Estimating Causal Effects of Innovative Treatments for Lower Grade Glioma Using Multinomial BART

Elisa Broserà
Lorenzo Cossiga
Marco Flavio Dello Russo
Marco Luigi Deviardi
Sarra Mars
Matthieu Varenne
Tutor: Alessandro Carminati

Politecnico Di Milano

February 14, 2025

Context

Glioma is the most frequent brain tumor. Gliomas are classified (Grades I-IV, WHO) with Grade I being benign, while Grades II-III (LGG) are diffuse tumors that may progress to Grade IV (high-grade glioma).

This project focuses on analyzing the treatment outcomes of patients with **Lower Grade Glioma (LGG)**, using data from the paper [Pedone et al., 2024].

Patients dataset: 158 patients and 38 variables

Some important variables:

- **Treatment outcome first course** - defined using the RECIST criteria.
  1. Progressive Disease (PD)
  2. Partial Remission/Response (PR)
  3. Stable Disease (SD)
  4. Complete Remission/Response (CR)
- Our target variable:

$$\mathbf{Y} = \begin{cases} 0 \text{ (PD) worst outcome} \\ 1 \text{ (PR or SD)} \\ 2 \text{ (CR) best outcome} \end{cases}$$

- **New Treatment**: Targeted Molecular Therapy OR Radiation Treatment Adjuvant
- **Clinical data**: age, sex, tumor grade ...
- **Protein expressions**

Objective

- Make causal inference to understand the effectiveness of the innovative treatment
- Propose a model to estimate personalized causal effects

Causal inference

We are in an uncontrolled experiment setting.
For every patient i we have:

Binary new treatment Z

- $Z_i = 1$, new treatment
- $Z_i = 0$, control

Potential outcome Y

- $Y_i(1)$, if patient i received the treatment
- $Y_i(0)$, otherwise

We only observe one of these outcomes.

Confounding covariates X

- X_i, related both to the treatment and the outcome

Causal inference

We focus on estimating **causal effects**.

Some estimators for causal effect in the *continous* case are:

- the average treatment effect (ATE)

$$ATE = \mathbb{E}(Y_i(1) - Y_i(0)) \tag{1}$$

- the conditional average treatment effect (CATE)

$$CATE = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i) \tag{2}$$

We used other indicators, specific to *ordinal* outcomes:

$$\tau = \mathbb{P}(Y_i(1) \geq Y_i(0)), \qquad \eta = \mathbb{P}(Y_i(1) > Y_i(0)) \tag{3}$$

$$\tau(X_i) = \mathbb{P}(Y_i(1) \geq Y_i(0) \mid X_i), \qquad \eta(X_i) = \mathbb{P}(Y_i(1) > Y_i(0) \mid X_i)$$

Strong Ignorability

To identify the causal effects, we need to work under the assumption of **strong ignorability** of treatment assignment which consists of:

1. **Common support** assumption:

$$0 < \mathbb{P}(Z = 1 \mid X) < 1 \tag{4}$$

   i.e. there is no patient to whom we either surely give the treatment or surely do not.

2. **Unconfoundedness** assumption:

$$Y(0), Y(1) \perp\!\!\!\perp Z \mid X \tag{5}$$

   i.e. the probability of giving a patient the treatment does not depend on the potential effect the treatment would have.

Propensity Score

It is the estimated probability that any given individual receives the innovative treatment:

$$e(X) = \mathbb{P}(Z = 1 \mid X) \tag{6}$$

It helps reaching the strong ignorability assumption:

- we create groups of patients with similar propensity scores,
- we get that the hypothesis holds within each group.

S-Learner and T-Learner

S-Learner (Single Model)

- Uses **one model** to estimate both treatment and control outcomes.

$$p_{k\ell}(X_i) = \mathbb{P}(Y_i(1) = k, Y_i(0) = \ell \mid X_i)$$

T-Learner (Two Models)

- Trains **separate models** for treatment and control groups.

$$p_{\cdot\ell}(X_i) = \mathbb{P}(Y_i(0) = \ell \mid X_i)$$
$$p_{k\cdot}(X_i) = \mathbb{P}(Y_i(1) = k \mid X_i)$$

We estimate causal effect as:

$$\tau(X_i) = \sum_{k \geq \ell} p_{k\ell}(X_i)$$

S-Learner and T-Learner

From [Künzel et al., 2019] we learn that:

- an S-learner is the best choice when the treatment effect is small or simple
- a T-learner is the best choice when the treatment effect is highly heterogeneous

General framework: sequential regression models

Sequential regression models involve fitting multiple regression models in a stepwise manner

In any model in this project, dealing with a **categorical ordinal target variable** where K=3, we fit:

1. **Submodel 1** to estimate ρ_1: the probability of belonging to category 2 or 3

$$y_j = \begin{cases} 1 & \text{if the category is 2 or 3} \\ 0 & \text{if the category is 1} \end{cases}$$

2. **Submodel 2** to estimate ρ_2: the conditional probability of belonging to category 3 given not belonging to 1

$$y_j = \begin{cases} 1 & \text{if the category is 3} \\ 0 & \text{if the category is 2} \end{cases}$$

- Individual probability of belonging to category 1, 2 or 3:

$$\pi_{ij} = \begin{cases} (1 - \rho_{1i}) & \text{if j=1} \\ \rho_{1i} \cdot (1 - \rho_{2i}) & \text{if j=2} \\ \rho_{1i} \cdot \rho_{2i} & \text{if j=3} \end{cases}$$

BART model

$$Y = g(Z, X, T_1, M_1) + g(Z, X, T_2, M_2) + \cdots + g(Z, X, T_m, M_m) + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- Z: treatment
- X: vector of covariates
- T_j: tree j
- M_j: vector of mean responses associated to each tree's bottom nodes

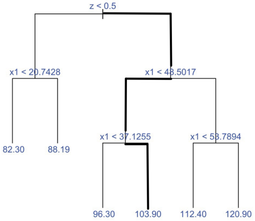


Figure 1: Example of tree realization [Hill, 2011]. Notice that $g(1, 40; T, M) = 103.9$

Updating in BART and its priors

The updating process of the trees T_j:

- A change (Grow, Prune, Change, Swap) to a random tree is proposed and accepted or not using a Metropolis Hasting algorithm
- Depth prior: probability of splitting for each node

$$p_{\text{split}} = \mathbf{base} \cdot (1+d)^{-\mathbf{power}}, \text{d: node's depth}$$

  Split prior: Dirichlet prior for variable selection

The updating of the values M_{ij} of the leaves and of the parameter σ^2 is standard:

- Priors

$$M_{ij} \sim \mathcal{N}(0, \tau)$$

$$\sigma^2 \sim \text{IG}(\alpha, \beta)$$

The BART model

for our model:

$$y_{ij} \mid p_{ij} \sim \text{Bernoulli}(p_{ij})$$

where $i \in S_j = \{i : y_{i1} = y_{i2}.. = y_{i,j-1} = 0\}$ and $j = 1, \dots, K-1$,

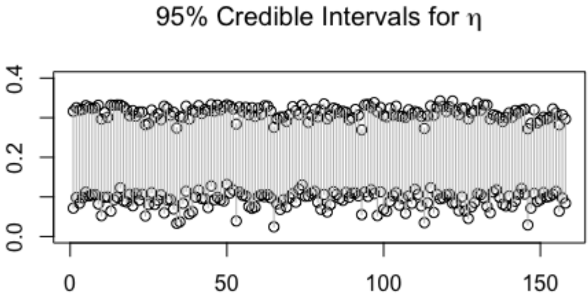
$$p_{ij} = \Phi(f_j(x_i)),$$

$$f_j \overset{\text{ind}}{\sim} \text{BART},$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

BART results

- we did some sensitivity analysis to set the hyper-parameters
- we estimated the causal effects with 95% CI for the values of η and τ



Bayesian regression model

- The model fits a logistic regression by estimating the model's coefficients while incorporating various **shrinkage priors**
- The Horseshoe prior on the β parameters is:

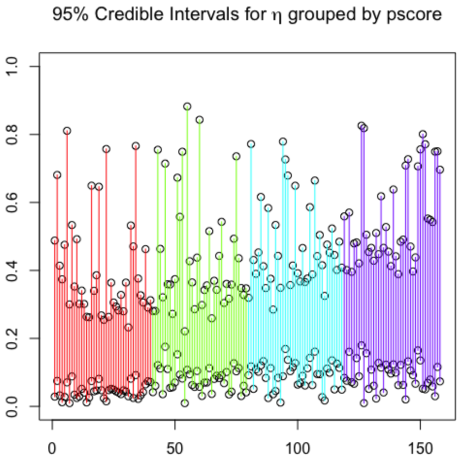
$$\beta_i \mid \lambda_i, \tau \overset{\text{ind}}{\sim} \mathcal{N}(0, \lambda_i^2 \tau^2)$$
$$\lambda_i, \tau \sim \mathrm{C}^+(0, 1)$$

  where $\mathcal{C}^+(0,1)$ is a half-Cauchy distribution, the λ_i's are local shrinkage parameters and τ is the global shrinkage parameter
- In formula:

$$\rho_{ij} = \frac{1}{1 + \exp(-\beta_0 - X_i^T \beta)} \quad j = 1, 2..K - 1$$

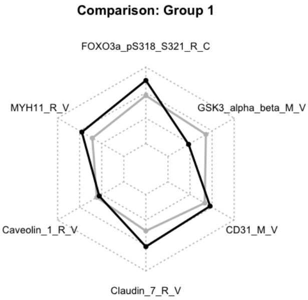
Bayesian sequential regression results

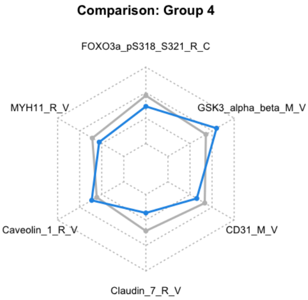
- We started with an S-learner approach, but we did not get any interesting results.
- We then moved to a T-learner: there was finally a slight pattern arising and more diverse credible intervals.



Bayesian sequential regression results

- We tried to characterize our groups to find meaningful covariates
- We found 6 proteins whose values change significantly across groups





Conclusion and further advancements

Conclusion:

- We conclude that the proposed treatment effect varies with respect to the values of 6 proteins.
- Our model can place a new patient in one of the groups to estimate the average effect and produce personalized credible intervals.

Further advancements:

- With more data, we could use a BART T-learner and obtain more precise intervals.
- We could try to isolate the effects of targeted molecular therapy and radiation treatment.

References

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009).
Handling sparsity via the horsehoe.
Proceedings of Machine Learning Research, page 8.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010).
Bart: Bayesian additive regression trees.
Annals of Applied Statistics, pages 266–298.

Ding, P. (2023).
A first course in causal inference.
pages 464–490.

Hill, J. L. (2011).
Bayesian nonparametric modeling for causal inference.
Journal of Computational and Graphical Statistics, 20:217–240.

Künzel, S. R., Sekhon, J. S., J., B. P., and Yu, B. (2019).
Metalearners for estimating heterogeneous treatment effects using machine learning.
Proceedings of the National Academy of Sciences, 116(10):4156–4165.

Linero, A. R. (2018).
Bayesian regression trees for high-dimensional prediction and variable selection.
Journal of the American Statistical Association, 113(522):626–636.

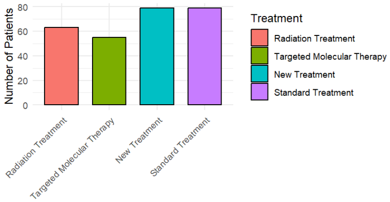
Lu, J., Ding, P., and Dasgupta, T. (2018).
Treatment effects on ordinal outcomes: Causal estimands and sharp bounds.
Journal of Educational and Behavioral Statistics, 43(5):540–567.

Pedone, M., Argiento, R., and Stingo, F. C. (2024).
Personalized treatment selection via product partition models with covariates.
Biometrics, 80(1).

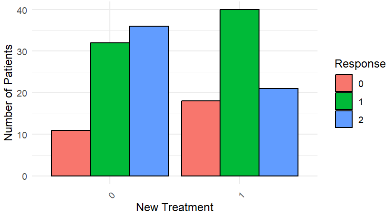
Rigon, T., Durante, D., and Torelli, N. (2019).
Bayesian semiparametric modelling of contraceptive behaviour in india via sequential logistic regressions.
Journal of the Royal Statistical Society, pages 225–247.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021).
Nonparametric machine learning and efficient computation with bayesian additive regression trees: The bart r package.
Journal of Statistical Software, 97:1–66.

Analysis of the dataset



(a) Patients Distribution by treatment type



(b) Contingency Table of New Treatment vs. Response