Applying frontal cortex metabolites quantified by 7-Tesla ^{1}H-MRS to predict multiple sclerosis subtype through recursive partitioning and conditional inference trees

Isabel Abonitalla[1], Kelley M. Swanberg[1,2], Hetty Prinsen[2], and Christoph Juchem[1,2,3,4]

[1]Department of Biomedical Engineering, Columbia University Fu Foundation School of Engineering and Applied Science, 1210 Amsterdam Ave., New York, NY 10027, United States; [2]Department of Radiology and Biomedical Imaging, Yale University School of Medicine, 330 Cedar Street, New Haven, CT 06520, United States; [3]Department of Radiology, Columbia University School of Medicine, 622 W 168 Street, New York, NY 10032, United States; [4]Department of Neurology, Yale University School of Medicine, 330 Cedar Street, New Haven, CT 06520, United States

Synopsis

Multiple sclerosis (MS) is an autoimmune disease that impairs the central nervous system by attacking the myelin sheaths of neurons. It affects more than 2.3 million people worldwide. One potential key to understanding the metabolic differences in the brains of MS patients and potentially treating MS is investigating the relationship between different metabolites and the types of MS. Data on the different metabolite levels of patients with relaxing-remitting MS (RR-MS), progressive MS (P-MS), and controls have been previously obtained from the frontal cortex of adults using ^{1}H-MRS at 7 Tesla. We created classification trees and performed tree-boosting on these data using the R programming language. Our results exhibited 43% accuracy in predicting whether an individual had relapsing-remitting, progressive, or no MS; they also showed that glycine and total choline were the strongest factors in this classification. Our data highlight the importance of continued investigation in the potential applications of machine learning in analyzing data from biomedical research.

Purpose

Multiple Sclerosis (MS) is an autoimmune disease that impairs the nervous system by attacking the myelin sheaths of nerve cells[1]. Although it affects more than 2.3 million people worldwide[1], the only way it is diagnosed is either through the McDonald Criteria or through a differential diagnosis. The McDonald Criteria falls short in that it cannot be applied to patients who do not present symptoms during an MR scan. A potential supplement to this is using levels of different metabolites potentially implicated in MS pathogenesis as a predictor for the subtype of MS that a patient has. In this study, we used machine learning to analyze the metabolite levels measured in both RR-MS and Progressive P-MS patients. Machine learning is unique in that it finds patterns in the data not easily found by humans. Furthermore, it is unbiased and is able to learn from previous iterations of the algorithms it uses[2].

Methods

Data Collection

The analysis was applied to a dataset from a previous experiment that [summarize experiment here][3]. Out of 68 patients, 25 were healthy controls, 26 had relapsing-remitting multiple sclerosis (RR-MS), and 21 had progressive multiple sclerosis (P-MS). Out of the healthy controls, 12 were male and 13 were female; their average age was 44 years. Out of the RR-MS patients, 18 were female and 8 were male; they also had an average age of 44 years. Out of the P-MS patients, 12 were female, 9 were male; their average age was 55 years.

Proton signals of small molecule metabolites from a $3\times3\times3$ cm voxel in the prefrontal cortex were measured using a macromolecule-suppressed STEAM sequence (TE, TM, TR) at 7 Tesla and quantified as concentration ratios referenced to 10 mM total creatine using linear combination model fitting[5] to simulated basis functions in INSPECTOR[4], as previously reported.

Analysis

Our data were randomly separated into a training set (with 18 healthy controls, 19 RR-MS, and 13 P-MS) and a test set (with 6 healthy controls, 6 RR-MS, and 6 P-MS). Concentration ratios with Cramer-Rao lower bounds greater than 20% were removed from both training and test sets. Exploratory analyses were performed on the data, which include creating boxplots for all metabolites in relation to the patient conditions. The function rpart[6] (recursive partitioning) was used on the training set to produce binary prediction trees, which were pruned to avoid overfitting. Pruning reduces the size of the prediction tree by removing the branches that are too complex to provide accurate classification[7]. Afterwards, the function ctree[8] was used on the training set to produce similar conditional inference trees. These trees did not require pruning as the ctree algorithm already includes parameters to prevent overfitting. Cross-validation (10-fold; 100 iterations) was performed on both trees to estimate accuracy. Moreover, tree-boosting[9] was performed to find the relative influence of each metabolite.

Results

The accuracy of the function r-part was 60% before cross-validation and 34% after. The accuracy of the function ctree was 58% before cross-validation and 38% after. The results of tree-boosting are yet to be measured for accuracy.

$$Accuracy = \frac{correctly\ predicted\ patient\ type}{total\ number\ of\ patients} * 100\%$$

On the test set, whose accuracy would be most definitive of the performance of an algorithm, rpart had an accuracy of 56% while ctree exhibited an accuracy of 50%.

Gradient boosting machine algorithms showed glycine had the most influence on the classification of patient types. GBM creates multiple decision trees and averages out those trees to create the optimum prediction tree. The metabolite that has the strongest effect in classifying patients is calculated using the equation below. j refers to a variable in tree T with L splits.[9]

$$Influence_j(T) = \sum_{i=1}^{L-1} I_i^2 1(S_i = j)$$

Conclusions

Recursive partitioning exhibited the best results in terms of predicting patient conditions using metabolite levels. An accuracy of 60% is not good enough for many, but it is much better than chance, so there is potential for further studies in using machine learning to analyze the results of neurological studies. Other potential improvements and further research options include studying tree-boosting further as initial results showed it having a high accuracy. Furthermore, we should look at other algorithms that are not similar to k-means clustering as it is not appropriate for our data.

Acknowledgements

Special thanks are due to Ms. Yvette Strong at the Yale Center for Clinical Investigation and the nursing staff at the Yale-New Haven Hospital Interventional Immunology Clinic for their invaluable help in patient recruitment. This research was made possible by the National Multiple Sclerosis Society (NMSS) grant “In Vivo Metabolomics of Oxidative Stress with 7 Tesla Magnetic Resonance Spectroscopy“ (RG 4319) and falls within the purview of Yale Medical School Human Investigation Committee protocol #1107008743.

References

[1] WHO. Atlas Multiple Sclerosis resources In The World 2008. WHO Press 2008:56. doi:ISBN 978 92 4 156375 8.

[2] Alpaydın E. Introduction to machine learning. vol. 1107. 2014. doi:10.1007/978-1-62703-748-8-7.

[3] Swanberg KM, Prinsen H, Coman D, Graaf RA De, Juchem C. In vivo quantification of glutathione T 2 in the human brain at 7 Tesla using echo time extension with variable refocusing selectivity and symmetry. J Magn Reson Imaging 2017.

[4] Swanberg KM, Prinsen H, Fulbright RK, Pitt D, Bailey M, Juchem C. Towards in vivo neurochemical profiling of multiple sclerosis with MR spectroscopy at 7 Tesla : Cross-sectional assessment of frontal-cortex glutathione , GABA , and glutamate in

individuals with relapsing-remitting and progressive multiple sclerosis. J Magn Reson Imaging 2017;25:0–5.

[5] Prinsen H, de Graaf RA, Mason GF, Pelletier D, Juchem C. Reproducibility measurement of glutathione, GABA, and glutamate: Towards in vivo neurochemical profiling of multiple sclerosis with MR spectroscopy at 7T. J Magn Reson Imaging 2017;45:187–98. doi:10.1002/jmri.25356.

[6] Therneau TM, Atkinson B, Ripley BD. rpart: Recursive Partitioning. Rpart Packag Man 2006.

[7] Fürnkranz J. Pruning Algorithms for Rule Learning. Mach Learn 1997;27:139–72. doi:10.1023/A:1007329424533.

[8] Hothorn T, Hornik K, Zeileis A. ctree: Conditional Inference Trees. CranAtR-ProjectOrg 2006.

[9] Ridgeway G. Generalized Boosted Models: A guide to the gbm package. Compute 2007. doi:10.1111/j.1467-9752.1996.tb00390.x.

Figure 1. Metabolite concentration ratios were measured from the frontal cortex of healthy controls as well as patients with multiple sclerosis. Macromolecule-suppressed STEAM (TE, TM, TR) and the values were quantified using linear combination modeling relative to 10 mM of total creatine.

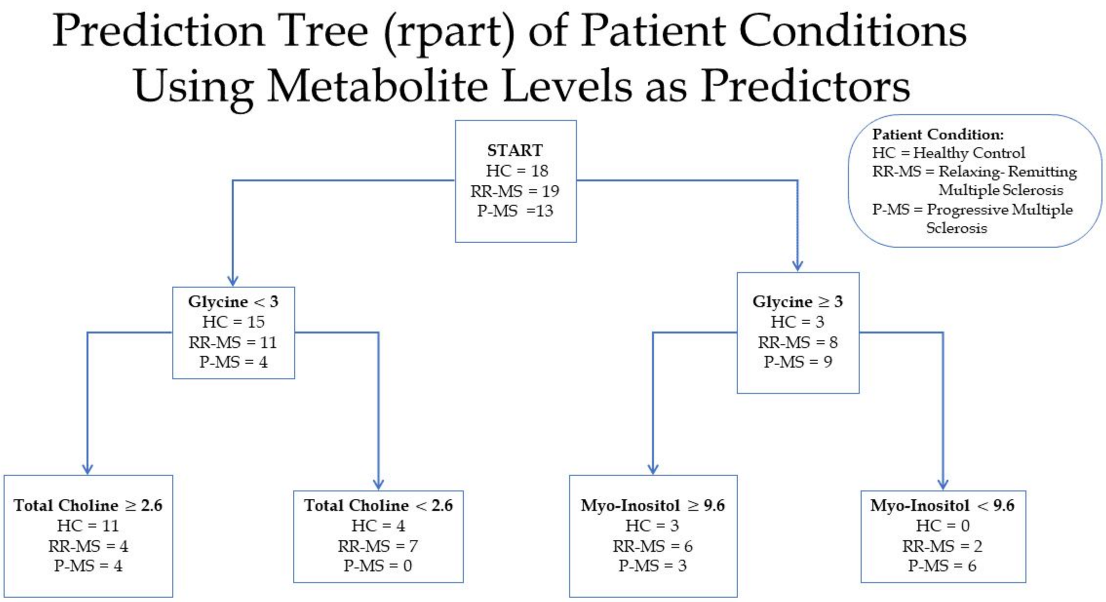

Figure 2. The function rpart was used in R to produce a prediction tree, which was pruned and cross-validated. It shows that glycine, myo-inositol and the combination of choline and phosphocholine to be the best predictors of the type of MS a patient has. The model exhibited 56% accuracy used on the test set.

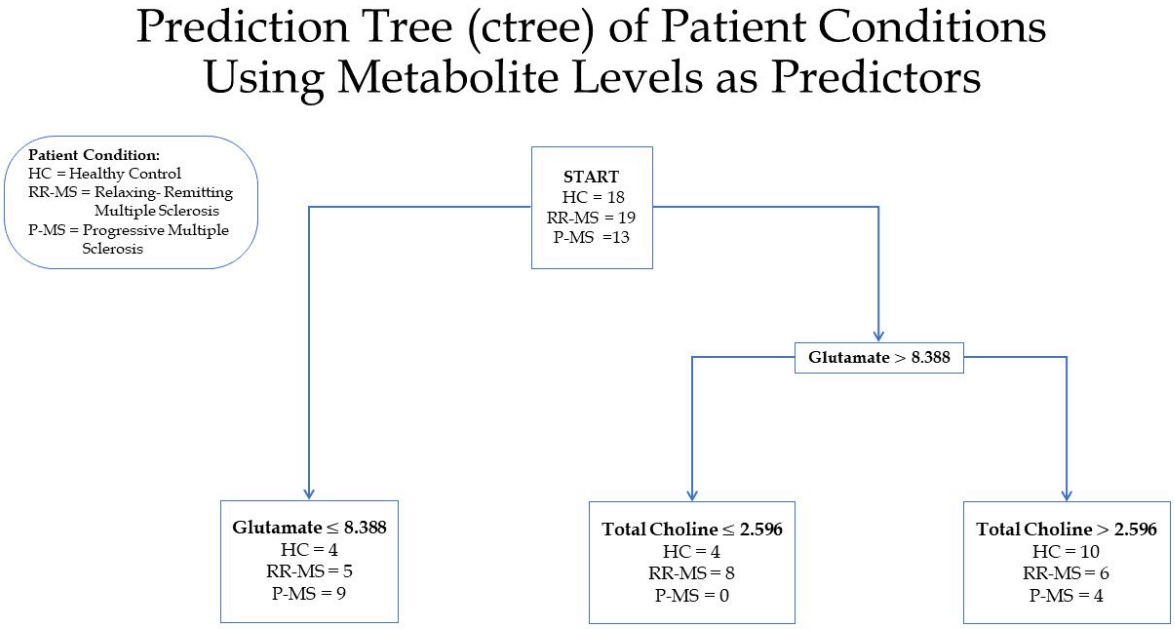

Figure 3. The function ctree was used to create a prediction tree, which was then cross-validated (10-fold; 100 iterations). The model had a 50% accuracy when used on the test set.