Digital Public Health Seminar Series 2024

How Machine Learning Can Be Utilized in Causal Inference

Momenul Haque Mondol

PhD Student, SPPH, UBC

Co-author:

Dr. Mohammad Ehsanul Karim

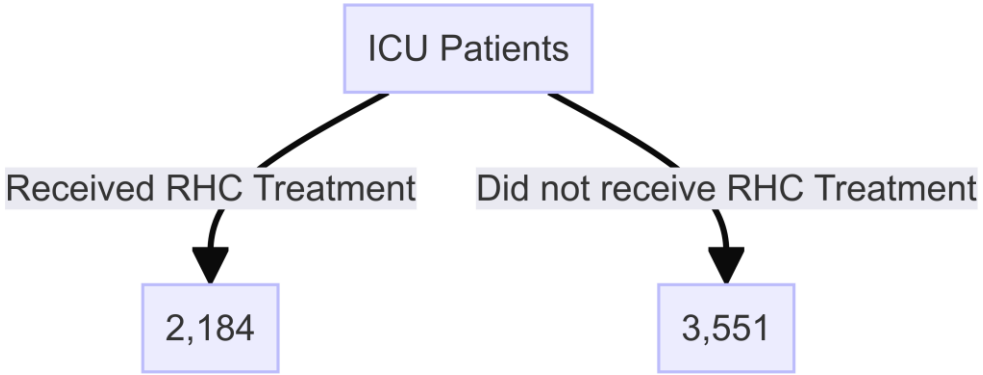
Funding:

- Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant
- UBC Work Learn program

Background

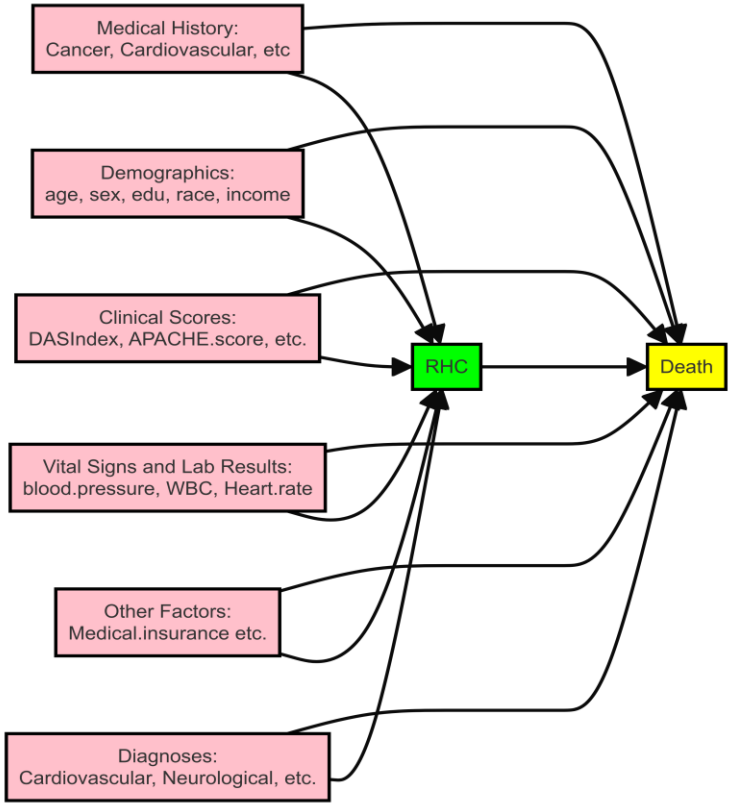
- Right heart catheterization (RHC) measures heart and pulmonary artery pressures in critically ill patients. This procedure can guide treatment decisions, improve outcomes, and potentially reduce ICU mortality risk.

- 5,735 ICU patients from USA hospitals between 1989 and 1994. (Connors et al., 1996)

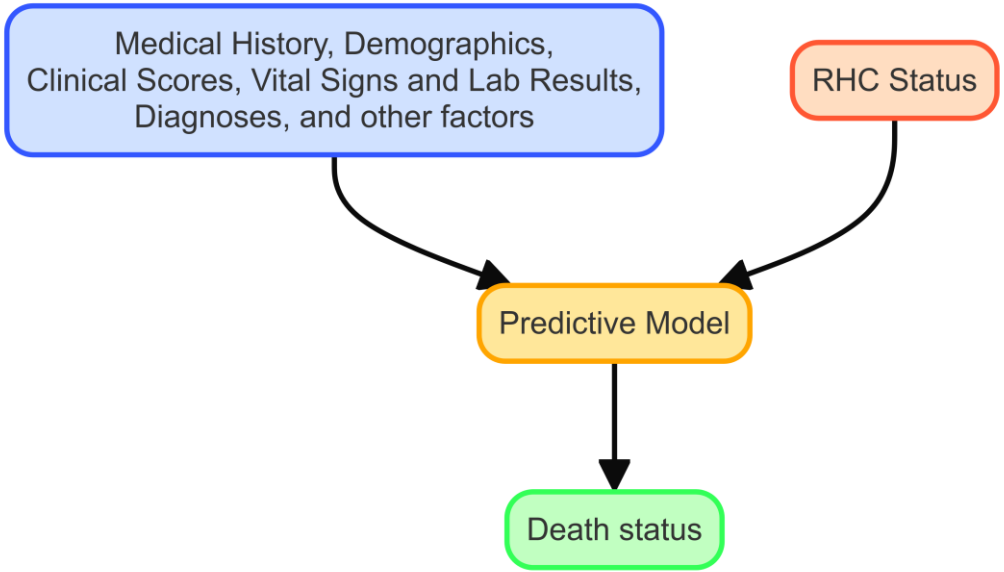


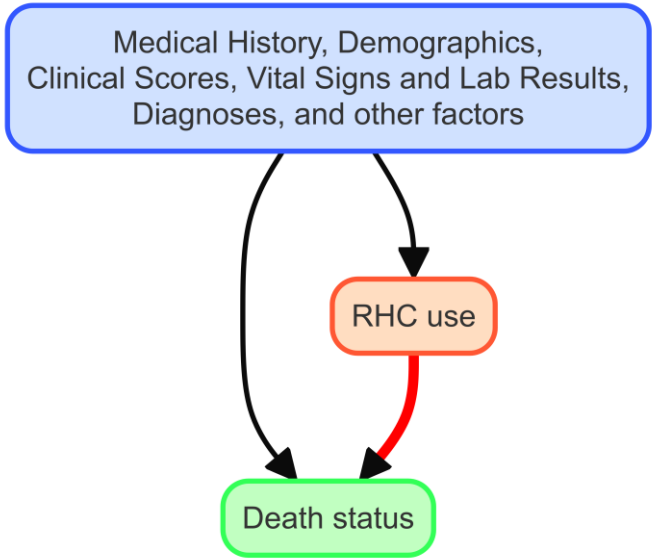
Background

- 49 characteristics of ICU patients were measured at baseline.



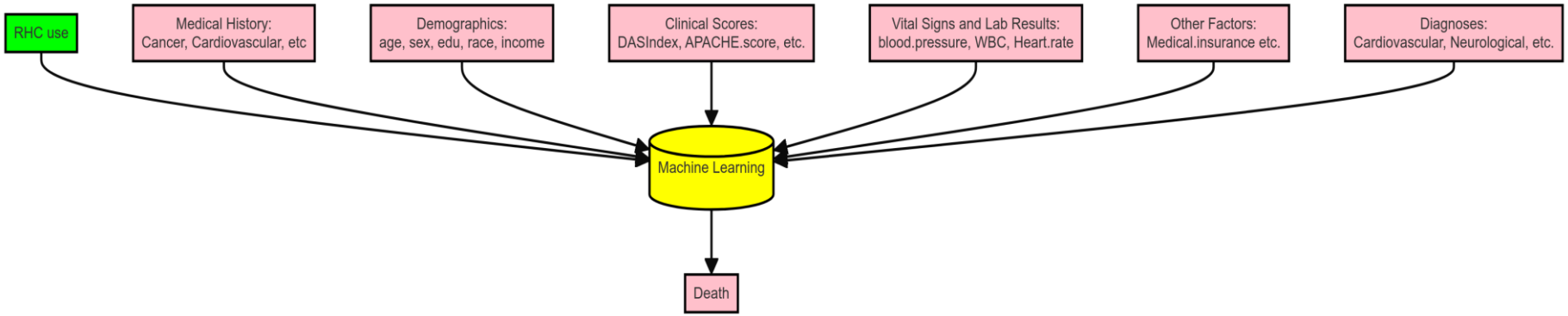
Prediction vs Causal Question





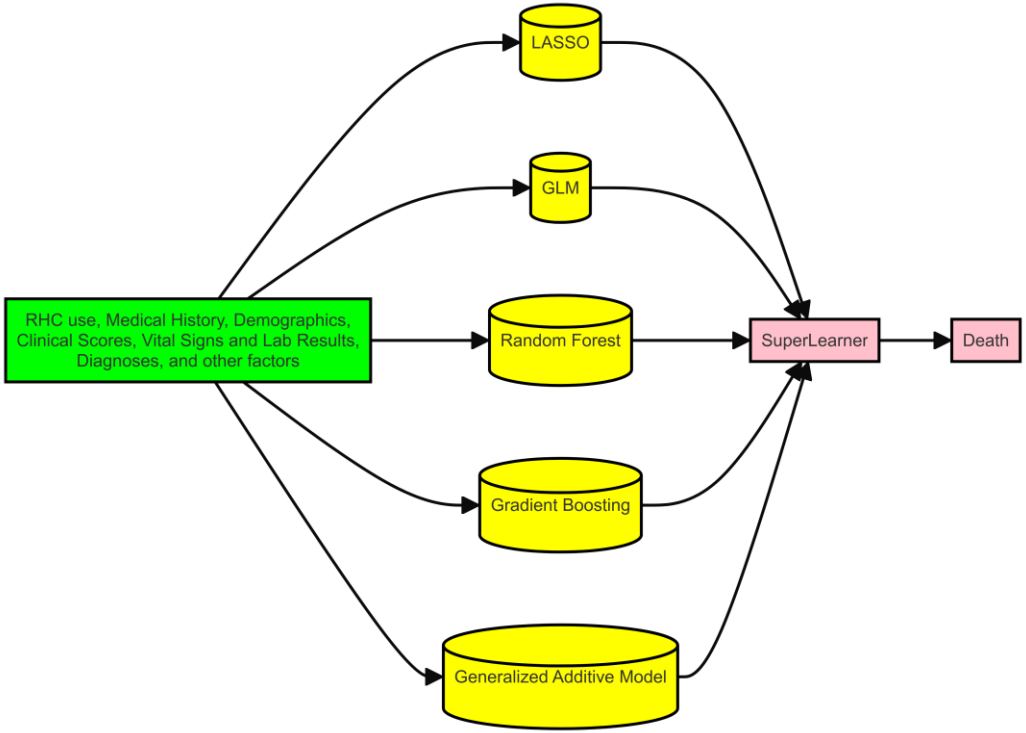
- We estimate death status (in prediction), but estimate RHC→Death association (in causal inference)

Use of Machine Learning



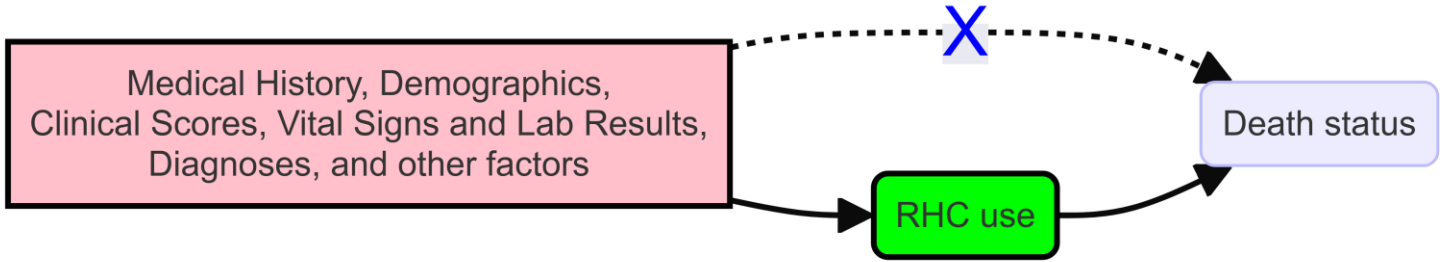
- Single machine learning algorithm may not be adequate.
- Multiple algorithms can offer more diverse learning experience (Super learner).

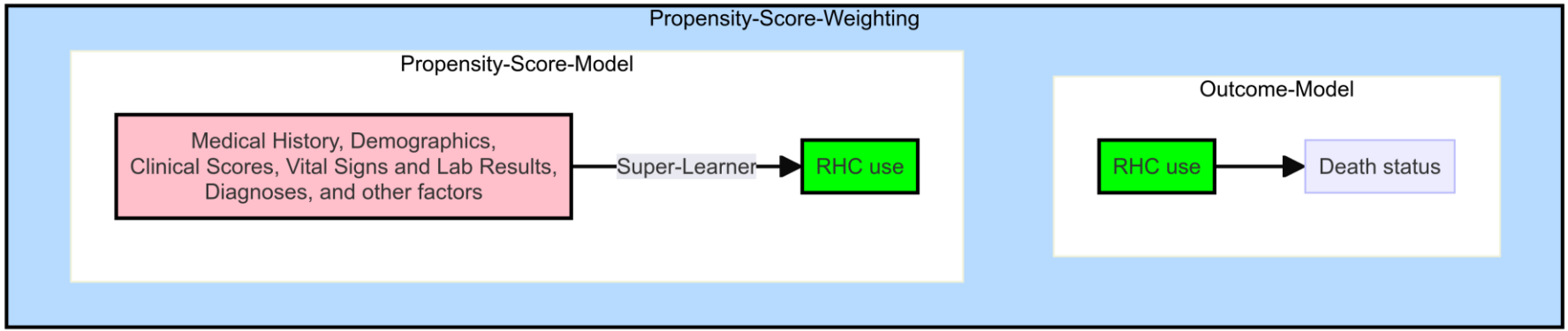
Super Learner



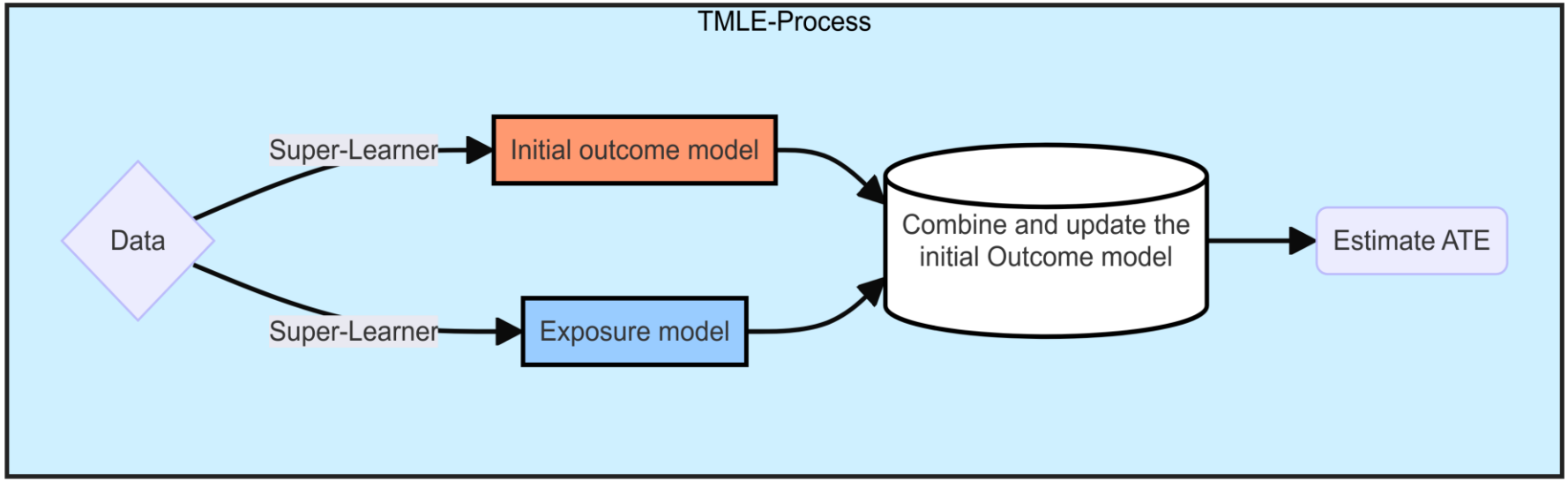
- Super-Learner is a weighted prediction of several machine learning algorithms.

Use of Machine Learning





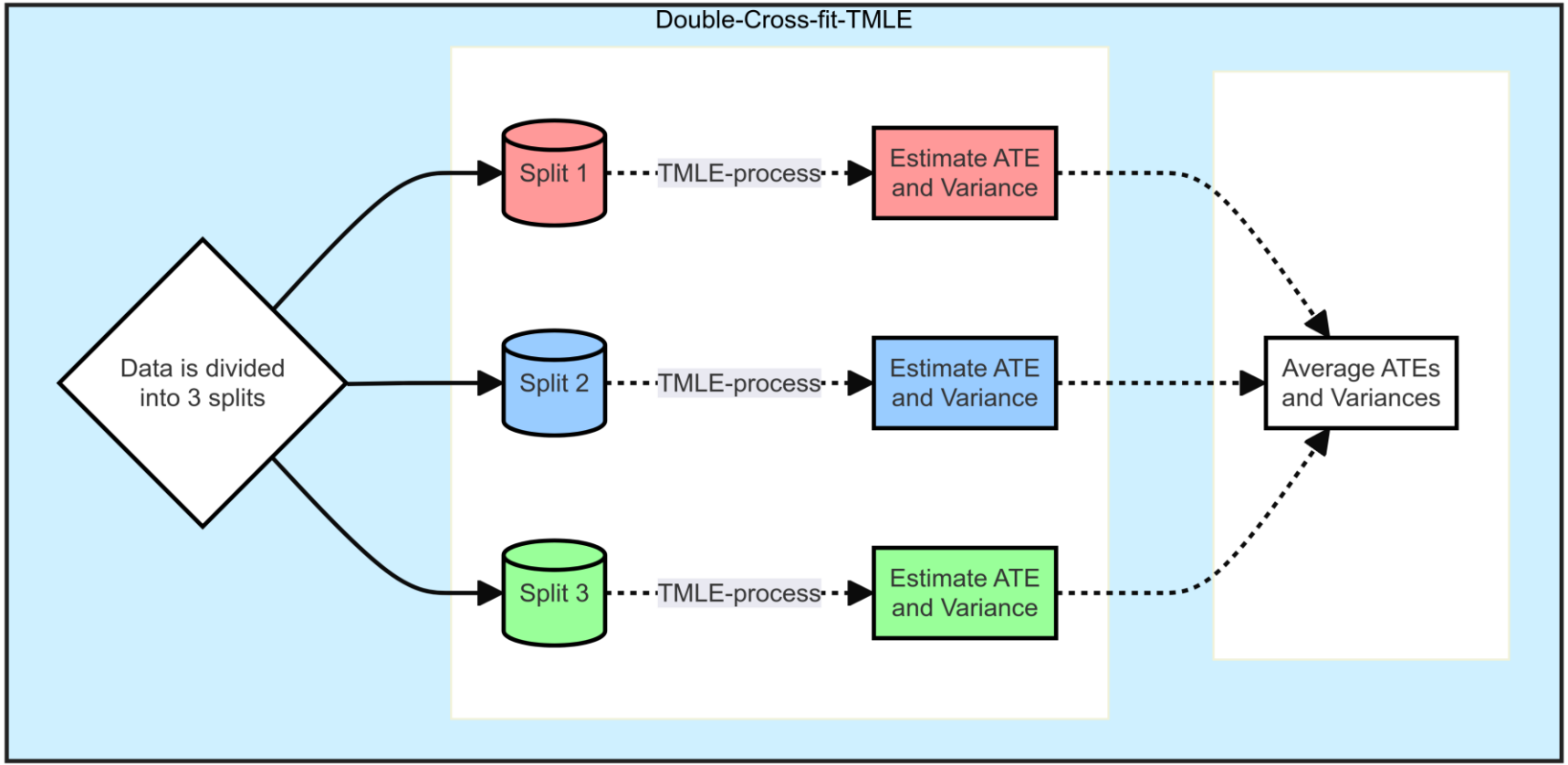
TMLE



- Same data is utilized for model development, prediction, and ATE calculation

- Flexible machine learning is prone to overfitting that may result in bias and under-coverage.(Naimi et al., 2021)

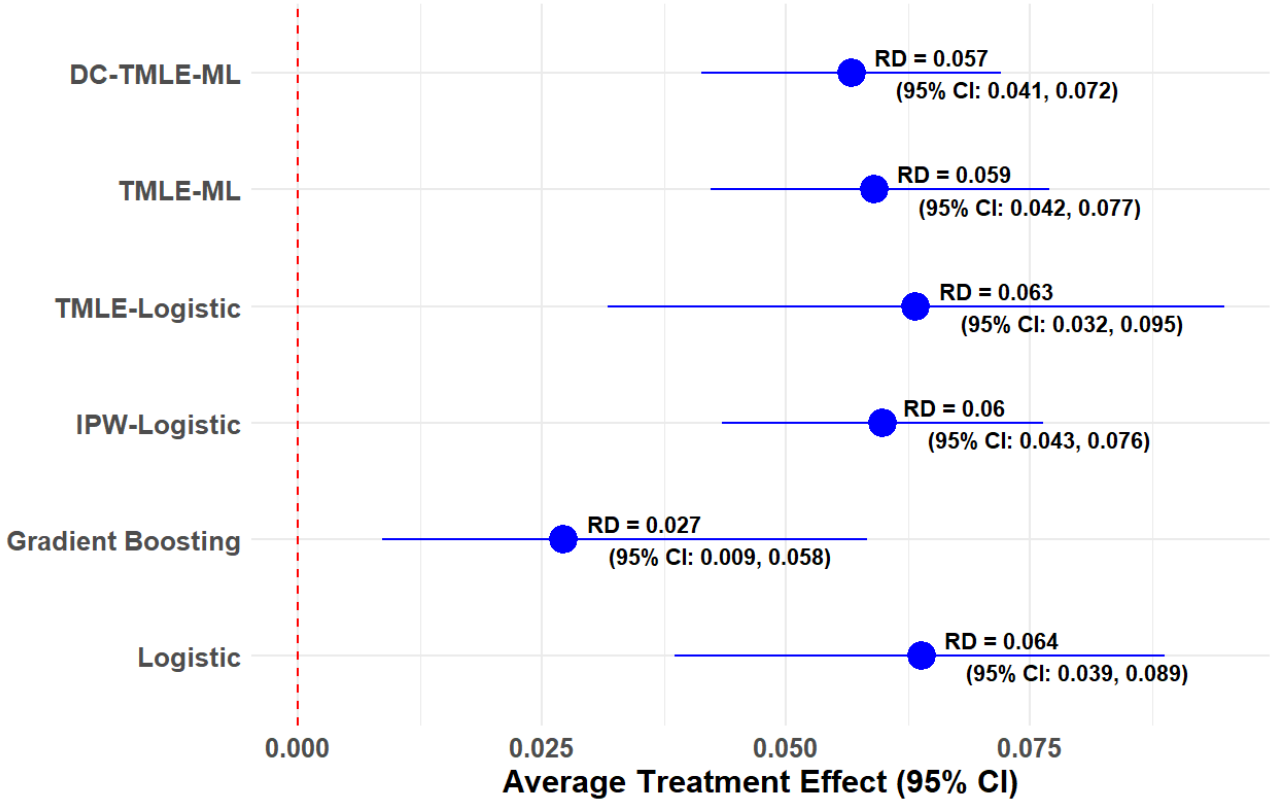
Double Cross-fit TMLE (Zivich 2021)



Apply Double Cross-fit TMLE

- Double Cross-fit TMLE can be applied by R package “crossfitTMLE”. (Mondol and Karim, 2023)
- We applied:
  - 3 splits (although >3 splits is possible)
  - LASSO, Random Forest, and Gradient Boosting
  - Repetition 100
- Estimated Risk Difference (RD) 0.057 (95% CI: 0.041, 0.072)
- The analysis took 1 hour and 36 mins to execute without parallel computing.

Comparison



ML: LASSO, Random Forest, Gradient Boosting learners under Super-Learner

Summary

Machine learning tools are alternative to standard statistical models (and sometimes better) in causal inference. However but we need to use them appropriately-

- Double robust methods (e.g., TMLE) and double cross-fit
- A diverse set of learners instead of single learner
- Include necessary interactions and polynomial terms of confounders

References

1. Description of the question and data set (DOI: 10.1001/jama.276.11.889)
2. Double cross-fit TMLE is described here (DOI: 10.1097/EDE.0000000000001332)
3. Flexible machine learning requires cross-fitting (DOI: 10.1093/aje/kwab201)
4. R package for double cross-fit TMLE (https://github.com/momenulhaque/crossfitTMLE)

Appendix

- DC-TMLE can be applied as follows (Mondol and Karim, 2023)

```
library(crossfitTMLE)
library(SuperLearner)
DC_tmle <- crossfitTMLE(data = ObsData,
                        exposure = "A",
                        outcome = "Y",
                        covarsT = L,
                        covarsO = L,
                        family.y = "binomial",
                        learners = c("SL.glmnet", "SL.randomForest",
                                     "SL.xgboost"),
                        control = list(V = 5, stratifyCV = FALSE,
                                       shuffle = TRUE, validRows = NULL),
                        num_cf = 10,
                        n_split = 3,
                        seed = 2575,
                        conf.level = 0.95, stat = "median")
```