Politecnico di Milano

Bayesian Statistics

A.Y. 2024/2025

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

Authors:

Elisa Brosera

Marco Flavio Dello Russo

Marco Luigi Deviardi

Lorenzo Cossiga

Matthieu Marie Jean Pierre Varenne

Sarra Mars

Tutor: Alessandro Carminati

Contents

1 Introduction

Glioma is the most frequent brain tumor. Gliomas are classified by the World Health Organization (WHO) into Grades I-IV, with Grade I being benign and Grades II-III being classified as Lower Grade Glioma (LGG), which are diffuse tumors that may progress to Grade IV (high-grade glioma). This project aims to analyze the treatment outcomes of patients diagnosed with LGG using data from the study by Pedone et al. [1].

2 Dataset

2.1 Dataset Description

The dataset contains clinical data for 158 patients with LGG and 38 variables. Key variables include:

Treatment Outcome: Defined using the RECIST criteria:

- Progressive Disease (PD)
- Partial Remission/Response (PR)
- Stable Disease (SD)
- Complete Remission/Response (CR)

Target Variable: The target variable is defined as:

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

Treatments: Targeted Molecular Therapy or Radiation Treatment (Adjuvant).

Clinical Data: Age, sex, tumor grade, and other clinical factors.

Protein Expressions: Protein expression levels are also included in the dataset, which may play a role in treatment outcomes.

2.2 Dataset Preprocessing

Preprocessing steps included:

Feature selection: Removing irrelevant or redundant columns:

- `bcr_patient_barcode`
- `treatment_outcome_first_course`, which is explained in the target variable
- `radiation_treatment_adjuvant`
- `targeted_molecular_therapy`

patients with one of this two last therapies were all considered as patients with new treatment.

Response variable extraction: Since most of our code requires integer values, the response was shifted and converted, resulting in levels:

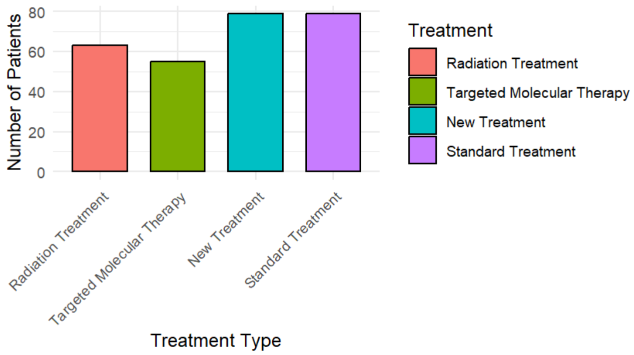
- 1 (PD), 2 (PR/SD), and 3 (CR)

3 Exploratory analysis

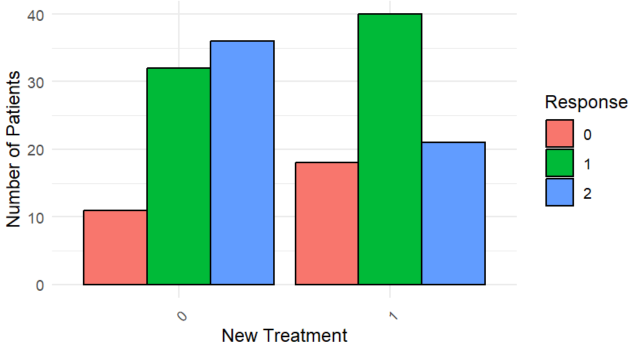
The dataset exploratory analysis focuses on two key aspects: the distribution of patients by treatment type and the contingency relationship between the new treatment and patient response.

In Figure (a), the bar chart illustrates the number of patients across four treatment categories: Radiation Treatment, Targeted Molecular Therapy, New Treatment, and Standard Treatment. We observe that there are 79 patients receiving the standard treatment and 79 patients receiving the new treatment, indicating equal representation for these two categories.

In Figure (b), the plot reveals notable variation in response levels for patients receiving the new treatment ($Z = 1$) compared to those who did not ($Z = 0$). Specifically, patients receiving the standard exhibit higher frequencies of the best outcome and lower frequencies of the worst outcome, suggesting that the new treatment may be less effective than the standard treatment.



(a) Patients Distribution by treatment type



(b) Contingency Table of New Treatment vs. Response

4 Problem Statement

The goal of this project is to analyze the causal effect of innovative treatments (targeted molecular therapy and radiation treatment) on the treatment outcomes for patients with LGG. We aim to estimate the treatment effect on the ordinal outcome (PD, PR/SD, CR) using a multinomial approach, specifically employing Bayesian Additive Regression Trees (BART).

5 Causal inference

We utilize causal inference methods to assess the effect of the innovative treatment. We define causal effects using the potential outcomes framework. In this framework an experiment has a treatment, and we are interested in its effect on an outcome. Note that most of theory is about treatment against no treatment while in our case we care about innovative treatment against standard treatment.

5.1 Potential Outcome framework

Consider a study with n experimental units indexed by $i \in 1, 2, \ldots, n$. For each unit i we have Z_i, the binary treatment indicator for unit i, vectorized as $\mathbf{Z} = (Z_1, \ldots, Z_n)$. The treatment has two levels: 1 for the treatment and 0 for the control. The observed outcome Y of unit i is a function of the potential outcomes and the treatment indicator:

$$Y_i = \begin{cases} Y_i(1), & \text{if } Z_i = 1 \\ Y_i(0), & \text{if } Z_i = 0 \end{cases}$$

Here, $Y_i(1)$ represents the outcome if the patient received the new treatment, while $Y_i(0)$ represents the outcome if they did not. The experiment reveals only one of unit i's potential outcomes with the other one missing. Then we have also a vector X that represents the confounding covariates, which, in our case, include the patients characteristics (age, sex, protein levels. . .) that influence both the treatment assignment and the outcome.
This definition of potential outcomes has two implicit assumptions:

no interference. Unit i's potential outcomes do not depend on other units' treatments.

consistency. We require that the treatment levels be well-defined, or have no ambiguity at least for the outcome of interest.

The causal effects we are looking for are functions of some difference between $Y_i(1)$ and $Y_i(0)$. Inferring individual causal effects is fundamentally challenging because we can only observe either $Y_i(1)$ or $Y_i(0)$ for each unit i.

5.2 Strong Ignorability Assumption

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

1. **the common support assumption**
$$0 < \mathbb{P}(Z = 1 \mid X) < 1$$
This implies that there are no patients who are certain to either receive or not receive the treatment.

2. **the unconfoundedness assumption**
$$Y(1), Y(0) \perp\!\!\!\perp Z \mid X$$
which requires that the potential outcomes vector must be independent of the treatment given the covariates. This implies that the probability of giving a patient the treatment does not depend on the potential effect the treatment would have.

This assumption of strong ignorability is quite restrictive, particularly the assumption of unconfoundedness, which appears to be unrealistic. We will later introduce propensity scores in order to achieve ignorability.

5.3 Key Causal Estimands

Let's now focus on the way we are going to measure the causal effects.

Some common estimators of causal effect are:

- the Conditional Average Treatment Effect ($CATE$):
$$CATE = \sum_{i=1}^{n} \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i) \tag{1}$$

- the Conditional Average Treatment Effect for the Treated ($CATT$):
$$CATT = \sum_{i:Z_i=1}^{n} \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i) \tag{2}$$

- the Average Causal Effect (ACE):
$$ACE = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0) \tag{3}$$

The problem with these is that they are not meant for ordinal outcomes. Some alternatives have been proposed by [2], such as:

$$\tau = \mathbb{P}(Y_i(1) \geq Y_i(0)) \qquad \eta = \mathbb{P}(Y_i(1) > Y_i(0)) \qquad \Delta_j = \mathbb{P}(Y_i(1) \geq j) - \mathbb{P}(Y_i(0) \geq j), \quad j = 1, 2 \tag{4}$$

In our work we mainly focused on eta and tau, because they are the ones with the most practical meaning: all the doctor cares about is whether or not a patient will be better off with some treatment or with some other.

6 Propensity Score

An important element which we will need from a certain point on in our analysis is the notion of propensity score. The propensity score [3] is the estimated probability - computed with a generalized linear model - that any given individual receives the innovative treatment given the values of the covariates:

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

What we can do with it, besides adding it to the model as a covariate, is to group observations by similar values of propensity score achieving the so called: *"stratified ignorability"*: what this means is that the strong ignorability assumption holds in each group.

So it's an additional information we can add to our data to help achieving the strong ignorability condition, which is mandatory to perform causal inference.

7 S-Learner and T-Learner

In causal inference, estimating the Conditional Average Treatment Effect (CATE) requires robust methods to model the relationship between covariates, treatment, and outcomes. The **S-Learner and T-Learner** are two used approaches for this purpose [4].

The **S-Learner** (Single Model Learner) uses a single supervised learning model to estimate both the treated and control outcomes by including the treatment indicator W as a feature of the model. The combined response function

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

is estimated using the entire dataset:

$$\{(X_i, Y_i, W_i)\}.$$

The **T-Learner** (Two-Model Learner) estimates the treatment effect by modeling the potential outcomes for the treated and control groups separately.

First, the control response function

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

is estimated using a supervised learning model (or base learner) trained on the control group data:

$$\{(X_i, Y_i) \mid W_i = 0\}.$$

Second, we estimate the treatment response function

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

using a model trained on the treated group data:

$$\{(X_i, Y_i) \mid W_i = 1\}.$$

We estimate causal effect as:

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

8 BART model

In this project, we use Bayesian Additive Regression Trees (BART) to model the causal relationship between the treatments and the outcome variable, considering the ordinal nature of the outcomes. This approach allows for flexible modeling of complex, non-linear relationships in the data and, at the same time, handles variable selection.

8.1 BART's structure

BART models the outcome Y as [5]:

$$Y = f(Z, X) + \epsilon,$$

where:
Z: Treatment variable ($Z = 0$ for control, $Z = 1$ for treatment),
X: Confounding covariates,
ϵ: Additive errors ($\epsilon \sim N(0, \sigma^2)$).
Under the strong ignorability assumption of the treatment assignment ($Y(0), Y(1) \perp Z \mid X$):

$$\mathbb{E}[Y(0) \mid X = x] = \mathbb{E}[Y \mid Z = 0, X = x] = f(0, x),$$

$$\mathbb{E}[Y(1) \mid X = x] = \mathbb{E}[Y \mid Z = 1, X = x] = f(1, x).$$

Let's now explain how BART works. BART models the target function as the sum of multiple decision trees. Each tree contributes partially to the overall prediction and consists of interior nodes decision rules splitting observations into subgroups and a set of terminal nodes. Moreover, the bottom nodes of each tree are associated with the vector of parameters M_i which represents the vector of mean responses. Each entrance in M_i is the mean of the y of the observations associated to that specific leaf.

Let's visualize a single tree to have a more precise idea:

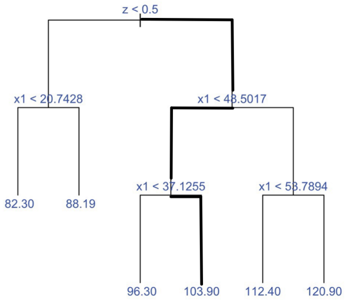


Figure 2: Example of tree realization [5]. Notice that $g(1, 40; T, M) = 103.9$

The tree above splits the observations at each decision node binarily. The decision rules are of the form: $\{X1 <= c\}$vs$\{X1 > c\}$. Notice, for example, that the pair (Z=1, X1=40) will follow the bold path in the tree and its predicted outcome would be 103.9, the mean of the observations dropping down the tree and hitting the same bottom node. We define g(Z=1, X=40; T, M) as such value. In general, $g(Z, X, T, M)$ is the contribution of tree T to the prediction of the response for (Z,X), based on the path followed by (Z, X) through the tree.

As mentioned above, BART is not built on a single tree, but sums over multiple trees to capture complex response surfaces:

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

$$\epsilon \sim N(0, \sigma^2).$$

The intuition behind this formula is that once the first tree in the model has returned its fit, $g(Z, X; T_1, M_1)$, the residuals, obtained by subtracting the fit to the original response, are explained by the second tree and so on, tree after tree.

In this model, the parameters are (T_j, M_j) and σ. A prior is imposed on the parameters and the posterior is computed using a backfitting Markov chain Monte Carlo (MCMC), in particular a generalization of a Gibbs Sampler. At each iteration of the algorithm, all parameters are recomputed. Here below we understand BART's priors and the updating process.

8.2 Updating in BART and its priors

The updating process in BART regards the trees (T_j), the values (M_j) of the leaves and the parameter σ.

1. Updating on the trees: At each iteration, a change to a random tree is proposed to maximize the posterior probability of the tree given the data and other parameters. One of the following operations are allowed:

   Grow: Split a randomly selected leaf node into two new leaves by adding a split on a predictor X_k at a threshold c.

   Prune: Merge two child leaf nodes back into their parent node, removing a split.

   Change: Modify an existing split by changing either the predictor X_k or the threshold c.

   Swap: Swap splits between adjacent internal nodes without changing the overall structure.

   The probabilities of these operations are predefined (e.g., 0.25 grow, 0.25 prune, 0.4 change, 0.1 swap). Then the modification is accepted or not according to a Metropolis Hasting algorithm [6]. Said so, the tree priors, that regularize the fit by keeping the individual tree effects from being unduly influential, are:

   Depth prior: the probability of splitting at a given node.

   $$p_{\text{split}} = \text{base} \cdot (1+d)^{-\text{power}}$$

   where d is the depth of the node and base and power are tunable parameters.

   Split prior: when choosing the decision rule to split, each covariate can be equally weighted or not. A Dirichlet prior is used for variable selection.

   $$(p_1, p_2, \ldots, p_p) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_p)$$

   where the parameters α_j control the probability of selecting each variable.

2. Updating on the leaf node parameters: Initially, the values associated with the bottom nodes of the tree j, M_{ji}, are assigned a Gaussian prior:

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

   The estimates for leaf node values are updated at each iteration, changing the mean and the variance of the normal distribution according to the new subsets of observations. Once a new distribution is found, a new value for M_{ji} is drawn.

3. Updating on the noise variance σ: The variance of the residuals, σ^2, is assigned an Inverse-Gamma prior:

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

   At each step, α and β are updated after having computed the residuals of the current model. In particular:

   $$\alpha = \alpha_0 + \frac{n}{2}$$

   $$\beta = \beta_0 + \frac{1}{2}\sum r_i^2$$

8.3 BART for binary outcomes

BART for binary outcomes is the following [7]:

$$y_i \mid p_i \sim \text{Bernoulli}(p_i), \quad \text{where } i = 1, \ldots, N,$$
$$p_i = \Phi(f(x_i)),$$
$$f \sim \text{BART},$$

where i indexes subjects, $i = 1, \ldots, N$, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

8.4 Multinomial BART for categorical outcomes

In this project, since we have 3 categories, we are interested in using Multinomial BART for categorical outcomes. The model is the following [7]:

$$y_{ij} \mid p_{ij} \sim \text{Bernoulli}(p_{ij}) \quad \text{where } i \in S_j = \{i : y_{i1} = y_{i2} = y_{i3}.. = y_{i,j-1} = 0\} \text{ and } j = 1, \ldots, K-1,$$
$$p_{ij} = \Phi(f_j(x_i)),$$
$$f_j \overset{\text{ind}}{\sim} \text{BART},$$

with i indexing subjects, $i = 1, \ldots, N$ and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

9 MBART results

9.1 MBART Model Setup

After some tries for hyperparameters tuning, the Multinomial BART model was fitted using the following configuration:

Response variable (`y`): Treatment outcomes (levels 1, 2, 3)

Predictors (`data`): The processed clinical features from the original dataset, excluding the response.

Hyperparameters:

- Number of Trees (`ntree`): 100
- Power Parameter (`power`): 2
- Base Parameter (`base`): 0.95
- Number of Posterior Draws (`ndpost`): 5000
- Burn-in Iterations (`nskip`): 1000
- Sparse Variable Selection (`sparse`): Disabled

Both training and test datasets were set to `data` for this analysis.

9.2 Sensitivity analysis

The MBART model produced posterior probabilities for each ordinal class. The predicted class for each patient was determined by selecting the class with the highest posterior probability, yielding the following results:

- Total correct predictions: **117 out of 158** patient. (Overall accuracy: **74.05%**)
- The confusion matrix (below) compares the true treatment outcomes to the predicted classes:

Predicted Class	True Class 1	True Class 2	True Class 3
Class 1	11	0	0
Class 2	15	68	19
Class 3	3	4	38

Table 1: Confusion matrix comparing predicted and true treatment outcomes.

The MBART model tends to favor predicting **Class 2** and **Class 3**, while under-predicting **Class 1**. This suggests a potential imbalance in model sensitivity across classes.

9.3 Credible Intervals for Misclassified Observations

We analyzed the **credible intervals (CIs)** for the posterior probabilities of the **correct and predicted classes** for **misclassified observations** in the ordinal regression task. We can see that for most of misclassified observations there is a high confidence interval for the second class.

For each misclassified patient, the analysis involves:

1. Extracting the posterior probabilities for:
   - The **correct class** (true class label `y`).
   - The **predicted class** (model's prediction `pred_class`).
   - The **third class**.
2. Computing the **credible intervals** (2.5th and 97.5th percentiles) for these probabilities to quantify uncertainty.
3. Visualizing the results with error bars to compare the model's confidence in the correct and predicted classes.

In the appendix there are traceplots of the MCMC of the estimated probabilities of some patients

9.4 Estimate of causal effects

The results of our analysis will be presented here, showcasing the estimated treatment effects and comparisons between different treatments.

The first thing we did was to run the BART's MCMC to obtain the probabilities for $Y_i(1)$ and $Y_i(0)$. We obtained - running 5000 iterations of the MCMC - for each patient i, two 5000x3 matrices of probabilities, the first one for $T_i = 0$ (i.e., old treatment), the second one for $T_i = 1$ (i.e., new treatment). For each iteration of the MCMC, we computed the value of the causal estimand ($delta_j$, tau, or eta) for each patient i. This way we ended up with a 5000x158 matrix for each causal estimand containing its values for each iteration and for each patient. Finally, for each patient, we computed the 95% credible interval for the causal estimands. This approach, even with different values of the hyperparameters of BART, didn't get interesting results because, for each patient, we were observing very similar credible intervals, making any causal inference impossible.

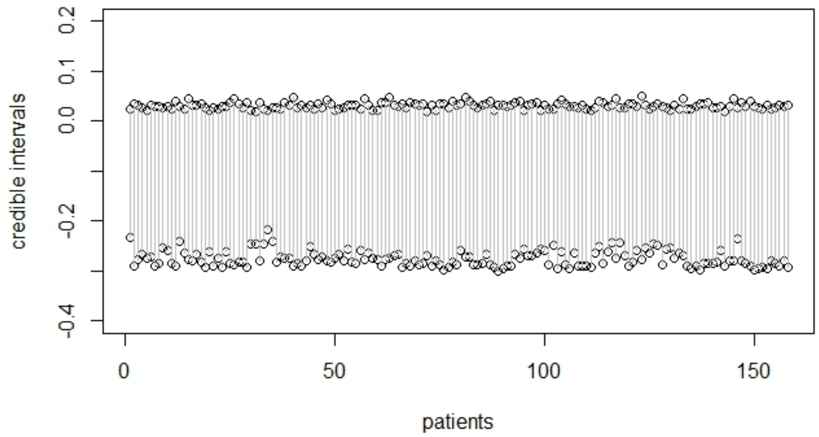


Figure 3: 95% Credible Intervals for $delta_{2i}$

This failure could have 3 possible causes:

1. Either the strong ignorability assumptions did not hold;
2. or the BART model is too complex for our data;

3. or the data itself does not present any causal effects to be found.

We then tried to make sure that our strong ignorability assumption held. To do this we both added the propensity score as a covariate in our BART model and we grouped the patients by propensity score.

A key element in this case is the selected number of groups.

If it is too small then we don't achieve ignorability, if it is too large then we do not have enough units within each group and the outermost groups have only advanced or standard units. Therefore, we face a trade-off. The number that looked fit for out data was 4, so from now on we are going to split our observations into 4 omogeneous groups where patients with lower propensity score are assigned to group 1, and the ones with higher propensity score to group 4.

We then calculated the propensity scores and rerun BART MCMC with propensity score as covariate, then we computed tau, eta, and delta stratified (starting from our 5000x158 matrix we averaged the rows by group, obtaining a 5000x4 matrix for each estimand), but once again the credible intervals of the 4 groups were almost identical for each group.

Apparently the BART model is too complex and we don't have enough data to properly train it, our only option at this point was to change our model: we went for Bayesreg.

It is important to note that to make sure that the problem was in the data and not in our model/code we ran the same model on a simulated dataset where we forced a clear and obvious causal effect, by setting all observations with the new treatment to group 2, and all observations with the standard treatment to groups 0 and 1. In this case, BART was able to recognize the positive effect of the treatment, with confidence intervals of the causal effects showing an improvement in outcomes for patients treated with the new treatment. We also used a second simulated dataset to see if BART would recognize the evident effect of a modified covariate. To do this, we increased the numerical value of a protein, in our case the FOXO3a-pS318-S321-R-C, for patients who had received the treatment and were in group 3. When calculating the causal effects in this scenario, the model identified the protein as significant: individuals with a high value of said protein showed a higher credible interval for eta and tau: they are expected to respond better to the treatment.

10 Bayesian Regression

We subtituted the BART model with Bayesian Regression, which is a model that uses the *Bayesreg* package in R, which allows fitting a logistic regression model by estimating the model's coefficients while incorporating various shrinkage priors for variable selection and complexity control. Specifically, the Horseshoe prior was used as explained in [8], where the prior on the β parameters is:

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

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.

For our model we called the function *bayesreg* twice: once to compute p_{i1}, and once to compute p_{i2}. The way we did that is we computed two one-hot encoded vectors, the first with all 1s for patients belonging to the 0 group and 0 elsewhere and the second one with all 1s for patients belonging to the 1 group and 0 elsewhere. Bayesreg estimates β_0 and all the β_i for each iteration, so we compute $p_{i1}$and p_{i2} calling twice the function

$$p_{ij} = \frac{1}{1 + \exp(-\beta_0 - X_i^T \beta)} \tag{5}$$

and p_{i3} as 1-p_{i1}-p_{i2}.

This approach ended up not giving interesting results, mostly because of a simple but very important issue: this idea of one-hot encoded vectors does not give any information about the ranking of the groups, which is a key element in our data. What we need is a way to build our model in such a way that this ordering is preserved. The way to do this is explained in [9]: the idea is to first choose between group 1 and 2-3 combined then, if we don't choose 1, we have the second model that discriminates between 2 and 3. Note that this approach can be extended to any number of (ordered) groups.

The actual implementation needed two auxiliary variables ρ_1 and ρ_2. We estimated ρ_1 as the probability of belonging to category 2 or 3 (so we set $y_i = 1$ if the category is 2 or 3 and $y_i = 0$ if the category is 1). Then ρ_2

as the probability of belonging to category 3, conditioned on not belonging to 1 (so we do not use for this second regression the patients from category 1).
Then, from ρ values, computed as in (5), we obtained p_{ij} as:

- Probability of category 1: $(1 - \rho_{1i})$
- Probability of category 2: $\rho_{1i} \cdot (1 - \rho_{2i})$
- Probability of category 3: $\rho_{1i} \cdot \rho_{2i}$

At the end with this 3 matrices of p_{ij} we computed our causal estimands tau, eta and delta as always.

11 Bayesreg Results

Both models presented in 10 did not produce interesting results: the credible intervals for the patients were still very similar, and the variable selection did not work, as no significative covariates were found from the confidence intervals of the fits of bayesreg, which all included zero.

To understand why this approach failed it's useful to mention that there are some cases, for example when treatment effect varies significantly, where we need to use a T-learner instead of an S-learner.
This is precisely what we tried to do next. We follow the same idea of the last approach we described, but we used two different models: one for standard treatment and one for innovative treatment. This time, plotting the credible intervals for our target estimates, we started to see a pattern arise: some patients have a different credible interval than others!

From here we tried to group these observations by pscore to see if we could identify a pattern, and we actually could. We can see that on average the credible intervals move slightly upwards going from lower to higher pscore.

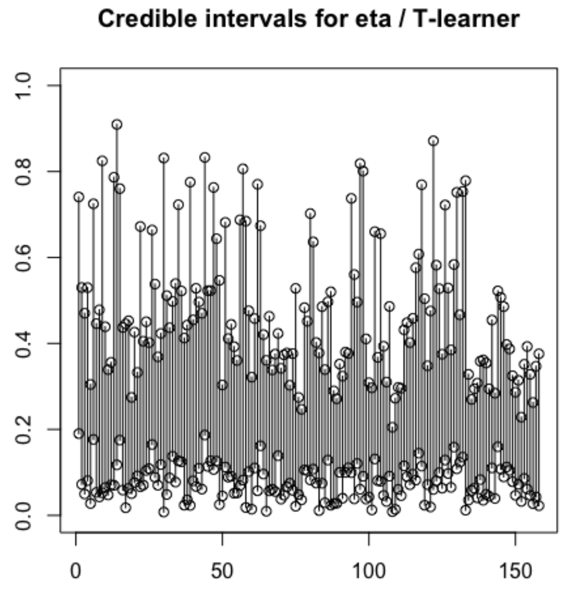


Figure 4: 95% Credible Intervals for η

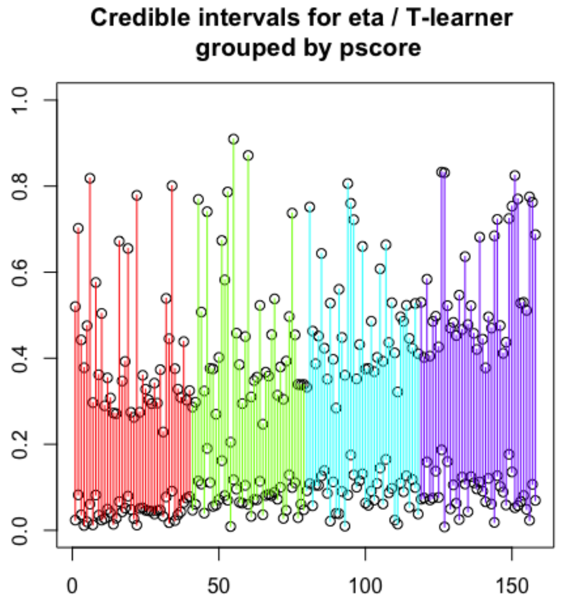


Figure 5: 95% Credible Intervals for η by pscore

To better understand this trend we computed the credible intervals of eta and tau among the 4 different groups, the results are shown in the table below and confirm the guess we had by looking at the plot.

P-score Group	**2.5% lower bound**	**97.5% upper bound**
1	0.045	0.415
2	0.079	0.450
3	0.089	0.482
4	0.088	0.563

Table 2: Credible intervals for η among propensity score groups

The last step of our work was to try to characterize the 4 groups we identified via pscore by looking for a pattern of some proteins that may have a particular trend among groups. We started by looking at the summary of the gls model for the pscore and we found the 3 most relevant proteins, which were "MYH11-R-V", "Caveolin-1-R-V" and "Claudin-7-R-V". Moreover, by checking at the plots of variable variations over the groups, we found other 6 variables that seemed to have a strong trend. To prove the relevance of these 9 variables we ran an anova test: out of our 9 variables 6 of them have a different mean among groups, our original 3 and 3 more: "FOXO3a-pS318-S321-R-C", "CD31-M-V" and "GSK3-alpha-beta-M-V". Below we can see some radarplots that show how these proteins are distributed among groups (we just show distribution for group 1 and 4 for simplicity, the rest can be found in the appendix).

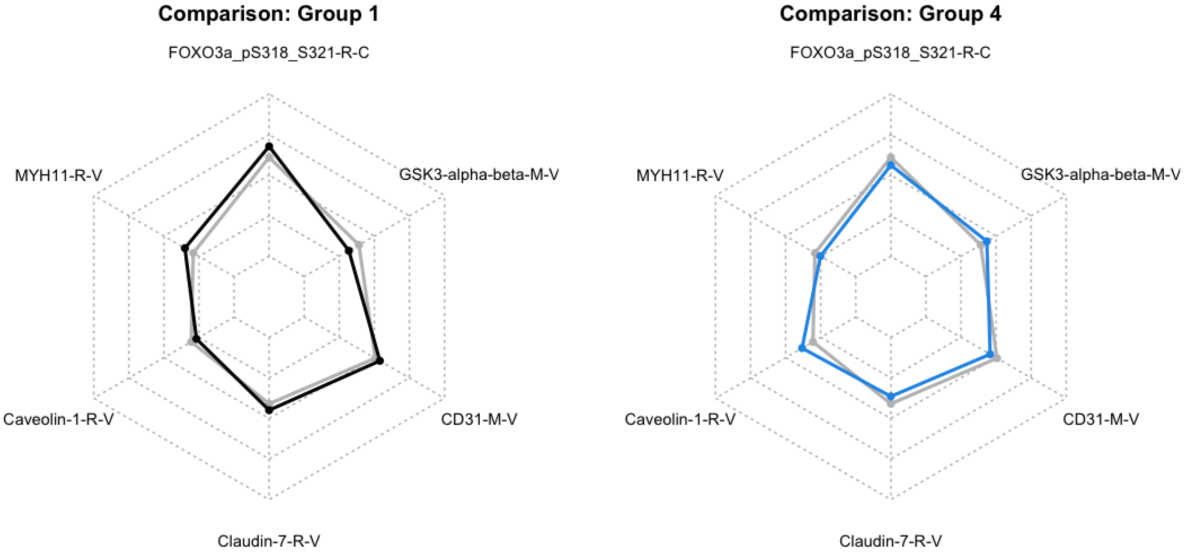


Figure 6: Radarplot for groups 1 and 4, mean shown in gray

The reason this is relevant is that, by just computing the values of a few proteins, a doctor knows which pscore group the patient belongs to and, maybe, choose to give the new treatment only to individuals from group 4 i.e. the ones who responded better to the treatment.

12 Conclusion and further works

We are satisfied with how our project turned out because, even though the data was pretty limited and we had issues finding a working model, we were still able to reach an interesting conclusion with the possibility to divide the individuals in groups and diverse credible intervals among patients, which is a big step towards personalized medicine i.e. we can choose whether or not to give a patient the innovative treatment according to a credible interval that depends on his own medical record.

If in the future, with more data, it could be interesting to try a T-learner approach for BART and explore separately the effect of targeted molecular therapy and radiation treatment so that, for each patient, we don't only have 2 choices (i.e. standard treatment or innovative treatment) but 4 (i.e. standard treatment or only targeted molecular therapy or only radiation treatment or both). The decision, just like we did, will be made with respect to the values of some clinical variables.

The full implementation of what we have done and additional results can be found at: `https://github.com/marcofdr/BayesianA2`.

13 Appendix

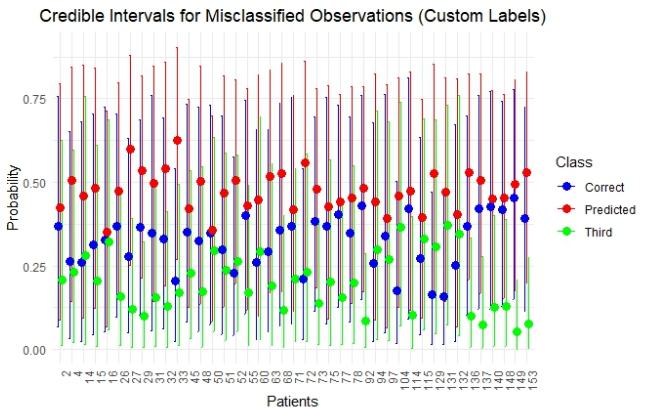


Figure 7: misclassification

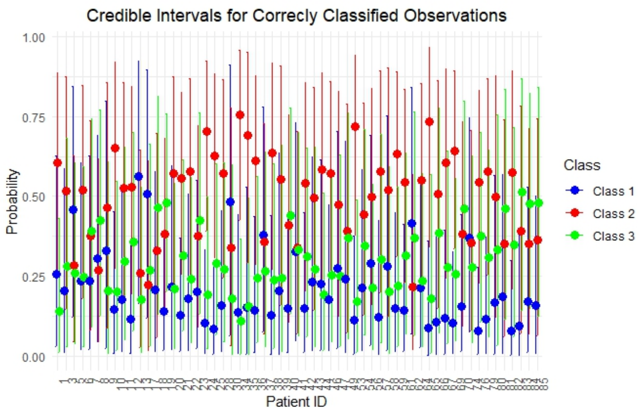


Figure 8: misclassification 2

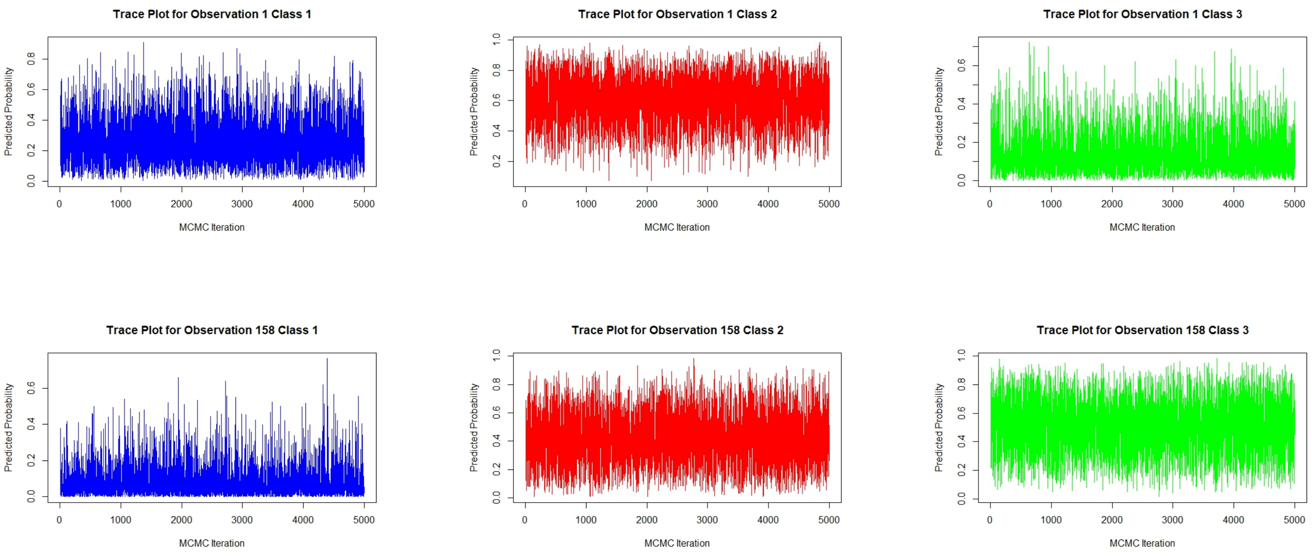


Figure 9: BART traceplots

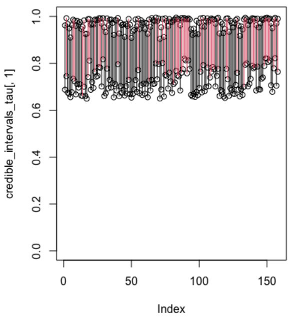


Figure 10: 95% Credible Intervals for τ for the simulated dataset

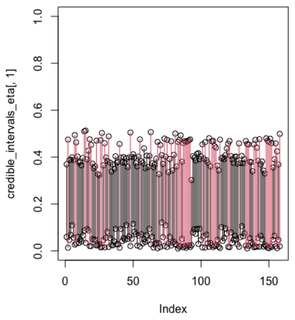


Figure 11: 95% Credible Intervals for η for the simulated dataset

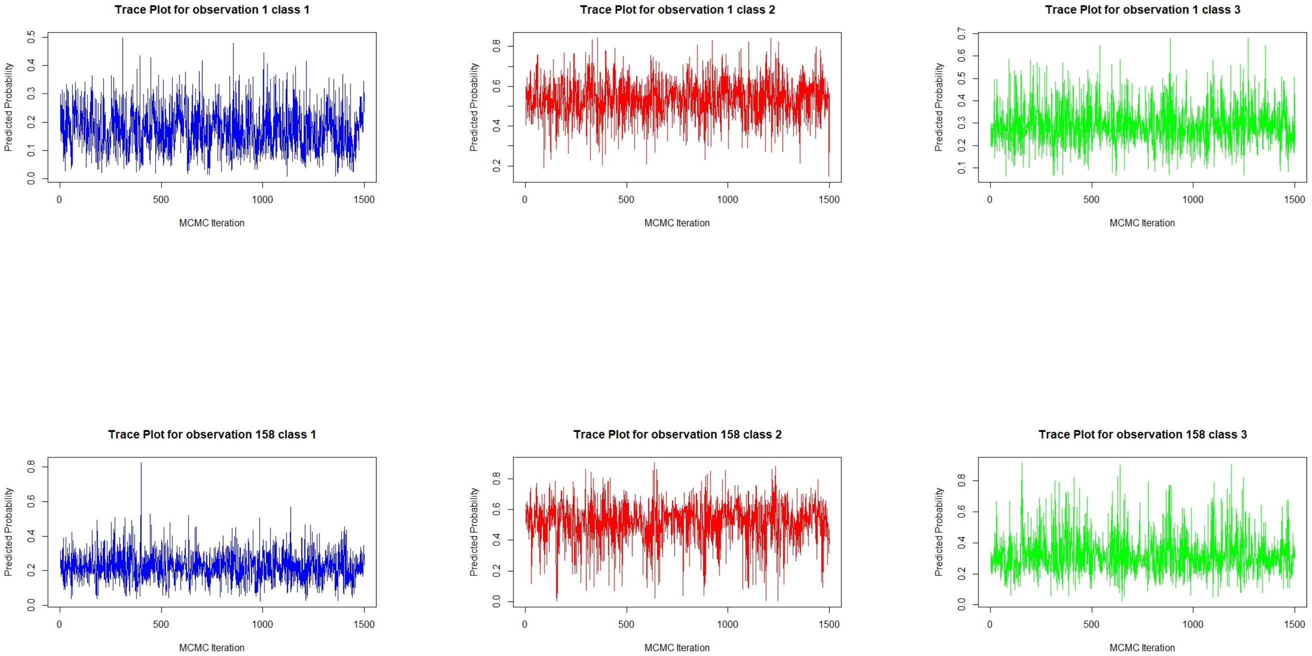
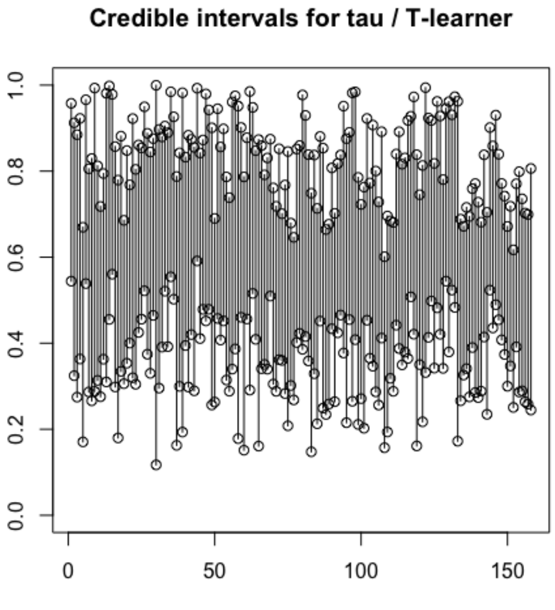


Figure 12: BayesReg traceplots



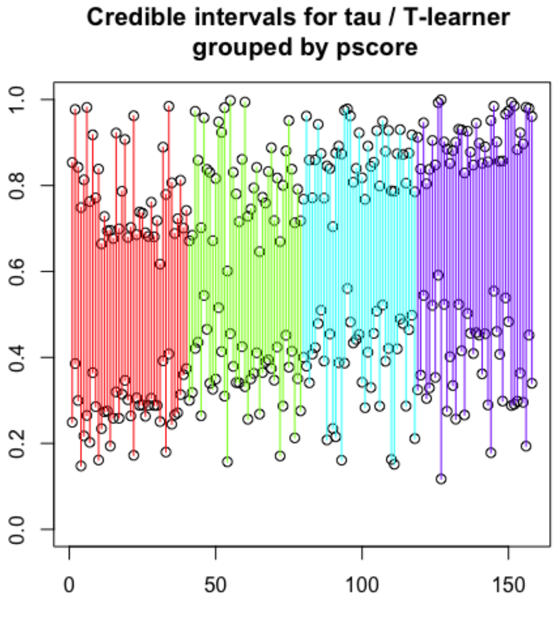


Figure 13: 95% Credible Intervals for τ

Figure 14: 95% Credible Intervals for τ by pscore

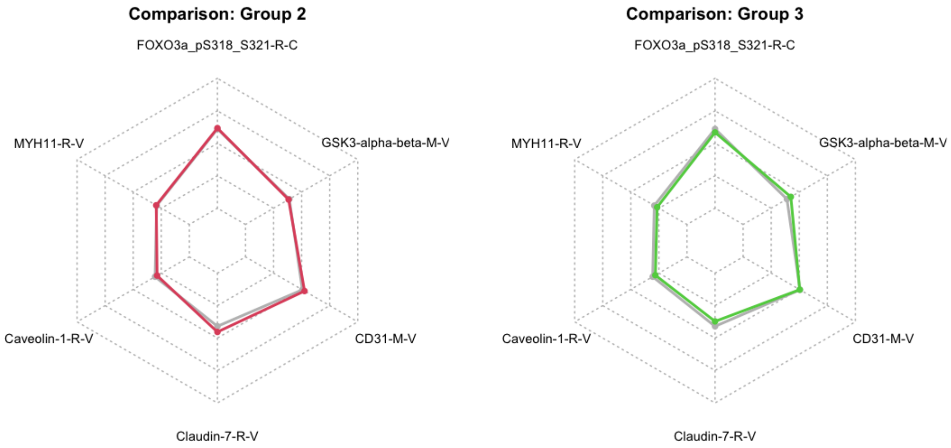


Figure 15: Radarplot for groups 2 and 3, mean shown in gray

References

[1] M. Pedone, R. Argiento, and F. C. Stingo. “Personalized treatment selection via product partition models with covariates”. In: *Biometrics* 80.1 (2024). DOI: 10.1093/biomtc/ujad003.

[2] J. Lu, P. Ding, and T. Dasgupta. “Treatment Effects on Ordinal Outcomes: Causal Estimands and Sharp Bounds”. In: *Journal of Educational and Behavioral Statistics* 43.5 (2018), pp. 540–567. DOI: 10.3102/1076998618776435.

[3] P Ding. “A First Course In Causal Inference”. In: (2023), pp. 464–490. DOI: 10.48550/arXiv.2305.18793. URL: https://doi.org/10.48550/arXiv.2305.18793.

[4] S. R. Künzel et al. “Metalearners for estimating heterogeneous treatment effects using machine learning”. In: *Proceedings of the National Academy of Sciences* 116.10 (2019), pp. 4156–4165. DOI: 10.1073/pnas.1804597116. URL: https://www.pnas.org/cgi/doi/10.1073/pnas.1804597116.

[5] J. L. Hill. “Bayesian Nonparametric Modeling for Causal Inference”. In: *Journal of Computational and Graphical Statistics* 20 (2011), pp. 217–240.

[6] H. A. Chipman, E. I. George, and R. E. McCulloch. “BART: Bayesian additive regression trees”. In: *Annals of Applied Statistics* (2010), pp. 266–298. DOI: 10.1214/09-AOAS285. URL: https://doi.org/10.1214/09-AOAS285.

[7] R. Sparapani, C. Spanbauer, and R. McCulloch. “Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package”. In: *Journal of Statistical Software* 97 (2021), pp. 1–66. DOI: 10.18637/jss.v097.i01.

[8] C. M. Carvalho, N. G. Polson, and J. G. Scott. “Handling Sparsity via the Horsehoe”. In: (2009), p. 8. URL: https://proceedings.mlr.press/v5/carvalho09a/carvalho09a.pdf.

[9] T. Rigon, D. Durante, and N. Torelli. “Bayesian Semiparametric Modelling of Contraceptive Behaviour in India Via Sequential Logistic Regressions”. In: *Journal of the Royal Statistical Society* (2019), pp. 225–247. DOI: 10.1111/rssa.12361. URL: https://doi.org/10.1111/rssa.12361.