Cognitive Remediation in Schizophrenia: Exploring Treatment Effect Heterogeneity to Improve Treatment Success

Ngoc Duong
Faculty Advisor: Caleb Miles, PhD

Introduction

Cognitive remediation therapy (CR) is a type of intervention implemented to target cognitive dysfunctions in patients with schizophrenia. Studies have been conducted to understand the effect of CR on cognitive outcomes, and its overall positive effect has been documented in reducing cognitive deficits of the disorder.[1,5,12,14,15] However, the degree of effectiveness, especially on functional and occupational outcomes, has not been shown to be consistently significant.[4,12,14,15] Therefore, we might have reason to believe that treatment effects could vary across strata of patients with respect to their characteristics or treatment environments and modifications, especially for functional and occupational outcomes.[3,11,13] Additionally, although some studies have noted positive effects of improved memory and cognitive processes on social functioning outcomes,[14,15] the mechanism(s) of how CR would affect these outcomes were still not studied carefully. We started out this project with three aims. First, we explored prognostic factors that might affect social functioning outcomes at follow-up. Second, we assessed any heterogeneity in treatment effects by estimating the conditional average treatment effect (CATE) and identify potentially important effect modifiers, which might help identify patient profiles that respond differently to the treatment. Lastly, we estimated the optimal individualized treatment rule and its value (the expected outcome under the implementation of this rule).

Methods

Data

We used data from three clinical trials in the NIMH Database of Cognitive Training and Remediation Studies. The first trial by Wykes et al. (1999) looked at CR

treatment effect on primary outcomes like memory and cognitive flexibility as well as other social behavioral measures. The other two trials by Wykes et al. in the same year of 2007 investigated similar outcomes, where functioning outcome such as SBS was treated as secondary outcome or as an outcome of changes in cognitive functions. All three studies were conducted in England.

Data preprocessing steps included: 1) dropping variables with more than 50% observations missing and with at least 95% of values that are similar (low variance), 2) converting variables with fewer than twelve unique observations to factors, 3) imputing missing values using Random Forest (using information from the whole feature space excluding the outcome of interest), and finally, min-max normalization to the range of 0-1 for all predictors. After preprocessing, the analytical sample size consisted of 142 patients from the original 160. The Social Behavioral Scale (SBS) total score at follow-up was used as the outcome of interest (mean = 10.9, median = 8, SD = 8.8). Higher scores indicate more problematic behaviors. The treatment variable is binary (1 for treatment and 0 for no treatment). In this sample, there were 75 randomized to receive treatment and 67 randomized to receive no treatment. There were 97 predictors including demographic variables (age, sex, race, education, parents' education level, age of psychiatric symtom onset, etc.), and items from different scales used to measure symptoms and assess individuals with schizophrenia: Social Behavioral Scale (SBS), Positive and Negative Syndrome Scale (PANSS), The Brief Psychiatric Rating Scale (BPRS), Scale for the Assessment of Negative Symptoms (SANS) and Scale for the Assessment of Positive Symptoms (SAPS), among some other scales.

Statistical analyses

For the first aim, we created a predictive model of the total SBS score at follow-up and identified important prognostic factors using the R package SuperLearner.7 The CV.SuperLearner function implements a V-fold Nested Cross-Validation framework, which provides a way to maximize the use of information in this dataset while preventing overly optimistic estimation of the error. The package also constructed a weighted SuperLearner, which could improve on individual learners' performances (under certain settings) by combining multiple estimators into an improved estimator.8 We evaluated the cross-validated loss of all individual learners specified in the library together with the SuperLearner itself to determine the best performing learner. The library consisted of the following learners: the marginal mean predictor, the LASSO, Elastic Net, Random Forest, XGBoost, BART Machine, Support Vector Regression, and MARS. Among these, the LASSO, BART Machine, RF, and SVR models were tuned using caret before being evaluated in competition with other learners. Additionally, we also customized some learners with pre-specified hyperparameters such as Elastic Net (mixing parameters 0.325 and 0.1), Random Forests (number of randomly selected variables mtry set to be 24 and 32 with 1000 trees), and XGBoost (max depth of 1 and 2, step size shrinkage 0.001, with 3000 trees). A variable screening step using $p < 0.1$ for the pairwise correlation of each predictor with the outcome

was also used. After picking the best performing learner, we then looked at variable importance under this learner and partial dependence plots (PDP) of most important predictors.

For the second aim, we first estimated the ATE by taking the difference in means between the treated and untreated group, which should be unbiased given the randomized treatment assignment. On first assessment of the effect size and its statistical significance, a small ATE does not necessarily imply there is little or no treatment effect for anyone at all. We then would be wondering if this was the case or if there could be treatment effect heterogeneity. We attempted to explore the heterogeneous treatment effects by estimating the CATE, which required making individual-level predictions of both counterfactual outcomes given a set of baseline individual-specific covariates. Specifically, the predicted counterfactual outcome for individual i under treatment can be attained by running the best learner in the treated group $\widehat{Q_1}$ on their set of covariates W_i, and similarly, we can use the best learner within the untreated group $\widehat{Q_0}$ to estimate the same individual's counterfactual outcome under no treatment:

$$\widehat{Q_1}(W_i) = \hat{E}(Y_i|A_i = 1, W_i)$$
$$\widehat{Q_0}(W_i) = \hat{E}(Y_i|A_i = 0, W_i)$$

Accordingly, we can estimate the CATE for each patient by simply taking the difference between their counterfactual outcomes had they been treated versus had they not been treated:

$$\widehat{CATE} = \hat{E}(Y_{i1}|W_i) - \hat{E}(Y_{i0}|W_i)$$
$$= \hat{E}(Y_i|A_i = 1, W_i) - \hat{E}(Y_i|A_i = 0, W_i)$$

This framework of deriving counterfactual outcomes is possible given the assumptions of exchangeability and positivity, which are satisfied under randomization, and SUTVA, which is reasonable since it is unlikely any patient's outcomes would depend on others' treatment. Then, the CATE distribution can help assess any heterogeneity underlying the small observed average treatment effect. Once we have specified an estimate of the CATE distribution based on $\widehat{Q_1}$ and $\widehat{Q_0}$ and individual-specific set of baseline covariates Wi, we can again assess variable importance and use PDP to explore the relationships between important predictors and the CATE.

For the last aim - estimation of the optimal ITR, a reasonable way to determine this from the estimated CATE is:

$$\widehat{d(W)} = I(\widehat{CATE(W)} < 0)$$

In other words, if the CATE for a particular patient is less than 0, which means treatment is effective, then the patient should be assigned to treatment, and to no treatment otherwise.

Then the value of the optimal ITR can be defined as follows:

$$V(\hat{d}) = E(Y^{\hat{d}}) = E(\frac{\hat{d}(W)A}{\pi(W)}Y + \frac{[1-\hat{d}(W)](1-A)}{1-\pi(W)}Y$$

which is unbiased when the propensity score of being assigned treatment $\pi(W)$ is known, and given the randomization setup, this value is approximately 0.5 (or 0.528 in this sample). According to this, the outcome of those with the same recommended and actual treatment assignments were upweighted while those who got different recommendations from their actual assignments received a weight of 0.

However, overfitting might occur as we used the best-performing learner to predict on the original sample, a large portion of which was used as training data. This might in turn, yield an overly optimistic result. To prevent this, we used sample splitting to prevent overfitting and devise the treatment rule accordingly. The idea is to only predict on unseen data (or held-out data) by taking advantage of the nested cross-validation framework. The estimation process under sample splitting can be summarized in the following steps:

The estimation process under sample splitting can be summarized in the following steps:

1. Perform 5-fold Nested Cross Validation and determine the learner with the lowest CV loss across 5 CV rounds
2. Use the learner created in each validation round to predict the outcome on the test data instead of using one single best-performing learner to make predictions.
3. Repeat 1-2 on the untreated group and obtain the CATE distribution from both learners' predictions
4. Devise the optimal ITR using [*]

In short, from the process above, the version of ITR we are using to estimate the value is:

$$\hat{d}_{(-i)} = I(\hat{Q}_{(-i)}(A=1, W_i) < \hat{Q}_{(-i)}(A=0, W_i)], i \in 1, 2, 3, 4, 5$$

where $(-i)$ indicates the held-out fold that was not used for training and cross-validation. We can test whether the difference in means between the value of ITR and the outcome under random treatment, all treatment, and no treatment would be different from 0 using a simple statistical test such as a paired t-test

which corresponds to the Wald test based on an application of the central limit theorem.

Statistical learning models and visualizations were carried out using packages SuperLearner and $DALEX^2$ in R (v.4.0.1).

Results

For the first aim, a comparison of the learners in the library can be found in Figure 1. The best performing learner was found to be MARS, which had 5-fold cross-validated MSE of 63.16 (SE = 10.53). For this learner, the most and only important variable, as indicated in the variable importance plot Figure 2 was total SBS score at baseline. Among other learners that had relatively good performances like Elastic Net and Random Forest, total baseline SBS score was also consistently ranked as most important. The PDP showed the positive marginal effect of total SBS score at baseline on the outcome Figure 3 across learners, which is reasonable due to the positive correlation between the baseline and follow-up values. Overall, for MARS, the PDP takes the form of a piecewise linear function – up until a certain point, SBS score at baseline had no relationship with the outcome at follow-up, after which point the relationship gets much steeper. Qualitatively, this might suggest SBS score at baseline could be a strong prognostic factor for more severe patients, while being less so for more stable patients at baseline.

In the second aim, the estimated ATE in this sample is 0.65. The results from independent-sample t-test gives 95% CI = [-2.23, 3.59], p = 0.65. Considering higher SBS score indicates worse outcome, the positive sign in the estimate suggests a harmful effect associated with treatment qualitatively, although the effect size might be deemed clinically small (considering the scale of the SBS total score) on top of being statistically insignificant. This ATE, corroborated by finding from existing literature, warrants the speculation of potential effect heterogeneity through the examination of the CATE distribution. Among the treated, we used Elastic net with mixing parameter alpha 0.1 with CV MSE 56.44 and relatively low variability. Among the untreated group, MARS seemed to be the best learner with the lowest CV MSE of 74.1 (Figure 4). Using counterafactual outcome predictions from these two learners, we estimated the CATE distribution (Figure 5), which had a mean of 1.24 (median = 1.02, SD = 3.82). The spread in the estimated CATE distribution might suggest heterogenous treatment effects underlying the small observed average treatment effect, although a robust statistical test is necessary to make a statistically sound claim. We also found that total SBS score at baseline was ranked among the most important variables in explaining $\widehat{CATE}$, and PDP showed a negative marginal effect between scaled SBS score at baseline and $\widehat{CATE}$ (Figure 6). We can see that for higher value of total baseline SBS score, the estimated CATE decreases, which

might indicate larger treatment effect for more severe patients. Accordingly, this might suggest that treatment effect for patients with lower SBS score at baseline might be different from the treatment effect for patients with higher SBS score at baseline. Overall, this could be informative to determine the optimal treatment rule if limited information about the patients at baseline is available to clinicians.

For the third aim, under sample splitting, the estimated CATE suggested the following estimated optimal ITR - 78 recommended to no treatment 64 recommended to treatment. The table below shows the difference between the estimated optimal ITR and the actual random assignment:

	Randomized to no treatment	Randomized to treatment
Recommended no treatment	31	47
Recommended treatment	36	28

We found the outcome under the implementation of the estimated optimal ITR had a mean of 9.3, in comparison with the mean observed outcome under randomization of 10.9. This gives the difference in means of 9.3-10.9 = -1.6. Qualitatively, the negative sign in the estimate indicates a reduction in mean total SBS at follow-up under this rule. However, we tested the hypothesis that the difference is different from 0, and as shown through 95% CI = [-4.51, 1.39] and p = 0.29, the difference in means is not statistically significant at 5% significance level. This suggests a lack of statistical evidence for treatment effect heterogeneity in this sample. We also compared this value with the outcome had everyone been treated: ES = -2, 95% CI = [-4.45, 0.45], p = 0.11, and with the outcome had everyone not been treated: ES = -0.89, 95% CI = [-3.35, 1.57], p = 0.48. The conclusions were largely the same, and considering the range of the total score, this also might not be clinically significant, although more domain knowledge is necessary to determine this. In the absence of treatment effect heterogeneity, our devised optimal ITR did not imply superiority over randomly treating people, treating everyone, or treating no one.

Discussion

Identifying subgroups of patients who benefit differently from CR is important in optimizing medical resource allocation, as the treatment itself is both costly and demanding. Exploring the relationships between certain covariates and outcome of interest or treatment effect through variable importance and partial dependence plot might be helpful. Clinicians could use this information to tailor treatment targeting particular prognostic factors that can have large effects on the outcome of interest. This could be another way to help clinicians given limited information about a patient. However, as it is difficult to visualize partial

dependence plots of more than two predictors simultaneously, this one-covariate-at-a-time approach might be less straightforward in terms of informing clinical decision-making when more information about patients is available, or when the intersection of covariates gets more granular. For example, the level of one covariate might suggests treatment but the value of another covariate might suggest otherwise.10 In this project, we also found lack of evidence for treatment effect heterogeneity. However, the results could be inconclusive due to lack of power, variabilities in estimations/predictions, and the inability to capture effect heterogeneity by unobserved variable(s).

Strengths and Limitations

A strength of the project is that data came from clinical trials, which guaranteed conditions held that are needed for identification of estimands in the counterfactual framework. Using the SuperLearner package, we also compared a variety of non-parametric statistical learning methods with parametric models to improve predictive performances in the event of potential violations of modeling assumptions. Although we found no statistical evidence for treatment effect heterogeneity, the results are still not conclusive, as the project does have its limitations that might affect the power and validity of the findings. Since we used data from more than one clinical trial, a big challenge was systematic missing data across studies. We used random-forest-based imputation to address this, which might induce some variability in imputed values and in turns, predictions.

In terms of future directions, we could include more data from similar clinical trials to address the limitation of the potential lack of power. Assuming there is heterogeneity of treatment effects, more generalizable and interpretable optimal individualized treatment rule can be devised. Finally, since we only explored heterogeneity in treatment effects as a function of baseline covariates, adding a more longitudinal component (another timepoint post-treatment but before the social functioning outcome was measured) can offer more insights into the treatment mechanism. Further investigations including mediation analyses treating cognitive outcomes as mediators might be warranted to understand how the treatment works.

References

1. Barlati, S., Deste, G., De Peri, L., Ariu, C., & Vita, A. (2013). Cognitive remediation in schizophrenia: current status and future perspectives. Schizophrenia research and treatment, 2013, 156084

2. Biecek, P., Maksymiuk, S., Baniecki, H. (2018). DALEX: explainers for complex predictive models. Journal of Machine Learning Research. 19: 1-5

3. Cella, M., Reeder, C., Wykes, T. (2015). "Cognitive remediation in schizophrenia – now it is really getting personal." Current Opinion in Behavioral Sciences. 4: 147-151

4. Combs D.R., Tosheva A., Penn D.L., Basso M.R., Wanner J.L., Laib K. (2008). Attentional-shaping as a means to improve emotion perception deficits in schizophrenia. Schizo Research. 105(1-3): 68-77

5. McGurk, S. R., Twamley, E. W., Sitzer, D. I., McHugo, G. J., Mueser, K. T. (2007). "A Meta-Analysis of Cognitive Remediation in Schizophrenia" Am J Psychiatry. 164:12

6. Penadés R, Catalán R, Salamero M, Boget T, Puig O, Guarch J, Gastó C. (2006) Cognitive remediation therapy for outpatients with chronic schizophrenia: a controlled and randomized study. Schizophr Res. 87(1-3):323-31.

7. Polley, E., LeDell, E., Kennedy, C., Lendle, S., van der Laan, M. (2019) SuperLearner: Super Learner Prediction. The Comprehensive R Archive Network (CRAN) https://CRAN.R-project.org/package=SuperLearner

8. Polley, E.C., Rose, S., van der Laan, M.J., (2011). Super Learning in Prediction. Targeted Learning (Chapter 3). Cambridge University Press.

9. Stearea, R. (2015). The Relationship Between Social Cognition and Functional Outcomes in Schizophrenia. Social and Behavioral Sciences. 187: 256-260.

10. VanderWheele, T., Luedtke, A. R., van der Laan, M.J., Kessler R. (2019). Selecting Optimal Subgroups for Treatment Using Many Covariates. Epidemiology. 30(3): 334-341

11. Wu CS, Luedtke AR, Sadikova E, Tsai HJ, Liao SC, Liu CC, Gau SS, VanderWheele TJ, Kessler RC. Development and Validation of a Machine Learning Individualized Treatment Rule in First-Episode Schizophrenia. JAMA Network Open. 2020 Feb 5;3(2): e1921660.

12. Wykes, T., Reeder, C., Landau, S., Everitt, B., Knapp, M., Patel, A., & Romeo, R. (2007). Cognitive remediation therapy in schizophrenia: Randomized controlled trial. British Journal of Psychiatry, 190(5): 421-427

13. Wykes, T.; Huddy, Vyv (2009) Cognitive remediation for schizophrenia: it is even more complicated, Current Opinion in Psychiatry. 22(2): 161-167

14. Wykes, T., Newton, E., Landau, S., Rice, C., Thompson, N., Frangou, S., (2007) Cognitive remediation therapy (CRT) for young early onset patients with schizophrenia: An exploratory randomized controlled trial. Schizophrenia Research. 94(1–3): 221-230,

15. Wykes T, Reeder C, Corner J, Williams C, Everitt B. (1999). The effects of neurocognitive remediation on executive processing in patients with schizophrenia. Schizophr Bulletin;25(2):291-307.

Figures

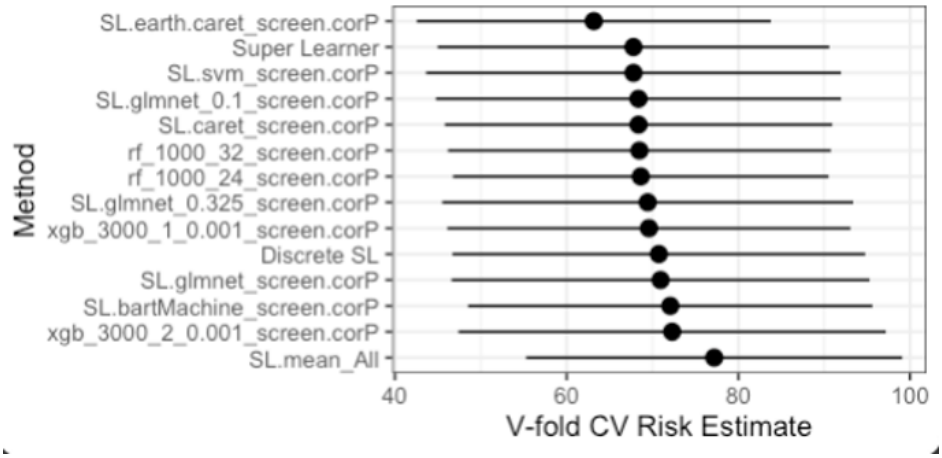

Figure 1: 5-fold Nested Cross-validated Risk Estimates for all learners in the library

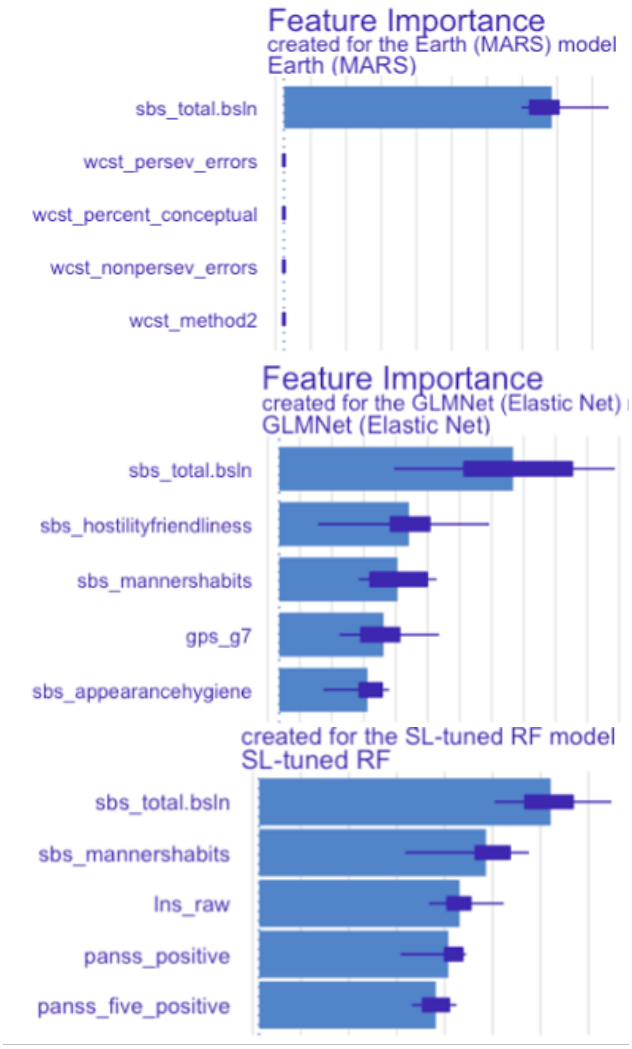

Figure 2: Variable Importance Plots for 3 learners (in order: MARS, Elastic Net, Random Forest)

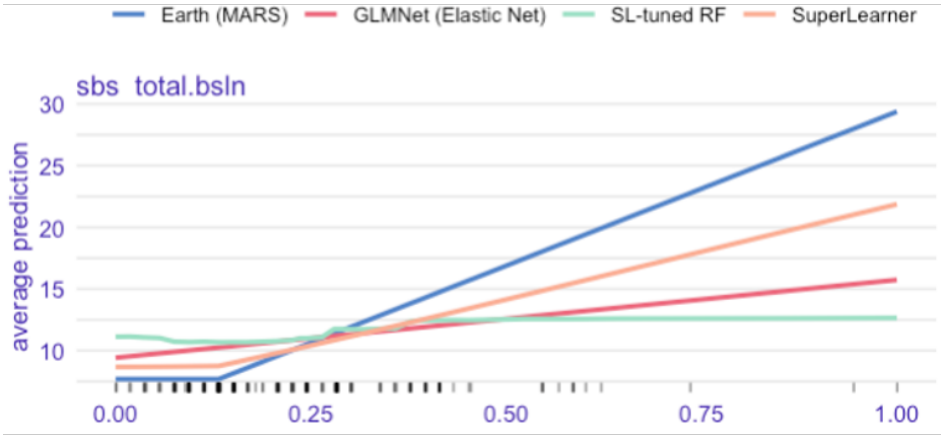

Figure 3: Partial Dependence Plots for the marginal effect of scaled total SBS score at baseline on SBS score at follow-up from different models used. Y-axis shows the mean predicted values of the SBS score at follow-up

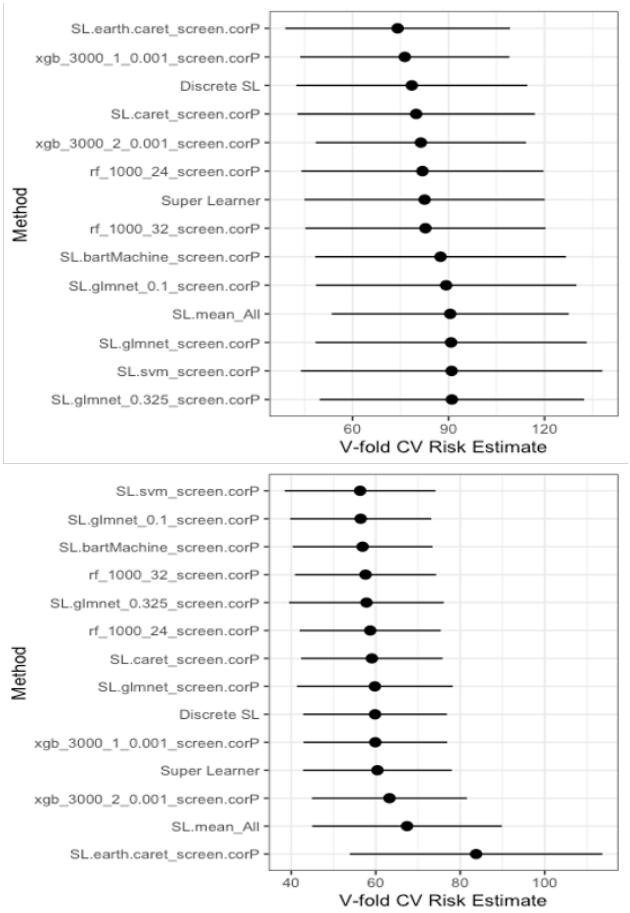

Figure 4: Cross-validated risks of selected learners among the treated group (top) and the untreated group (bottom))

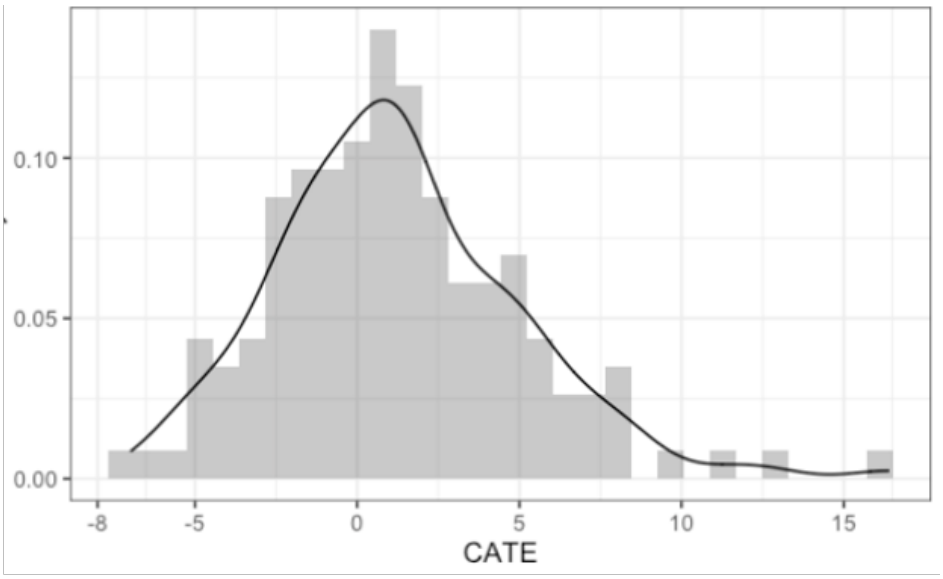

Figure 5: Distribution of the estimated Conditional Average Treatment Effects (CATE) computed from predicted counterfactual outcomes using the best-fitting learners

Figure 6: Partial Dependence Plot for the marginal effect of scaled total SBS score at baseline on the estimated CATE using Dalex. Y-axis shows the mean predicted values of the CATE