Benchmarking Six Causal Inference Methods for ATE Estimation: A Comprehensive Simulation Study

Zung-Ru Lin*
University of Pennsylvania, Philadelphia, PA, USA
Email: lin61302@gmail.com

April 5, 2025

Abstract

Estimating the average treatment effect (ATE) from observational data is a fundamental challenge in medical and policy research. Numerous methods have been proposed to address this task, ranging from classical interpretable techniques (e.g., propensity score matching and weighting) to more flexible machine learning approaches. However, the relative performance and trade-offs of these methods in realistic settings remain incompletely understood. In this study, we conduct an extensive simulation-based comparison of six commonly used ATE estimation approaches, including two propensity score matching strategies, two propensity score weighting schemes (one using logistic regression and one using a boosted decision tree model), and two doubly robust estimation methods (targeted maximum likelihood estimation (TMLE) and its doubly robust extension, DRTMLE). The simulation design emulates a plausible clinical scenario with a known true ATE, enabling unbiased evaluation of each method's accuracy, precision, and coverage across repeated trials. Our results reveal substantial discrepancies among the methods. Approaches that incorporate flexible modeling of the outcome or propensity score (such as the use of boosting within the propensity score estimation or nonparametric outcome regression in TMLE) tend to achieve lower bias and more stable estimates compared to simpler parametric or matching methods. At the same time, the estimates produced by these advanced techniques remain interpretable as causal effects, which is crucial for real-world decision-making. This study highlights the importance of simulation-based benchmarking to understand method performance and provides practical guidance on selecting appropriate, interpretable causal inference tools.

Keywords: causal inference, average treatment effect, propensity score, matching, doubly robust estimation, targeted maximum likelihood estimation

*This research was conducted independently and was not supported by the author's affiliated institution.

1 Introduction

In medical and epidemiological research, it is crucial to accurately determine whether a treatment truly improves patient outcomes, especially when randomized trials are infeasible or unethical. In previous research, we analyzed observational health data and identified smoking as a dominant risk factor for costly and deadly conditions. Smoking contributes to diseases such as emphysema, chronic bronchitis, and lung cancer, which carry a high risk of mortality. Meanwhile, some individuals have unscrupulously claimed that their medications can cure such serious diseases. According to *The Guardian*, *"More than 110 countries have reported more than 2,000 cases of bad drugs over WHO's global surveillance and monitoring system,"* noted Michael Deats.[1] In other words, driven by tremendous potential profit, some actors promote ineffective or harmful treatments—jeopardizing lives and eroding public trust in the medical environment. In such an environment, rigorous evaluation of treatment effectiveness is critical to ensure that only genuinely beneficial interventions are advocated.

This tension motivates our research comparing different methodologies to examine treatment effects. We illustrate this approach in the context of a lung cancer scenario, chosen for its high mortality and well-understood risk factors (such as smoking) to ground our simulation in a realistic clinical setting. Specifically, we explore the ability to estimate the true ATE using the following methods: nearest neighbor matching on propensity scores, full matching on propensity scores, propensity score weighting by logistic regression, propensity score weighting by a generalized boosted model, targeted maximum likelihood estimation, and doubly robust targeted maximum likelihood estimation. Given the true ATE from simulation, we can compare the accuracy of each method. The outcomes give credence to our research, yielding meaningful insights for practitioners.

In this study, we present a rigorous simulation-based comparison of these six interpretable methods under a unified, realistic scenario. By combining matching, weighting, and doubly robust estimators in one framework, our evaluation highlights each method's relative strengths, common pitfalls, and biases in causal effect estimation. With the true ATE known from the simulated population, we directly assess the bias and variability of each estimator. The results provide insight into how each technique performs under realistic conditions and form the basis for practical recommendations. In summary, by evaluating matching, weighting, and doubly robust strategies side by side, we offer a comprehensive comparison of popular ATE estimation methods and actionable guidance on choosing among them in applied research.

2 Methods

2.1 Data Generating Process

We simulate data to reflect a real-world scenario of patients suffering from lung cancer. Each patient is monitored for five years to record whether they survive or not. We consider smoking as a strong predictor of death for lung cancer patients, based on prior analysis. However, we aim to quantify the effect of smoking on the odds of death. Our simulation incorporates covariates inspired by those in Luque-Fernandez's tutorial on targeted maximum likelihood estimation for a binary treatment,[5] combined with additional factors found in online sources, which are believed to