Comparing Machine Learning Methods for Estimating Individualized Treatment Effects: A Simulation Study

Authors: Diane Vincent, MSc[1], **Antoine Movschin, MSc[1]**, Tristan Fauvel, PhD[1]
Affiliation: [1]Quinten Health, Paris, France

MSR58
ISPOR Europe
2025

INTRODUCTION

- Randomized controlled trials (RCTs) are the gold standard for estimating average treatment effects (ATE) due to their rigorous design, which effectively minimizes bias. However, they are constrained by limited representativeness and capacity to provide estimates at an individual level[1-3] .
- The increasing availability and richness of real-world data (RWD) presents an opportunity to estimate individual treatment effects (ITE) in real-world situations. However, use of RWD comes with various biases[1-3] .
- To this end, many causal inference methods for estimating the ITE based on machine learning (ML) have emerged. They rely on various assumptions and present strengths and weaknesses depending on the context. Moreover, no consensus exists as to how they should be evaluated[1,5-6] .

OBJECTIVE

This work aims at providing guidance for selecting the best ITE estimation approach across use cases through extensive simulations, diverse scenarios, and evaluation of representative methods using comprehensive metrics.

BACKGROUND

Various methods (Table 1) and metrics (Table 2) were identified for estimating ITE.

Notations and definitions

$\boldsymbol{T}$: Treatment assignment, $T \in \{0,1\}$
$\boldsymbol{X}$: Patients characteristics, covariates
$\boldsymbol{Y(T=0), Y(T=1)}$: Potential outcomes
$\boldsymbol{Y}$: Observed outcome (binary in our study)
$\boldsymbol{\tau(X)} = \mathbb{E}[Y(1) - Y(0)|X]$: Conditional average treatment effect (CATE)
$\boldsymbol{\pi(X)}$: Propensity score
$\boldsymbol{\mu(X)} = \mathbb{E}[Y|X]$: Conditional average outcome
$\boldsymbol{n(x)}$: Nearest neighbor of patient with covariate $X = x$ from the opposite treatment group

Table 1: List of selected methods for the simulation study

Family	Method
Baseline (ATE)	Adjusted Difference in Means (ADM)[5]
Meta-learners	S-learner[1], T-learner[1], X-learner[1], DR-learner[1], Double Machine Learning (DML)[1], R-learner[2]
Tree-based methods	Causal Forest (CF)[1]
Bayesian methods	BART[3], BCF[5]
Deep Learning	CFR Wasserstein[5], CEVAE[4], GANITE[5]

Python package: [1]EconML, [2]Causallib, [3]pyMC, [4]Pyro, [5]Other

Table 2: List of selected metrics for the simulation study

Type	Metric	Formula: $\mathbb{E}(\cdot)$
Oracle metrics only accessible in a simulation study	PEHE	$\left(\hat{\tau}(X) - \tau(X)\right)^2$
	Coverage	$\mathbb{I}(\tau(X) \in [\hat{\tau}^{low}(X), \hat{\tau}^{high}(X)]$
	Interval width	$\hat{\tau}^{high}(X) - \hat{\tau}^{low}(X)$
	Policy risk	$1 - Y(T = \mathbb{I}(\tau(X) > 0))$
Observable metrics accessible in real-world situations	IF-PEHE	$\left(\hat{\tau}(X) - \tilde{\tau}(X)\right)^2 + IF_{\tilde{\tau}}(X, T, Y; \hat{\tau})$
	PEHEnn	$\left(\hat{\tau}(X) - (1-2T)\left(Y_{n(X)} - Y\right)\right)^2$
	R-Loss	$\left(Y - \tilde{\mu}(X) - \hat{\tau}(X)\left(T - \tilde{\pi}(X)\right)\right)^2$
	Policy risk (Factual)	$1 - Y\|T = \mathbb{I}(\tau(X) > 0)$
	FO Brier Score	$(\hat{\mu}(X) - Y)^2$

- PEHE: precision of estimating heterogeneous effects,
- FO: factual outcome,
- IF: influence function,
- nn: nearest neighbor
- ^ refers to the estimate of interest
- ~ refers to plugin estimates,
- low and high refer to the bounds of the generated confidence intervals.

METHODS

ITE estimation methods were compared through a simulation study (Fig. 1).

1. Data generation

To generate a representative set of constraints typical of healthcare data, we defined a set of scenario-varying constraints (Table 3). For each scenario, we generated 100 low-dimensional datasets with independent random seeds. The data-generating process (DGP) enables control and knowledge of the true ITE.

2. Modeling

We covered a representative and diverse set of ITE estimation methods (Table 1).

3. Evaluation

To compare the methods, we used various metrics[1,3-6] (Table 2) and robustness checks[1].

Implementation details available upon request.

Figure 1: Simulation study

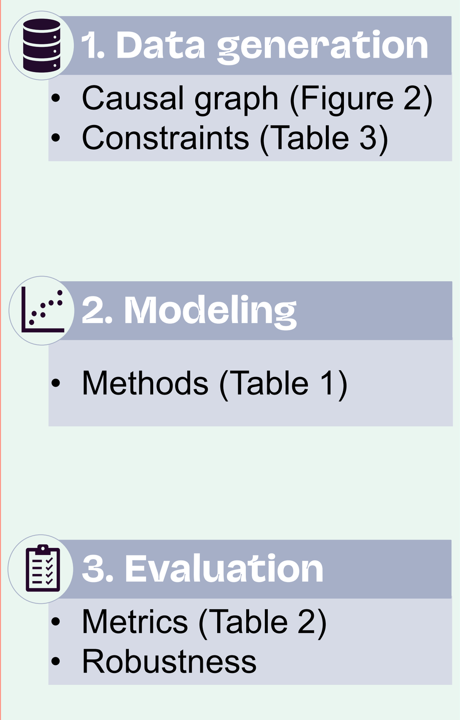
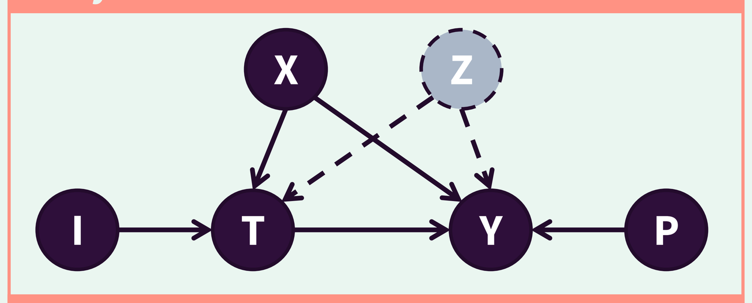


Table 3: List of scenario-varying parameters

Parameter	Options
Sample size	100, **1000**, 5000
Treatment effect heterogeneity	None, **Monotonic**, Complex
Covariate overlap	Low, **High**
Treatment prevalence	**50%**, 90%
Outcome frequency	**50%**, 90%
Confounding strength	**Medium**, Strong
Unobserved confounding	With, **Without**

Options in **bold** correspond to the baseline scenario.

Figure 2: Causal graph of the simulation study



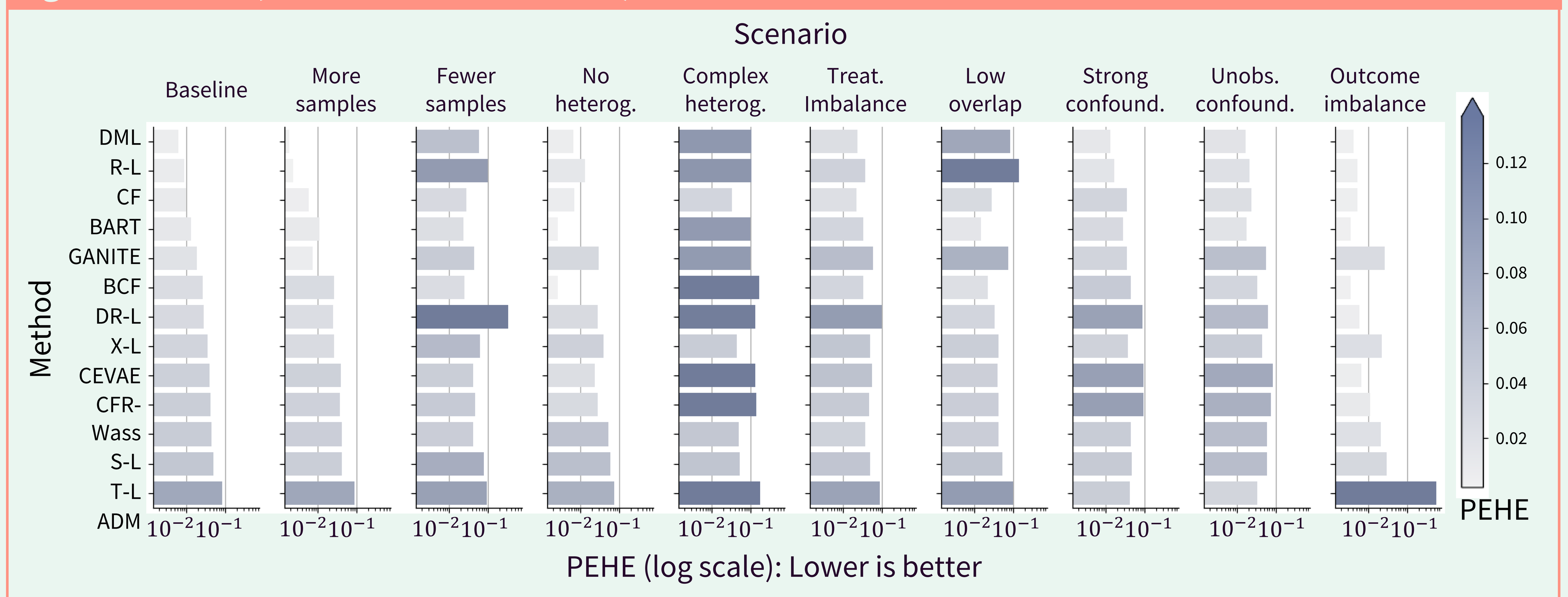
T: Treatment assignment
Y: Outcome
X: Confounder
I: Instrumental variable
P: Predictive variable
Z: Unobserved confounder

RESULTS

- Methods are generally accurate (low PEHE) but struggle with lower sample size and complex heterogeneity, their individual performance varies with the scenarios (Fig. 3).
- No method clearly dominates over all scenarios and metrics (Fig. 3)[4].
- In average, the best estimator for each scenario has a PEHE 27,5 times lower than the baseline ATE estimator (ADM) (Fig. 3), demonstrating the benefits of using CATE estimators.
- Built-in confidence intervals (when provided) are overly wide, limiting their practical use*.
- Robustness tests confirm the methods do not produce spurious treatment effects*.
- Observable metrics are not consistent with oracle metrics and thereby unreliable* (Fig. 4)[5,6].

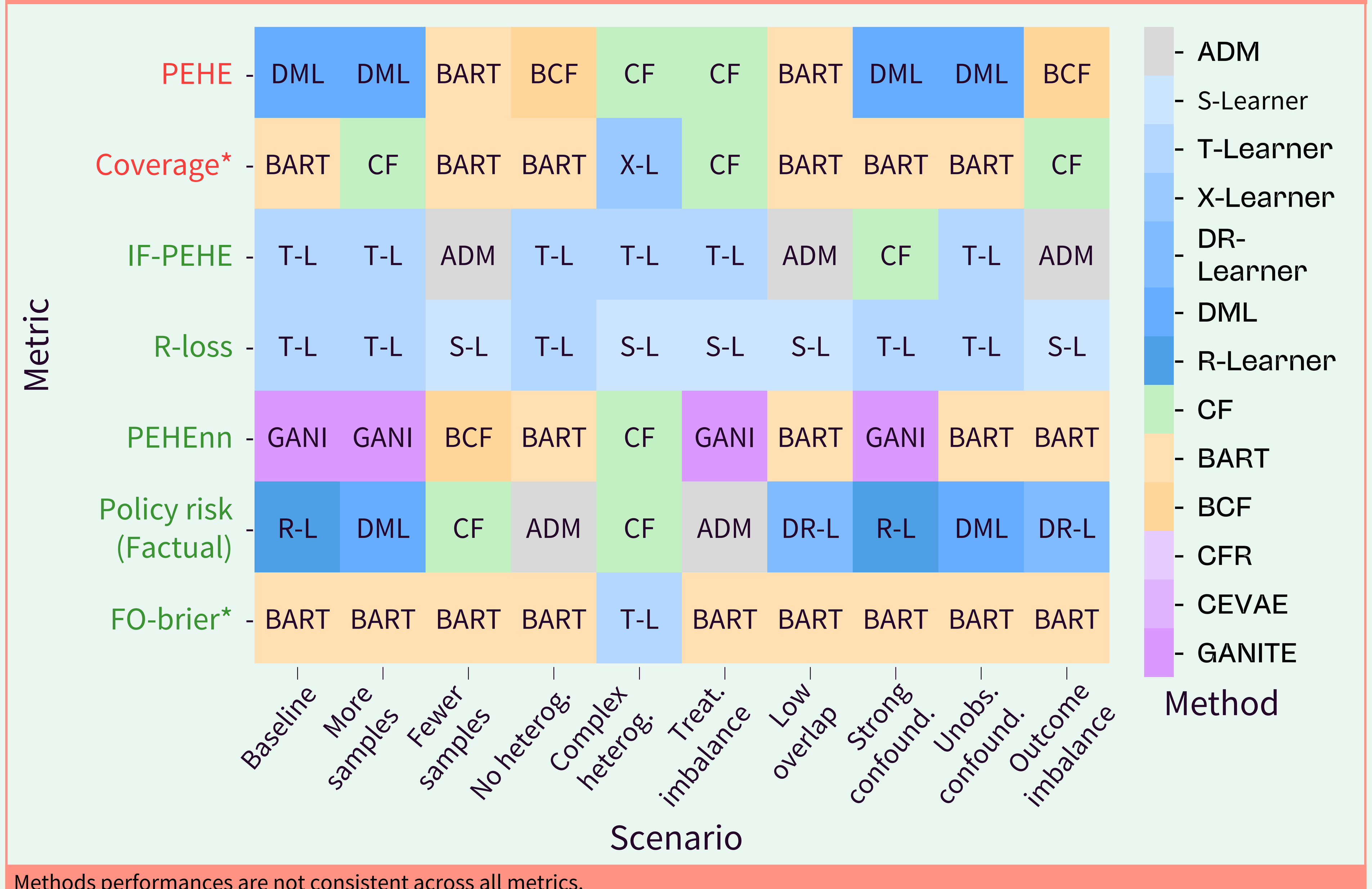
* Details available upon request

Figure 3: PEHE (and associated 95% CI) for each method and scenario



Methods performance varies with scenarios.

Figure 4: Best selected method for each metric and scenario



Methods performances are not consistent across all metrics.
*Not all methods provide the prediction associated to this metric.

DISCUSSION

- Most ITE estimation methods demonstrate reasonable accuracy but face limitations due to unreliable confidence intervals, limiting practical use. The lack of universal observable metrics and their inconsistency with oracle metrics hinders broader adoption of these methods. Future work should prioritize adressing these issues[1-2,5-6] .
- The DGP of the simulation study has notable limitations, including low-dimensionality, lack of time dependency and potentially insufficiently restrictive constraints.

CONCLUSION

Our simulation study maps the performance of ITE estimation methods across diverse settings and evaluation metrics, providing guidance for method selection in real-world contexts, for specific use-cases.

CONFLICT OF INTEREST: NA
CONTACT INFO: a.movschin@quinten-health.com

REFERENCES

1. Feuerriegel, S. et al. Causal machine learning for predicting treatment outcomes. Nat Med 30, 958–968 (2024).
2. Bica, I., Alaa, A. M., Lambert, C. & Van Der Schaar, M. From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges. Clin Pharma and Therapeutics 109, 87–100 (2021).
3. Ling, Y., Upadhyaya, P., Chen, L., Jiang, X. & Kim, Y. Emulate randomized clinical trials using heterogeneous treatment effect estimation for personalized treatments: Methodology review and benchmark. Journal of Biomedical Informatics 137, 104256 (2023).
4. Caron, A., Baio, G. & Manolopoulou, I. Estimating Individual Treatment Effects using Non-Parametric Regression Models: a Review. Journal of the Royal Statistical Society Series A: Statistics in Society 185, 1115–1149 (2022).
5. Curth, A., Svensson, D., Weatherall, J. & van der Schaar, M. Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation. (2021).
6. Mahajan, D., Mitliagkas, I., Neal, B. & Syrgkanis, V. Empirical Analysis of Model Selection for Heterogeneous Causal Effect Estimation. in Proceedings of the 12th International Conference on Learning Representations (2023).

quinten health