Project GitHub

# `causal-curve`: python tools to preform causal inference on continuous treatments

Roni Kobrosly PhD
Head of Data Science, DrFirst, Washington, D.C.

## Package Goal

The “causal inference” python ecosystem is rapidly growing. To our knowledge, this is the first python package able to conduct causal inference on continuous treatments. This enables users to estimate the possibly non-linear causal (not just correlational) effect of treatments such as:

- Minutes per week of aerobic exercise
- Product price changes
- Neighborhood income inequality (as measured by the continuous Gini index)
- Customer wait time in seconds

## Package structure

**Generalized Propensity Score (GPS) Tool**
Employs generalized additive models to estimate the GPS (a score that captures how confounders influence treatment “assignment”), and then produces a curve of how the treatment relates to outcome, controlling for the GPS. computationally efficient, better suited for large datasets, but produces significantly wider confidence intervals

**Targeted Maximum Likelihood Estimation (TMLE) Tool**
Uses gradient boosting and a “targeting” step to correct for covariate imbalance and to estimate an unbiased causal effect. TMLE method is double robust against model misspecification, produces significantly smaller confidence intervals, but is less computationally efficient and not recommended for big datasets.

**Mediation Tool**
Allows user to estimate the, dynamic, direct and indirect contributions of a mediator between a continuous treatment and outcome.

**Project Build**
Project contains both unit and integration tests and is built with CI/CD (Travis-CI) including linting and code coverage checks.

**Project Usability**
The API was designed to resemble that of `scikit-learn`, as this is a Python predictive modeling framework familiar to most machine learning practitioners. All the major classes contained in causal-curve readily use Pandas DataFrames and Series as inputs, so this package easily integrates with the modern Python data stack.

## Case Study: Understanding the impact of childhood lead exposure

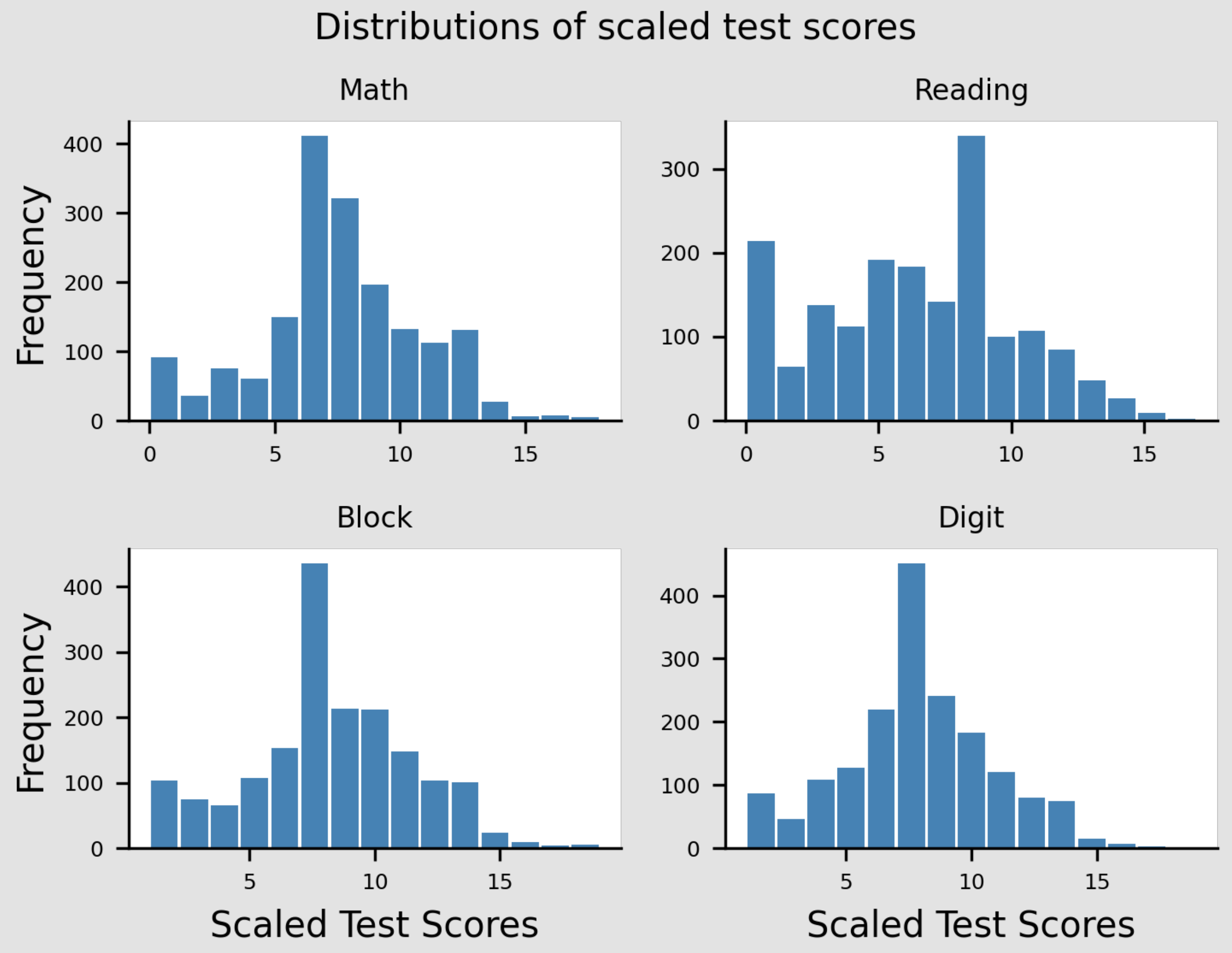
Despite the banning of the use of lead-based paint and the use of lead in gasoline in the United States, lead exposure remains an enormous public health problem for children and adolescents. This is particularly true for poorer children living in older homes in inner-city environments.

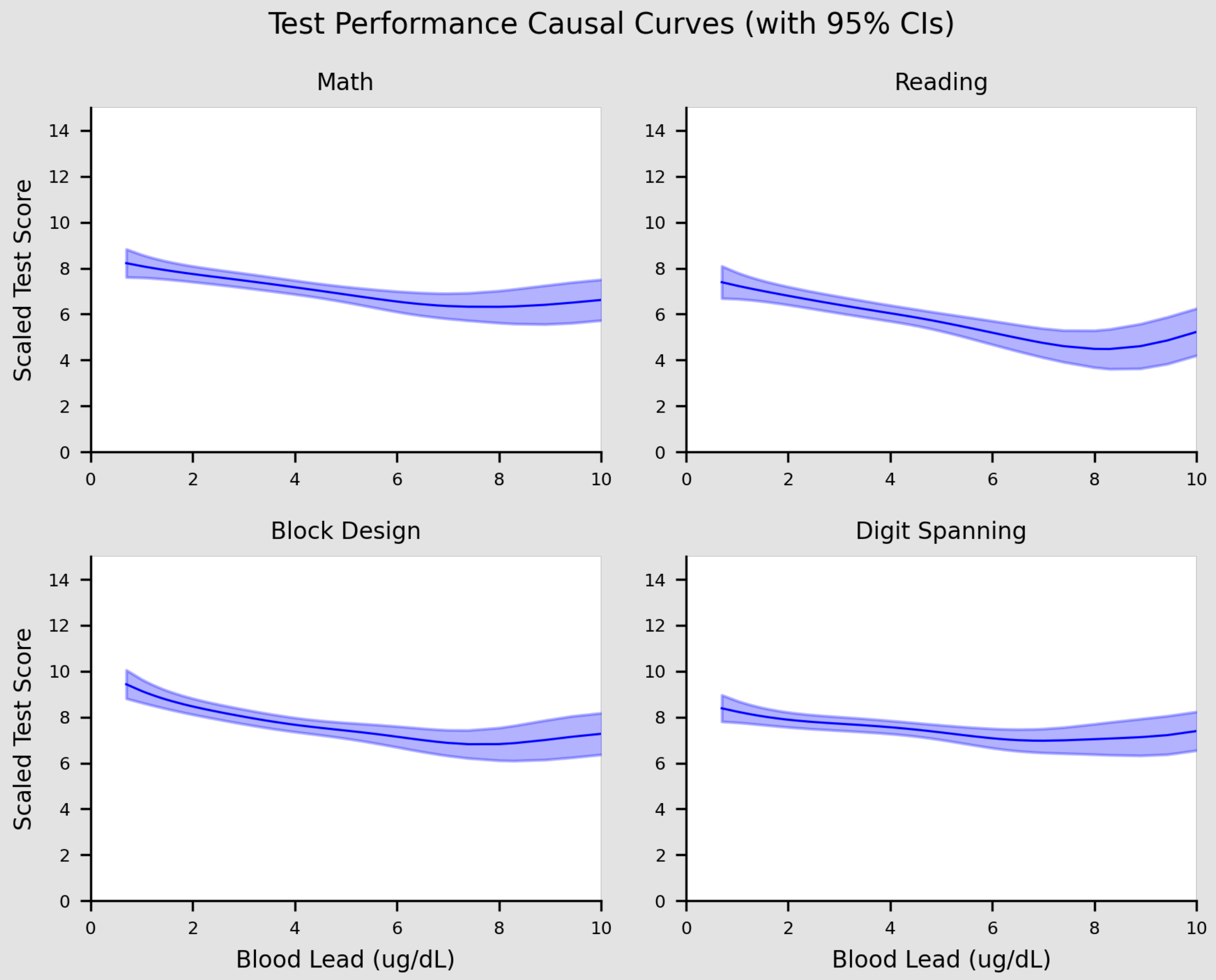
For children, there is no known safe level of exposure to lead, and even small levels of lead measured in their blood have been shown to affect IQ and academic achievement, permanently.

In terms of public policy, it would be helpful to understand how childhood cognitive outcomes would be affected by reducing BLLs in children.

We analyzed data collected from the National Health and Nutrition Examination Survey (NHANES) III. This was a large, national study of families throughout the United States, carried out between 1988 and 1994. Children ages 6 – 12 years were involved in extensive interviews, medical examinations, and provided biological samples. As part of this project, blood lead levels were measured, and four scaled sub-tests of the WISC/WRAT cognitive test were carried out. These included:

- a math test
- a reading test,
- a block design test (a test of spatial visualization ability and motor skill)
- a digit spanning test (a test of memory).





This data is de-identified and publicly available on the Centers for Disease Control and Prevention (CDC) government website.

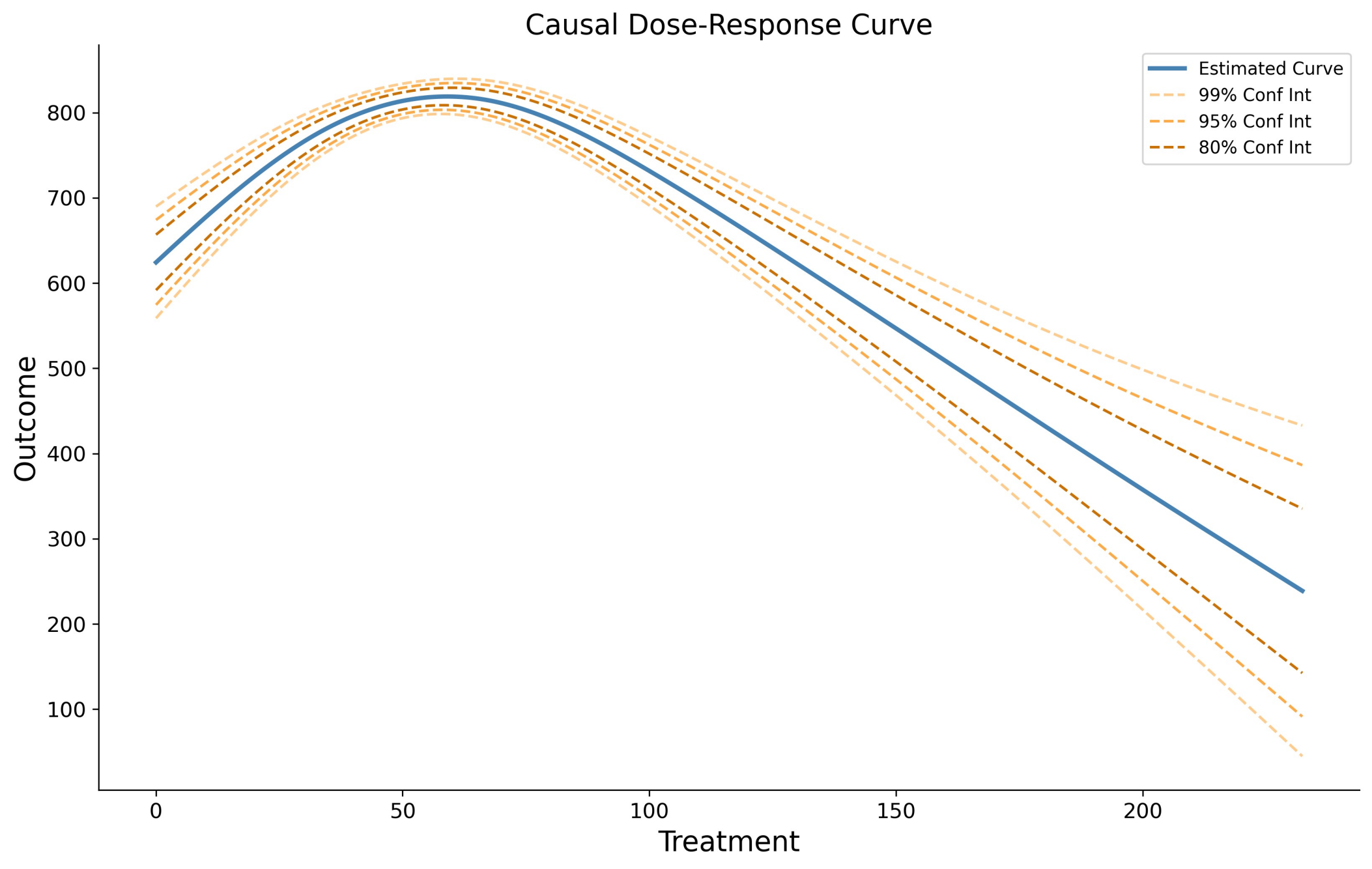
We used the following features as potentially confounding “nuisance” variables:

- Child age
- Child sex (in 1988 - 1994 the CDC assumed binary sex)
- Child race/ethnicity
- The education level of the guardian
- Whether someone smokes in the child’s home
- Whether the child spent time in a neonatal intensive care unit as a baby
- Whether the child is experiencing food insecurity (is food sometimes not available due to lack of resources?)

In the case of the math test, **these results indicate that by reducing BLLs in this population to their lowest value would cause scaled math scores to increase by around 2 points, relative to the BLLs around 10 ug/dL.** Similar results are found for the reading and block design test, although the digit spanning test causal curve appears possibly flat (although with the sparse observations at the higher end of the BLL range and the wide confidence intervals it is difficult to say).

## Conclusion

Although significant research and implementation effort has gone towards methods in causal inference to estimate the effects of binary treatments (e.g. what was the effect of treatment “A” or “B”?), much less has gone towards estimating the effects of continuous treatments. This is a shame as the number of academic and business use cases involving the analysis of continuous treatments is countless. We hope this empowers users to make sound, data-driven decisions in their day-to-day work.



## References

Galagate, D. Causal Inference with a Continuous Treatment and Outcome: Alternative Estimators for Parametric Dose-Response function with Applications. PhD thesis, 2016.

Hirano K and Imbens GW. The propensity score with continuous treatments. In: Gelman A and Meng XL (eds) Applied bayesian modeling and causal inference from incomplete-data perspectives. Oxford, UK: Wiley, 2004, pp.73–84.

Imai K, Keele L, Tingley D. A General Approach to Causal Mediation Analysis. Psychological Methods. 15(4), 2010, pp.309–334.

Kennedy EH, Ma Z, McHugh MD, Small DS. Nonparametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society, Series B. 79(4), 2017, pp.1229-1245.

Moodie E and Stephens DA. Estimation of dose–response functions for longitudinal data using the generalised propensity score. In: Statistical Methods in Medical Research 21(2), 2010, pp.149–166.

van der Laan MJ and Gruber S. Collaborative double robust penalized targeted maximum likelihood estimation. In: The International Journal of Biostatistics 6(1), 2010.

van der Laan MJ and Rubin D. Targeted maximum likelihood learning. In: U.C. Berkeley Division of Biostatistics Working Paper Series, 2006.

We acknowledge the valuable feedback from Miguel-Angel Luque, Erica Moodie, and Mark van der Laan during the creation of this project.
Prepared for SciPy 2021