College of Health

COMPARING LOGISTIC REGRESSION AND SUPERLEARNER FOR CAUSAL INFERENCE ON OBSERVATIONAL DATA

Carson Mowrer

Table 1: Patient Characteristics

Characteristic	**No RHC**, N = 3,551[1]	**RHC**, N = 2,184[1]
Age	61.76 (17.29)	60.75 (15.63)
Sex		
Female	1,637 (46%)	906 (41%)
Male	1,914 (54%)	1,278 (59%)
Disease Category		
ARF	1,581 (45%)	909 (42%)
CHF	247 (7.0%)	209 (9.6%)
Cirrhosis	175 (4.9%)	49 (2.2%)
Colon Cancer	6 (0.2%)	1 (<0.1%)
Coma	341 (9.6%)	95 (4.3%)
COPD	399 (11%)	58 (2.7%)
Lung Cancer	34 (1.0%)	5 (0.2%)
MOSF w/Malignancy	241 (6.8%)	158 (7.2%)
MOSF w/Sepsis	527 (15%)	700 (32%)
Cancer		
Metastatic	261 (7.4%)	123 (5.6%)
No	2,652 (75%)	1,727 (79%)
Yes	638 (18%)	334 (15%)
Immunosupression	907 (26%)	636 (29%)
Medical Insurance		
Medicaid	454 (13%)	193 (8.8%)
Medicare	947 (27%)	511 (23%)
Medicare & Medicaid	251 (7.1%)	123 (5.6%)
No insurance	186 (5.2%)	136 (6.2%)
Private	967 (27%)	731 (33%)
Private & Medicare	746 (21%)	490 (22%)

[1] Mean (SD); n (%)

INTRODUCTION

While randomized controlled trials (RCTs) are often considered the gold-standard for establishing causality, they are ethically or practically challenging to conduct for many relationships of interest. Particularly as rich observational datasets, such as electronic medical records, are increasingly available for research purposes, methods for estimating causal effects from observational data are gathering attention. This analysis aims to demonstrate the use of propensity score matching (PSM), a methodology for reducing the impact of confounding in observational data and allowing for estimates of causal effects, and to compare results using traditional logistic regression and the ensemble machine learning algorithm SuperLearner.

METHODS

- Dataset of medical records of 5,735 adults receiving care in an ICU between 1989 and 1994
- Treatment of interest: right heart catheterization (RHC)
- Outcome of interest: length of stay in the hospital
- A selection of patient demographics are shown in Table 1, stratified by treatment group

Three analytic approaches:

- MLR with no PSM
- MLR with logistic regression-based PSM
  - 1:1 matching, 0.2 times logit of SD of PS caliper width
- MLR with SuperLearner-based PSM
  - Base learners: GLM, regularized GLM, Random Forest, Bayesian additive regression tree classifier, regularized gradient boosting

RESULTS

Figure 2
Propensity Scores using Logistic Regression

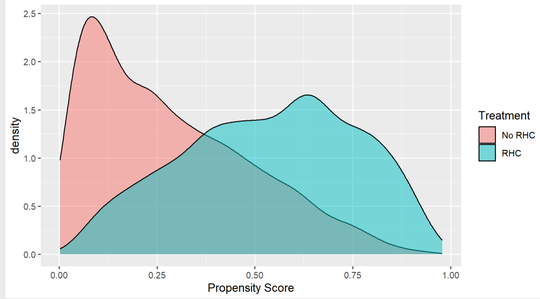


Figure 3
Propensity Scores using SuperLearner

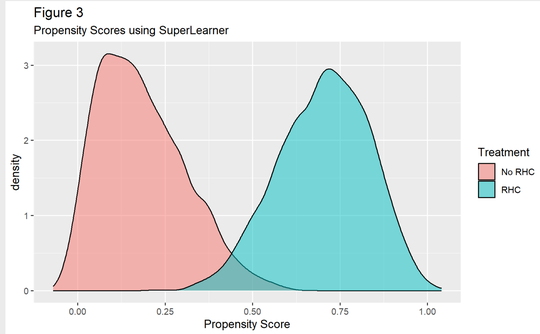


Figure 4
Covariate Balance with Logistic Regression and SuperLearner

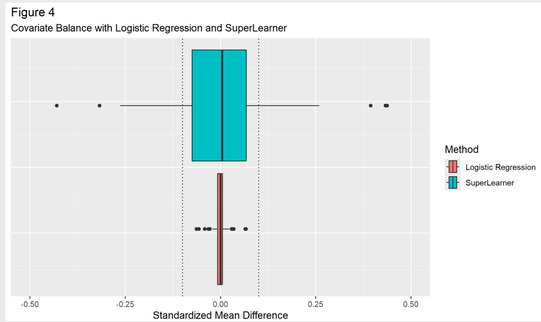
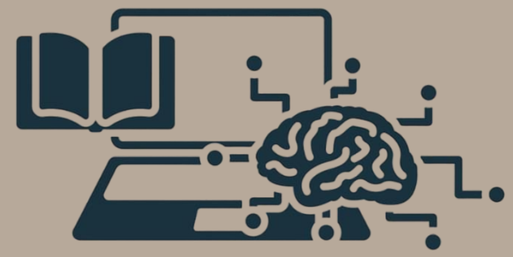


Table 2	No PSM	PSM with Logistic Regression	PSM with SuperLearner
Effect of RHC on Length of Stay (no covariate adjustment)	5.33 days (3.96, 6.70)	4.02 days (2.21, 5.83)	-2.90 days (-8.94, 3.15)
Effect of RHC on Length of Stay (with covariate adjustment)	2.84 days (1.31, 4.36)	3.41 days (1.66, 5.16)	-0.48 days (-9.27, 8.32)



DISCUSSION

- Propensity score estimation using logistic regression resulted in considerable overlap in the distribution of propensity scores, suggesting a large set of exchangeable patients
- Estimation using SuperLearner shows less overlap, indicating a smaller exchangeable population
- Covariate balance was significantly better using logistic regression
  - 69 of 69 covariates balanced with logistic regression
  - 43 of 69 covariates balanced with SuperLearner
- SuperLearner may be detecting key non-linear relationships and interactions not captured by logistic regression
- Covariate balance is assessed on each variable in isolation, meaning the complex relationships identified by SuperLearner may not be well assessed by individual covariate balance
- These results suggest the dataset may lack enough truly exchangeable individuals, leading to reduced power and wider confidence intervals using SuperLearner-based matching

REFERENCES

Data obtained from http://hbiostat.org/data courtesy of the Vanderbilt University Department of Biostatistics

Oregon State University