Evaluating causal inference methods for survival data in large-scale observational studies

Shiyao Xu[1], Akihiko Nishimura[1], Elizabeth Ogburn[1]

[1] *Johns Hopkins University, Baltimore, MD 21205, USA*

June 23, 2022

Background

With the increasing availability of large-scale electronic health record (EHR) and administrative claims databases, there has been increasing interests in comparing the performance of various causal survival methods for analyzing large-scale observational data[1, 2, 3, 5, 6] under different simulation settings. However, most of the simulations were performed under relatively simplified scenarios that do not reflect the reality of observational data, and assumptions such as proportionality and random censoring were made for the data generative process that might not hold in reality. We aims to evaluate the performance of three categories of survival causal methods in estimating difference in survival probability at discrete time t in realistic simulation setting with computationally feasible algorithm implementations: inverse probability weighting (IPW), propensity score (PS) methods and the recent development of doubly robust methods. The methods considered in this study accounts for informative censoring and violations of the proportional hazards assumption.

Methods

In this section, we will briefly describe the three categories of methods under evaluation, and introduce the simulation framework. We provide both the simulation framework and the algorithms as an open-source R package `CausalSurvival` (https://github.com/zeger-nishimura-lab/CausalSurvival).

IPW[2] adjust for confounding and informative censoring by weighting via the inverse of propensity score and censoring probability. The consistency of the IPW estimator depends on the consistent estimation of both PS and censoring probability.

For PS methods, we evaluate the performance of stratified cox model in [4] that assumes confounding could be adjusted via adjusting for PS strata. To estimate difference in survival probability with stratified cox model, we performed pooled logistic regression with treatment, PS strata indicator, time and interaction between PS strata indicator and time component to allow baseline hazards to vary across PS strata. The consistency of the estimator relies on the correct specification of the baseline hazards and the assumption on confounding adjustment. We also

perform weighted pooled logistic regression by weighting observations with inverse of censoring probability to account for informative censoring. The consistency of the weighted stratified cox model further relies on the correct estimation of the censoring probability.

Doubly robust estimators combines the outcome regression and inverse weighting estimators in a way that the estimator remains consistent when either the outcome regression model or the inverse weighting estimators are estimated consistently[2, 5]. Targeted Maximum Likelihood Estimation (TMLE) and augmented IPW[2] are the two doubly robust estimation methods under evaluation in this study.

All methods require the estimation of censoring probability and PS, and doubly robust methods further require the estimation of survival probability. We focus on the practical implementation of L1/L2-regularized logistic regression model adjusted for linear combinations of baseline covariates to estimate PS, survival and censoring distribution while performing automatic variable selection for confounding adjustment for all methods.

For the simulation study, we use data from the IBM MarketScan Commercial Claims and Encounters database to compare the efficacy of two of the most common first-line hypertension treatments, ACE inhibitor and thiazide. To preserve the complex structure of confounding, we use the exposure status and baseline covariates from the real data in constructing a logistic regression model to simulate survival time. We derive empirical estimates of the baseline hazards and treatment effect used in the simulation model by fitting a L1/L2-regularized logistic regression model to the real data. Our simulation framework reserves potential informative censoring and non-proportionality in the data.

Results

The dataset includes 1,065,745 individuals, among which about 5000 individuals experienced major cardiovascular event and 7890 baseline covariates were collected. We coarsen survival time into 50 non-uniform time intervals with each containing roughly 100 events to account for low event rate and skewed survival time distributions. For the simulation study, we will compute the following set of metrics to compare and evaluate the performance of the estimation of difference in survival probability at all 50 discrete time points: mean bias, square root variance, estimated asymptotic variance, coverage of 95% pointwise confidence intervals and width of confidence interval at 95% coverage.

Conclusion

We implemented state-of-art causal survival methods that accounts for informative censoring and non-proportionality, and demonstrated the advantages of doubly robust methods in obtaining consistent estimations. We also provide a simulation framework that simulate realistic large-scale

observational data to evaluate the performance of various methods. This work adds to the OHDSI methods library in population-level estimation and explore alternative causal estimands such as difference in survival probabillity and difference in restricted mean survival time.

References

[1] Robin Denz, Renate Klaaßen-Mielke, and Nina Timmesfeld. A comparison of different methods to adjust survival curves for confounders, 2022. URL `https://arxiv.org/abs/2203.10002`.

[2] Iván Díaz, Elizabeth Colantuoni, Daniel F. Hanley, and Michael Rosenblum. Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime Data Anal*, 25:439–468, 2019. doi: https://doi.org/10.1007/s10985-018-9428-5.

[3] Xiaoqing Tan, Shu Yang, Wenyu Ye, Douglas E. Faries, Ilya Lipkovich, and Zbigniew Kadziola. When doubly robust methods meet machine learning for estimating treatment effects from real-world data: A comparative study, 2022. URL `https://arxiv.org/abs/2204.10969`.

[4] Yuxi Tian, Martijn J Schuemie, and Marc A Suchard. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International Journal of Epidemiology*, 2018. doi: 10.1093/ije/dyy120.

[5] Ted Westling, Alex Luedtke, Peter Gilbert, and Marco Carone. Inference for treatment-specific survival curves using machine learning, 2021. URL `https://arxiv.org/abs/2106.06602`.

[6] Donglin Zeng. Estimating marginal survival function by adjusting for dependent censoring using many covariates. *The Annals of statistics*, 32(4):1533–1555, 2004. doi: 10.1214/009053604000000508.