MULTIPLE SCLEROSIS: FROM PREDICTIONS TO SENSITIVITY ANALYSIS AND ROBUST MODELS

15.095 Machine Learning under a Modern Optimization Lens

Suzana Iacob **El Ghali Ahmed Zerhouni**

siacob@mit.edu egaz@mit.edu

Abstract

Multiple Sclerosis(MS) is a disease that disrupts the flow of information within the brain. It affects 1 million people in the US [1] and remains incurable. MS treatments can cause side effects and impact the quality of life and even survival rates. Based on an existing research study [2], we investigate the risks and benefits of three treatment options based on methylprednisolone (a corticosteroid hormone medication) prescribed in (1) high-dose, (2) low-dose or (3) no treatment. The study currently prescribes one treatment to all patients as it has been proven to be the most effective on average. We aim to develop a personalized approach by building machine learning models and testing their sensitivity against changes in the data. We first developed an unsupervised predictive-prescriptive model based on k-means clustering in addition to three predictive models. We then assessed the models' performance with patient data perturbations and finally developed a robust model by re-training on a set that includes perturbations. This increased the models' robustness in highly perturbed scenarios (+10% accuracy) while having no cost in scenarios without perturbations. We conclude by discussing the trade-off between robustification and its interpretability cost.

1 Problem Statement

The exact causes of MS are still unknown in the medical community and treatment methods are an area of active research. As there is no known cure and the disease affects patients differently, prescribing effective treatments is of paramount importance. The variation of symptoms and treatment response represents a strong motivation for a personalized treatment approach. Studies such as [2] advocate a certain treatment based on best results obtained by the treatment on average, across a pool of patients. We use the dataset from this study to develop a robust machine learning approach to treatment prescription.

2 Motivation

MS treatments are used to manage disease progression often throughout the life of the patients. We believe machine learning approaches can provide accurate treatment choices based on individual patient characteristics, leading to higher chances of symptoms remission. This project focuses on building models and testing them against perturbation in the underlying data. Patient data is variable, difficult to measure and prone to human input errors. Moreover, incorrectly predicting MS treatment can cause serious side effects for the patient. For this reason, we believe robust models bring an important benefit to treatment analysis.

3 Data

The dataset was obtained from the study [2] and it consists of 10,000 observations of patient utility data. The target variable is *best treatment*, a categorical variable taking 3 possible values: (1) high-dose, (2) low-dose or (3) no treatment. The predictors are the utilities (i.e. risks) of each side effect (e.g. 0.8 utility of cardio-pulmonary distress). They differ as patients are impacted differently by the same side effect (e.g. an older patient will have a higher negative impact from a cardiac arrest). A higher utility means a higher risk and worse impact of a particular side effect. These utilities are calculated in the study [2] with qualitative information and depend on relapse severity, the difference between lethal and non-lethal outcomes, as well as individual patient characteristics. We used stratified sampling to split the data into training (7,000 observations) and testing (3,000 observations) while preserving the proportions of the three values of the target variable (best treatment).

The treatment outcome is not directly observable in the data. It is not recorded how the patient reacted to the treatment, whether there were any side effects and how the MS disease progressed as a result. We can only observe the calculated utility that each of the three treatments has on each patient. For example, for patient i, the utility of high-dose treatment was 0.87, the utility of low-dose treatment was 0.65 and the utility of no treatment was 0.54. A higher utility of treatment means a greater benefit to the patient.

Moreover, the dataset itself has been obtained via simulations, and in reality, the patient utility data is difficult to calculate. Patient measurements could be imprecise (e.g. recorded incorrectly by the measurement device) or they could be inputted in the system incorrectly by nurses or physicians. Furthermore, they represent only a snapshot of a patient's state obtained when we take measurements such as blood pressure. Finally, future patients may not present the same characteristics as past patients. MS symptoms and treatment reactions vary widely across the patient population. In other words, we cannot fully assume that patient utilities in the train and test set are drawn from a common distribution. This represents a strong argument in favor of sensitivity analysis and robust models for treatment prescription.

4 Methods

4.1 K-means predictive-prescriptive models

As mentioned above, this dataset is unsuitable to an approach such as Optimal Prescriptive Trees due to the lack of treatment outcome. We developed instead an unsupervised prescriptive-predictive method based on k-means clustering. The process follows two steps:

- Perform k-means algorithm on the patients' utilities
- For each new patient, prescribe the most common treatment of the cluster he/she belongs to

This can be formulated via the following optimization problem:
We defined T as the number of new patients, here 3,000 (patients in the test set), and n the number of previous patients here 7,000 (patients in the train set), z_p the treatment to prescribe (1 for high-dose, 2 for low-dose, and 3 for no treatment) for new patient p, and y_i the best treatment evaluated for the previous patient i, hence we get the following formulation:

$$\begin{aligned} \max_{z_p \in \{1,2,3\}} \quad & \sum_{p=1}^{T} \sum_{i=1}^{n} \quad w_{pi} 1_{z_p = y_i} \\ \text{subject to} \quad & w_{pi} = 1 \quad \text{if p and i are in the same cluster} \end{aligned}$$

This formulation can be linearized with the introduction of additional variables and a sufficiently large constant M to obtain the following mixed-integer optimization problem:

$$
\begin{aligned}
\max_{z_p \in \{1,2,3\}} \quad & \sum_{p=1}^{T} \sum_{i=1}^{n} \quad w_{pi}(1 - q_{pi}) \\
\text{subject to} \quad & w_{pi} = 1 \quad \text{if p and i are in the same cluster} \\
& M q_{pi} \geq q_{pi}^{1}(z_p - y_i) + q_{pi}^{2}(y_i - z_p) \geq 0 \\
& q_{pi}^{1} + q_{pi}^{2} = 1 \\
& q_{pi}, q_{pi}^{1}, q_{pi}^{2} \in \{0, 1\}
\end{aligned}
$$

Results: We obtained a 0.64 out-of-sample accuracy with 20 clusters, which is equal to the baseline accuracy, as we will see. This value changes slightly with the number of clusters but remains within 2% of 0.64 for 5 to 100 clusters. The challenge of the unbalanced dataset in terms of the target variable remains prevalent even when splitting the patients into numerous clusters, hence a clustering approach did not bring significant value.

4.2 Predictive models: CART, OCT, OCT-H

For a personalized treatment approach we predict the best treatment based on adverse effects utilities. Study [2] simply proposes the most effective treatment on average (high-dose) to all patients, and will represent a baseline for our predictions. The accuracy of the baseline is 0.64.

We focus solely on interpretable models: CART, OCT and OCT-H. The initial challenge with all 3 models was that after cross-validation they were never predicting class=2 (low-dose treatment), as it is the lowest-occurring option in the dataset. This has been addressed by implementing penalty matrices or slightly compromising performance in favor of a model that predicts all three classes. We believe detecting low-dose cases is important, rather than defaulting to a high-dose which can be potentially dangerous for the patient or no-treatment which does not address the symptoms.

4.2.1 Classification Trees

We performed cross-validation and selected a model with minbucket=35 and cp=0.0025 (please see appendix). The model splits primarily on variables that indicate no serious adverse effects. Intuitively this means that the model chooses to differentiate patients based on whether they experience serious adverse effects (e.g. cardiac arrest, diabetes, cardio-pulmonary distress) or not, regardless of which specific adverse effect they experience. We further interpret that those with a high utility of no serious adverse effect should receive option 1 (high-dose), which makes intuitive sense. These patients are most likely to experience no adverse effects, and hence we can address their symptoms with the strongest treatment.

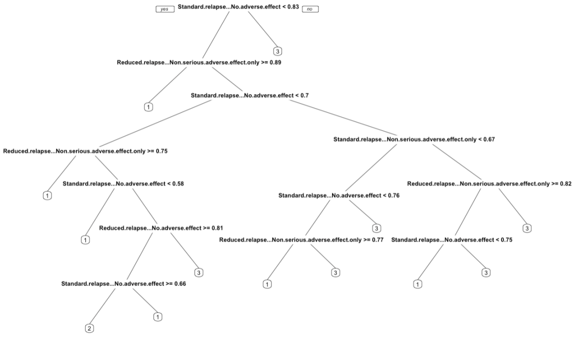

yes
Standard.relapse...No.adverse.effect < 0.83
no
3
Reduced.relapse...Non.serious.adverse.effect.only >= 0.89
1
Standard.relapse...No.adverse.effect < 0.7
Reduced.relapse...Non.serious.adverse.effect.only >= 0.75
Standard.relapse...Non.serious.adverse.effect.only < 0.67
1
Standard.relapse...No.adverse.effect < 0.58
Standard.relapse...No.adverse.effect < 0.76
Reduced.relapse...Non.serious.adverse.effect.only >= 0.82
1
Reduced.relapse...No.adverse.effect >= 0.81
3
3
Reduced.relapse...Non.serious.adverse.effect.only >= 0.77
Standard.relapse...No.adverse.effect < 0.75
3
Standard.relapse...No.adverse.effect >= 0.66
1
3
1
3
2
1

Figure 1: CART Model

We implemented a loss matrix to ensure class=2 (low-dose treatment) was being predicted by the model. As mentioned this is because we believe less invasive treatment to be beneficial. We over penalize misclassification of low-dose, and we also consider the trade-off between prescribing treatment when we should not versus underprescribing (please see appendix). The best accuracy of a CART model was 70%. The advantage of CART is that it is a simple model, runs fast, and in this case performs very similarly to OCT and OCT-H.

4.2.2 Optimal Classification Trees

We next attempted an OCT model and ran cross-validation, obtaining the parameters: minbucket=35, cp=0.003, and, max depth=4. Cross-validated accuracy was 0.71, however, we encountered the same issue of never predicting low-dose, therefore we decided on a final model of depth 5 with an accuracy score of 0.70.

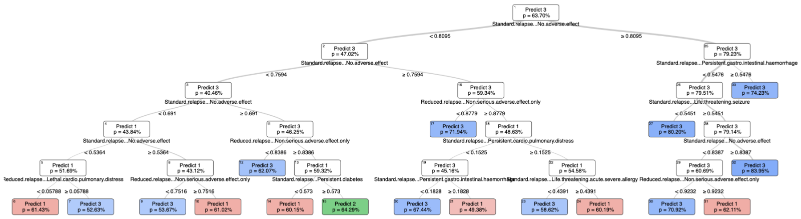


Figure 2: OCT Model

Similarly to CART, we split on the variables that detect the absence of serious adverse effects, yet we see new variables such as gastrointestinal hemorrhage, seizure, and cardio-pulmonary distress. OCT does not make greedy splits in the way CART does, enabling it to output splits across more features, providing us with richer information. OCT also is more interpretable than CART, having only 5 layers, as opposed to 7. Since the accuracy scores are the same, we favor OCT over CART.

4.2.3 Optimal Classification Trees with Hyperplanes

Finally, we looked at OCT-H hoping to improve performance, yet the model performed very similarly to both CART and OCT, improving only to 0.71 accuracy. The chosen model also has depth = 5, as opposed to the cross-validated model with depth = 2. We chose this for the same reason that the cross-validated tree did not predict low-dose treatment.

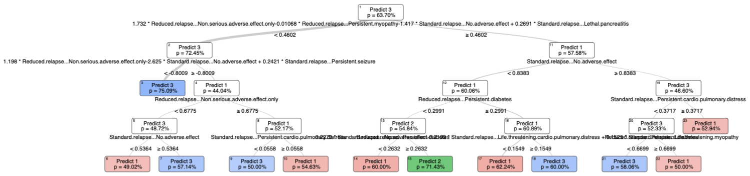


Figure 3: OCT-H Model

The drawbacks of OCT-H are the reduced interpretability (e.g. here most splits have 4 variables) and the significant runtime for cross-validation (over 3 hours).

4.2.4 Multinomial Logistic Regression

We also performed a multinomial logistic regression on this problem giving an out-of-sample accuracy slightly lower, at 0.59. The main advantage of this method is that it can be robustified adding a lasso regularizer term. This model has a lower performance than the baseline (0.64), however, it has proven itself useful in the robustification process.

5 Sensitivity analysis to perturbations

The goal of sensitivity analysis was to test the models against changes in the underlying data and check the impact on accuracy as we increase the magnitude of changes. We first calculated the means m_j and standard deviation σ_j of each predictor (patient utility) **j**. Then, we progressively perturbed the test dataset. For each feature column **i**, for each patient **j**, having the utility u_{ij}, we associate the perturbation p_{ij}. Hence, u_{ij} becomes $u'_{ij} = u_{ij} + p_{ij}$ where p_{ij} is generated randomly with the normal distribution of mean m_j and standard deviation $p\sigma_j$ where p in the range of perturbation, up to 3 in our case. By progressively perturbing within $1\sigma_j$, $2\sigma_j$ and $3\sigma_j$ we test how the accuracy changes at different stages.

5.1 K-means prescriptive-predictive model

The impact of the perturbation on the k-means predictive-prescriptive method gives us the following results:

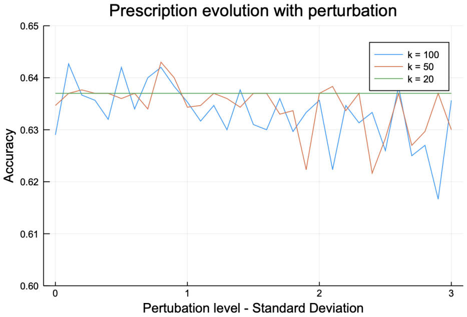


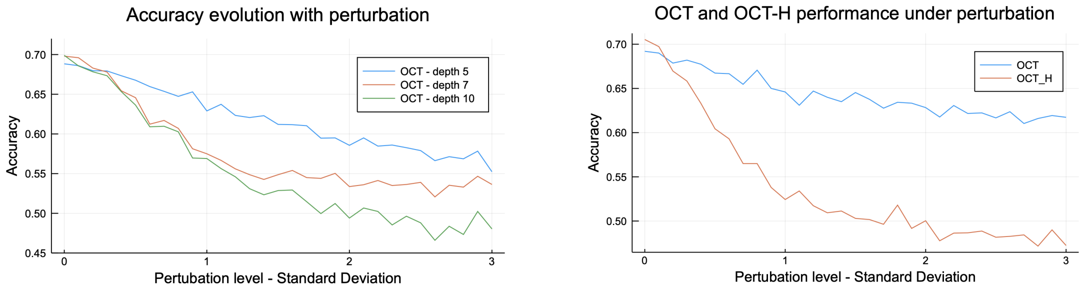
Figure 4: k-means perturbation impact

This shows an underlying characteristic of the problem: for low values of clusters, the most common treatment is the same for all clusters. This also corresponds to the reference study of this project where one treatment was given to all the patients. For a high number of clusters, the accuracy remains at levels of 0.62 and 0.64 which means we predict treatment 1 for most clusters, but not all.

5.2 Sensitivity Analysis of the Tree Models

5.2.1 Sensitivity evolution with the tree depth and complexity

After building the initial models, we vary their tree depth and assess their sensitivity to out-of-sample perturbation.



(a) OCT depth impact

(b) OCT vs. OCT-H performance

Figure 5: Perturbation impact

The OCT is highly sensitive to perturbation in the test set with the evolution of the perturbation range, accuracy decreasing from 0.70 to 0.55 when the data is perturbed within a range of 3σ. The perturbation impact is higher for deeper trees, which makes sense as the deeper the tree is, the more complex it is and sensitive to external perturbations. Similarly, OCT-H is more sensitive to perturbations than OCT. This also matches our intuition as OCT-H is more complex and fits better the data in-sample but is sensitive to out-of-sample perturbations.

5.2.2 Sensitivity evolution with the number of features perturbed

Another approach is to assess the sensitivity evolution with the number of features perturbed. A first step was to identify which features to perturb. We already have the feature importance from CART and OCT, as well as a random forest model that we ran:

Features	Mean Decrease Gini
Standard.relapse.no.adverse.effect	291
Standard.relapse.non.serious.adverse.effect.only	182
Reduced.relapse.non.serious.adverse.effect.only	173
Reduced.relapse.persistent.diabetes	71
Reduced.relapse.persistent.osteonecrosis	72

The top three features indicate no adverse effect or no serious adverse effect, conditioned by the progression of the MS disease (standard versus reduced relapse). In essence, high values in these features signal a relatively healthy patient with high chance of no adverse effect from the methylprednisolone treatment. This is different from the rest of the predictors, where high values indicate a high risk (e.g. of diabetes). The aim is to verify if there is a correlation between the feature's importance and the impact of its perturbation on the model's performance.

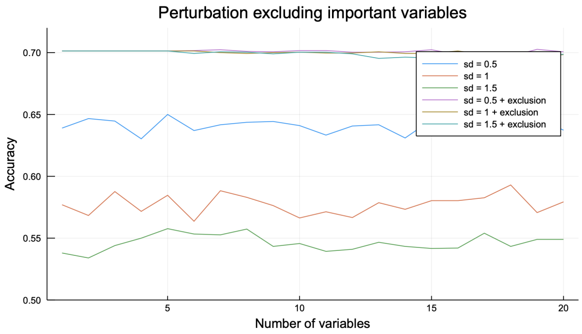


Figure 6: Perturbation by number of variables

In the plot above the main three lines (blue, red and green) corresponds to results of perturbing the three features with the highest importance (as taken from the table above) at different perturbation levels. The other lines at the top of the chart highlight perturbation in the rest of the features, excluding the three most important ones. This gives two very interesting insights, the first one being that only the three most important features influence the model's performance as there is almost no impact while perturbing a higher and higher number of features. This also could be seen in the perturbations including all the features, as this perturbation does not evolve a lot with the number of features, implying that the perturbation comes only from a sparse number of parameters. A second observation is that, when the important features are excluded, whatever the perturbation intensity is in terms of the number of standard deviations, the impact remains the same. These observations enable us to conclude that if we precisely control the three most important features our model will be very robust to out-of-sample perturbations.

We can also draw another interesting conclusion - the top three features are sufficient to predict the best treatment to a high accuracy (e.g. CART split on these 3 features alone and achieved 0.70 accuracy). Currently a high-dose treatment is prescribed to all patients, perhaps justified by the difficulty in measuring all 56 of the predictor variables. Since the top three features are sufficient, we could only measure those for each patient. In practice, this may prove challenging since the features are ”no adverse effect” which technically still implies we must look at all adverse effects, yet it is a valuable insight that if we find one single serious adverse effect this is enough information to make a prediction.

6 Models robustification and results

6.1 Impact of Multinomial Logistic Regression

First, we assessed the performance of the multinomial logistic regression by introducing a lasso regularizer. Hence, we implement the following formulation:

$$\max_{\beta,\beta_0} \quad -\sum_{i=1}^{n} log(1+e^{-y_i(\beta^T x_i+\beta_0)+\lambda||\beta||_1})$$

A cross-validation on the training set gave us 0.01 as an optimal value for λ which lead to an out-of-sample accuracy of 59% without perturbation. Now we perturb the data and see how the performance is evolving:

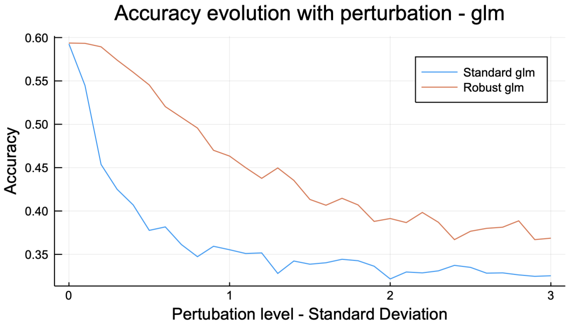
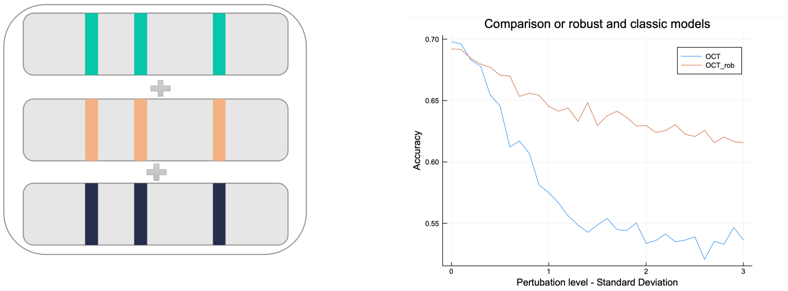


Figure 7: Robust Logistic Regression vs. standard performance

A cross-validation on the training set gave us 0.01 as an optimal value for λ which led to an out-of-sample accuracy of 59% without perturbation. Now we perturb the data and see how the performance is evolving:

6.2 Robustifying the training set - OCT

We robustify the model by generating a training set that includes perturbed scenarios. We first select the three most important variables. We generate three additional training sets, the first one where the three most important variables have been perturbed to a level of 0.5σ (in green), the second one to a level of σ (in orange) and the third one to a level of 1.5σ (in blue). We concatenate these three training sets with the initial one to build a final training set. Finally we retrain the model on this concatenated set.



(a) Building the robust training set (b) Robustification of OCT training set

Figure 8: OCT Robustification

With an OCT of depth 5 there is a clear improvement due to robustification. In highly perturbed scenarios the accuracy stays at 0.62 whereas the non-robust model falls to 0.55. It is very interesting to observe that this comes at no cost for low perturbations as both models stay at a level of 0.70. We also conclude that this model is far more robust than the regularized logistic regression as it maintains accuracy at a level of 0.60 whereas regularized logistic regression falls below 0.40. Lastly, in highly perturbed scenarios, the robust model remains at an accuracy similar to the baseline, while still predicting all 3 treatment classes (see appendix). This represents an improvement as we are able to differentiate patients and prescribe less-invasive methods, as opposed to current practice which prescribes high-dose to all patients.

7 Conclusion: the trade-off between model robustification and interpretability

To conclude, this project had two major contributions. The first one is building models allowing us to predict the best treatment to a patient with accuracy levels of 0.70 (improvement from 0.64 baseline) . This personalized approach improved previous practice of simply prescribing the most common effective treatment to all patients. We also note the three top features (i.e. presence of any serious adverse effects) which are sufficient to make highly accurate predictions.

The second contribution is the robustification of the OCT model by designing a new training set. This includes perturbed scenarios and it is based on a sparse number of features (the three most important ones). The major observation is that it came at no cost in unperturbed scenarios while remaining much more robust in perturbed ones. This outperforms models that are already robust such as regularized multinomial logistic regression.

Nevertheless, this robustification design came at a cost on the interpretability of our models. We can associate a price to the interpretability of models as defined in the paper [3]. Here interpretability is reduced since we train on a concatenated dataset which includes the same patient observation multiple times (with perturbations). An interesting area of research to explore is which price could be associated with this process and which formulation could be developed to include this price. For example, the interpretability penalty can be added as part of the MIO formulation of an OCT or as a regularizer term in logistic regression.

8 References

[1] National Multiple Sclerosis Society. https://www.nationalmssociety.org/What-is-MS/Who-Gets-MS. Published 2019. Accessed December 7, 2019.

[2] Caster, O., Edwards, R. Quantitative benefit-risk assessment of methylprednisolone in multiple sclerosis relapses. BMC Neurology, 2015

[3] Bertsimas D., Delarue A., Jaillet P., Martin S. The Price of Interpretability. July 9, 2019.

9 Appendix

The code for this project can be found at: `https://github.mit.edu/siacob/ml_project`

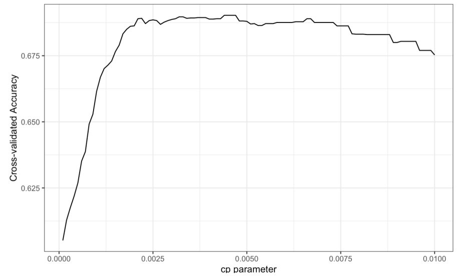
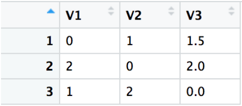


Figure 9: CART Cross-Validation cp parameter vs. Accuracy



	V1	V2	V3
1	0	1	1.5
2	2	0	2.0
3	1	2	0.0

Figure 10: CART Loss Penalty Matrix

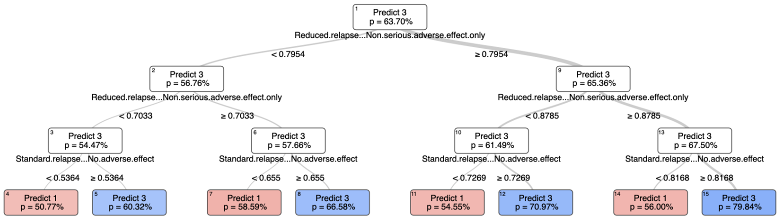


Figure 11: Cross-Validated OCT

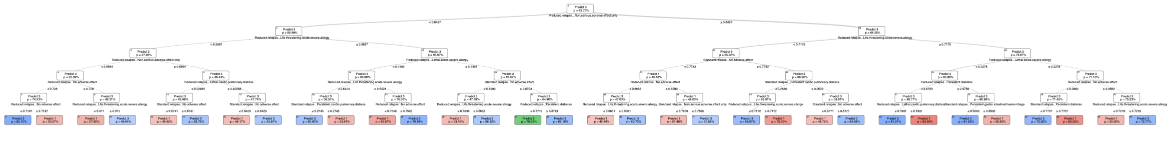


Figure 12: Robust OCT in high perturbation scenario