Efficiency gains in randomized trials for COVID-19 using specific covariates and performing variable selection with LASSO and random forests

Nicholas Williams, Iván Díaz, Michael Rosenblum

March 17, 2021

Abstract

1 Introduction

1.1 Covariate adjustment in RCTs

(Benkeser et al.)

2 Statistical Estimands and Estimators

(van der Laan and Rubin 2006) (Díaz et al. 2019) (Díaz et al. 2016)

- double-robust estimators
- consistent under correct specification of either treatment or outcome, guaranteed with the RCT.
- can use the LASSO and still get valid standard errors

2.1 Survival outcome

For a time-to-event outcome, we consider two measures of treatment effect: the difference in the restricted mean survival time (RMST) and the difference in survival probability (also referred to as the risk difference, RD). Let A be a binary treatment indicator where 1 indicates assignment to the treatment arm and 0 and indicates assignment to the control arm. In addition, let T denote a patients survival time.

Estimand 1 (Difference in restricted mean survival time)

The difference in restricted mean survival time is defined as the difference in the expected survival time between treatment and control arms when time is truncated at an arbitrary point, τ.

$$E(\min\{T,\tau\}|A=1) - E(\min\{T,\tau\}|A=0)$$

Estimand 2 (Difference in survival probability)

$$P(T > t|A = 1) - P(T > t|A = 0)$$

2.2 Ordinal outcome

We consider two treatment effect estimands for an ordinal outcome: the average log-odds ratio
(LOR) and the Mann-Whitney (MW) estimand. We define an ordinal outcome taking levels $1, ..., K$
and without loss of generality, we will assume that lower values of the outcome are preferred.

Estimand 1 (Average log-odds ratio)

If the proportional odds assumption is correct, this value is equal to the corresponding coefficient
from a proportional odds model (commonly referred to as ordinal logistic regression).

$$\frac{1}{K-1} \sum_{j=1}^{K-1} \log \left(\frac{\text{odds}(Y \leq j|A = 1)}{\text{odds}(Y \leq j|A = 0)} \right)$$

Estimand 2 (Mann-Whitney)

The Mann-Whitney statistic is interpreted as the probability that a patient drawn at random
from the treatment group has a greater outcome value than a patient drawn at random from the
control group, with ties broken at random.

$$P(Y^* > Y|A^* = 1, A = 0) + \frac{1}{2} P(Y^* = Y|A^* = 1, A = 0)$$

3 Simulation Methods

Our data generating distribution is based on a database of over 1,500 patients hospitalized at
Weill Cornell Medicine New York Presbyterian Hospital prior to 15 May 2020. For a description
of the clinical characteristics and data collection methods of the initial cohort sampling,
see Goyal et al. (2020).

For both outcome types, we evaluated the following auxillary covariates: age, sex, BMI, smoking
status, whether the patient required supplemental oxygen within three-hours of presenting to
the Emergency Department, number of comorbidities (diabetes, hypertension, COPD, CKD, ESRD,
asthma, interstitial lung disease, obstructive sleep apnea, any rheumatological disease, any
pulmonary disease, hepatitis or HIV, renal disease, stroke, cirrhosis, coronary artery disease,
active cancer), number of relevant symptoms, presence of bilateral infiltrates on chest x-ray,
dyspnea, and hypertension. Patient data were resampled with replacement to generate 1000 datasets
of size $n = 100, 250, 500$ and 1500. For each dataset, a treatment value was draw for each patient
with probability 0.5 from a Bernoulli distribution.

For each generated dataset, we evaluated the estimands of interest with the following conditions:
unadjusted; adjusting for each individual covariate; adjusting for all covariates; adjusting
for all covariates using the LASSO (Tibshirani 1996); adjusting for all covariates using random
forests. In scenarios using the LASSO or random forest, all nuisance parameters were estimated
with LASSO while only the outcome and censoring models were estimated using random forests and
an intercept only model was used to estimate the propensity. In addition, the use of random
forests was evaluated with and without cross-fitting (Chernozhukov et al. 2018). All simulation
scenarios were also repeated after shuffling covariates to remove any association between the
covariate and observed outcomes.

3.1 Survival outcome

For the survival outcome, the outcome is defined as the time from hospitalization to intubation or death. We focused on estimating the difference in RMST at 14 days after hospitalization and the difference in survival probability at 7 days after hospitalization. A positive treatment effect was simulated by adding an independent random draw from a χ^2 distribution with varying degrees of freedom (to produce different effect sizes) to each patient's observed survival time in the treatment arm. These effect sizes translate to a difference in RMSTs of 0.55 and 1.07, and RDs of 0.04 and 0.10, respectively. To simulate uninformative censoring, 5% of the patients were selected at random to be censored with censoring times drawn from a Uniform distribution, $U(1, 14)$.

3.2 Ordinal outcome

We defined a six-level ordinal outcome at 14 days post-hospitalization based on the WHO Ordinal Scale for Clinical Improvement (INSERT CITATION). The categories are as follows: 0, discharged from hospital; 1, hospitalized with no oxygen theraphy; 2, hospitalized with oxygen by mask or nasal prong; 3, hospitalized with non-invasive ventilation or high-flow oxygen; 4, hospitalized with intubation and mechanical ventilation; 5, dead.

A positive treatment effect was simulated by subtracting from the treatment arm an independent random draw from a four-parameter Beta distribution with support $x \in [0,5]$, Beta$(\alpha, 15)$; we sampled from distributions with the α parameter set to 1.5 and three to generate different effect sizes. These effect sizes corresponds to an LOR of 0.22 and 0.60, an a Mann-Whitney value of 0.49 and 0.46 respectively.

3.3 Performance criteria

4 Simulation Results

Tables containing the comprehensive results of simulations are presented in the Appendix.

- Supplemental O2 provides the greatest efficiency gains

- As sample size increases, the LASSO and RF provides largest efficiency gains

- Without variable selection, adjusting for non-prognostic variables results in efficiency losses

- The LASSO maintains efficieny when covariates are not predictive of the outcome

- The LASSO and RF handle zeros in an ordinal outcome while GLM becomes unstable

- Without cross-fitting, RF overfits but remains un-biased because the propensity is still correctly specified

- With cross-fitting, RF...

5 Recommendations

6 References

References

Benkeser, D., Díaz, I., Luedtke, A., Segal, J., Scharfstein, D., and Rosenblum, M. Improving precision and power in randomized trials for covid-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1-C68.

Díaz, I., Colantuoni, E., Hanley, D. F., and Rosenblum, M. (2019). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime data analysis*, 25(3):439-468.

Díaz, I., Colantuoni, E., and Rosenblum, M. (2016). Enhanced precision in the analysis of randomized trials with ordinal outcomes. *Biometrics*, 72(2):422-431.

Goyal, P., Choi, J. J., Pinheiro, L. C., Schenck, E. J., Chen, R., Jabri, A., Satlin, M. J., Campion, T. R., Nahid, M., Ringel, J. B., Hoffman, K. L., Alshak, M. N., Li, H. A., Wehmeyer, G. T., Rajan, M., Reshetnyak, E., Hupert, N., Horn, E. M., Martinez, F. J., Gulick, R. M., and Safford, M. M. (2020). Clinical characteristics of covid-19 in new york city. *New England Journal of Medicine*, 382(24):2372-2374.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267-288.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).