simply" applying predictor with finite sample

Sources: Rubin (1974); Lechner (2001); Rosenbaum and Rubin (1983); Wooldridge (2010)

Propensity scores are used to weigh $\mu_W(x)$, yielding augmented inverse probability weighted or "double-robust" scores

$$e(x) = p(W_i = 1 | X_i = x)$$

$$\hat{\tau}^*_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right)$$

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{W_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}_1(X_i)) - \frac{(1-W_i)}{1-\hat{e}(X_i)}(Y_i - \hat{\mu}_0(X_i))\right)$$

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\hat{\Gamma}_i$$

Propensity score: Gives the probability of receiving treatmnt based on X_i

The oracle IPW estimator is unbiased, but the rate of convergence is too slow when using machine learning methods

AIPW / double-robustness yields an acceptable rate of convergence ($1/\sqrt[4]{n}$) and limits bias

$\hat{\Gamma}_i$ marks an individual's double-robust score

Sources: Rosenbaum and Rubin (1983); Chernozhukov et al. (2018); Robins et al. (1994)

Literature (excerpt)

Bai, J., Fügener, A., Schoenfelder, J., & Brunner, J. O. (2018). Operations research in intensive care unit management: a literature review. Health Care Management Science, 21(1), 1–24. https://doi.org/10.1007/s10729-016-9375-1

Chan, C. W., Farias, V. F., Bambos, N., & Escobar, G. J. (2012). Optimizing Intensive Care Unit Discharge Decisions with Patient Readmissions. Operations Management, 60(6), 1323– 1341. https://doi.org/10.1287/OPRE.1120.1105

Diwas Singh, K. C., Scholtes, S., & Terwiesch, C. (2020). Empirical research in healthcare operations: Past research, present understanding, and future opportunities. *Manufacturing & Service Operations Management*, 22(1), 73–83. https://doi.org/10.1287/MSOM.2019.0826

Keskinocak, P., & Savva, N. (2019). A Review of the Healthcare-Management (Modeling) Literature Published in Manufacturing & Service Operations Management. *Manufacturing & Service Operations Management*, 22(1), 59–72. https://doi.org/10.1287/MSOM.2019.0817

Kramer, A.A., Higgins, T.L., Zimmerman, J.E. (2013). The association between ICU readmission rate and patient outcomes. Crit Care Med 41, 24–33. https://doi.org/10.1097/CCM.0B013E3182657B8A

Kramer, A. A., Higgins, T. L., & Zimmerman, J. E. (2016). Can this patient be safely discharged from the ICU? Intensive Care Medicine, 42(4), 580–582. https://doi.org/10.1007/s00134-015-4148-8

Niven, D. J., Bastos, J. F., & Stelfox, H. T. (2014). Critical care transition programs and the risk of readmission or death after discharge from an ICU: A systematic review and meta-analysis. *Critical Care Medicine*, 42(1), 179–187.

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228– 1242. https://doi.org/10.1080/01621459.2017.1319839