**Introduction to Study and Data**

Understanding the pathways that a particular disease takes is the first step in slowing and reversing its progression. In order to resolve these pathways, research must attempt to model the progression of the disease. This can be achieved by identifying factors that suggest a particular outcome. Alzheimer’s disease (AD) is not yet well understood so there is a demand for identifying these risk factors.

Predicting transitions from a particular diagnosis to another is of the utmost [try to avoid using such superlatives. Maybe just Predicting transitions from a particular diagnosis to another is a useful clinical goal and should help in modeling…] importance in modeling the progression of AD. The data from this study was taken from the Alzheimer’s Disease Neuroimaging Database, part of the NIH. There were three major diagnoses in this study. These patients were determined by rigorous mental health assessments to be “healthy” or to have “mild” or “severe” AD. The relevant transitions in this study were among these three classes. In addition to their diagnoses, dozens of measurements were recorded on each patient’s first visit, including genetic and hemic biomarkers, demographics, and environmental factors like educational history. Finally, patients were monitored over a period of six months to five years, and their transition or stability was observed.

It is no doubt of extreme [again, too superlative. Something like “An important clinical goal is to identify…” importance to identify the prognosis of a healthy patient. Nevertheless, the prognosis of a mild patient poses [not the right verb] a powerful clinical question as well. The current data were used to support the latter study, as the abundance of these patients allowed for a more robust study. Such is shown by the diagnostic distributions below. In this study, the patients whose diagnosis was initially mild were considered: some did not get worse (145 stable), while others transitioned to severe AD (164 transition).

*Table 1.* Patient populations.

|  |  |
| --- | --- |
| **Patient diagnosis over all visits** | **Number of patient**s |
| Stable healthy | 36 |
| Transition healthy to mild AD | 12 |
| Stable mild AD | 145 |
| Transition mild to severe AD | 164 |
| Stable severe AD | 90 |
| Total | 447 |

**Methods**

The goal of this study was to determine which factors were most closely linked with the progression of patients from the mild to the severe diagnosis.

The data for these patients, including their diagnoses and all available measurements (or “predictors”), was imported into MATLAB. Using MATLAB’s CART software, and labeling stable patients a ‘1’ and transitional patients a ‘2’, two types of decision trees were created. The first was the “full tree”, which uses all patients to train the tree and tests the tree’s performance on those same patients. The second tree was a “leave-out-one” tree, where all patients but one were used to train, and the tree’s performance was tested on the left-out patient (this process was iterated for each patient).

The performance of these trees was measured by their confusion matrices, as well as precision and recall. It was expected that the full tree would perform better on the data than the leave-out-one tree, because the same patients that were tested on the tree had been used to train the tree. However, a significant [substantial would be a better word here as significant has a specific statistical meaning that you do not really require here] difference between the performance of these trees might suggest overfitting. In this case, the reliability of the results would be questionable because the ultimate goal is to predict unknown prognoses from given predictors.

**Results**

The full tree has proven extremely accurate from the beginning. Below are the figures for both trees using all of the predictors, including the confusion matrices for each class and for each tree overall. Precision within a given class was calculated as the number of true positives divided by the union of all positive guesses for that class. Recall within a class was calculated as the number of true positives divided by the number of all truly positive patients in that class, i.e. true positives + false negatives. Total precision was calculated as the sum of the precision of each class times the frequency of that class in the larger population (i.e. p1\*f1+p2\*f2); total recall was calculated similarly.

Initial Confusion Matrices & Precision/Recall

*Table 2a-d Full Tree*

|  |  |  |
| --- | --- | --- |
| Trans. ‘2’ | Positive | Negative |
| True | 160 | 135 |
| False | 7 | 3 |

|  |  |  |
| --- | --- | --- |
| All | Positive | Negative |
| True | 295 | 295 |
| False | 10 | 10 |

|  |  |  |
| --- | --- | --- |
| Stable ‘1’ | Positive | Negative |
| True | 135 | 160 |
| False | 3 | 7 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mild stable | Mild transition | Total |
| Number of patients | 142 | 163 | 305 |
| Frequency | 0.466 | 0.534 | 1.00 |
| Precision | 97.8% | 95.8% | 96.8% |
| Recall | 95.1% | 98.2% | 96.7% |

*Table 3a-d Leave-out-one Tree*

|  |  |  |
| --- | --- | --- |
| Trans. ‘2’ | Positive | Negative |
| True | 85 | 76 |
| False | 66 | 78 |

|  |  |  |
| --- | --- | --- |
| All | Positive | Negative |
| True | 161 | 161 |
| False | 144 | 144 |

|  |  |  |
| --- | --- | --- |
| Stable ‘1’ | Positive | Negative |
| True | 76 | 85 |
| False | 78 | 66 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mild stable | Mild transition | Total |
| Number of patients | 142 | 163 | 305 |
| Frequency | 0.466 | 0.534 | 1.00 |
| Precision | 49.4% | 56.3% | 53.1% |
| Recall | 53.5% | 52.2% | 52.8% |

The total precision and recall for the full tree (96.8%, 96.7%) are vastly greater than those of the leave-out-one tree (53.1%, 52.8%). These data strongly indicate that the results were not credible in a clinical setting.

Further, two tools were used to test the null hypothesis. The first was to get an approximate p-value by randomly permuting each patient’s true prognosis, computing the precision and recall of this prediction, and comparing it to the precision and recall of either tree. This was repeated 1000 times and the results were given as the number of times the random permutation prediction was better than either of the trees.

*Table 4.* Random permutation vs. Trees

|  |  |  |  |
| --- | --- | --- | --- |
|  | Permutation Values | # times better than full tree | # times better than leave-out-one tree |
| Precision | 50.2% +/- 2.9% | 0/1000 | 156/1000 |
| Recall | 50.2% +/- 2.9% | 0/1000 | 229/1000 |

Although the full tree was successful, the leave-out-one tree was far from it. Random permutation was better by measure of precision 15.6% of the time, and better by measure of recall 22.9% of the time. The success of the full tree as well as the failure of the leave-out-one tree are a second strong indicator of overfitting.

The second test was to guess that all of the predictions were ‘1’s or ‘2’s. This mode of guessing represents [corresponds to] a crude tree with a single node and no splits. Precision and recall were measured for these trees below.

*Table 5a-b.* Single-valued tree statistics.

*a. Tree of all 1’s.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mild stable | Mild transition | Total |
| Number of patients | 142 | 163 | 305 |
| Frequency | 0.466 | 0.534 | 1.00 |
| Precision | 46.6% | 0% | 21.7% |
| Recall | 100% | 0% | 46.6% |

*b. Tree of all 2’s.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mild stable | Mild transition | Total |
| Number of patients | 142 | 163 | 305 |
| Frequency | 0.466 | 0.534 | 1.00 |
| Precision | 0% | 53.4% | 28.6% |
| Recall | 0% | 100% | 53.4% |

Any tree destined for a clinical setting must do much better in terms of precision and recall. This result is unsurprising, as this method simply guesses the same prognosis for every patient. However, this result does speak to the shortcomings of the leave-out-one tree: while the precision in 5a-b is low (21.7%, 28.6%), the recall in 5a-b (46.6%, 53.4%) is dangerously close or better than that in table 3d (52.8%).

**Predictor Importance & Optimizing Trees**

Ultimately, these results are only used to build confidence about our predicting power. The use in assuring predicting power, however, is to determine the profile of factors used to make that prediction. This profile was determined by software that quantifies predictor importance, which calculates the difference in mean squared error due to nodes associated with that predictor.

It was considered [We considered (more active voice please)] whether further resolution in the predictors would filter out extraneous or unimportant predictors, which might be responsible for the depth and overfitting of the trees. In order to obtain this resolution, the importance of each predictor was considered for the full tree. The importance for each predictor was then converted to deviations from the mean.

The patient tree [is this the full tree? If so we really should rerun all this on 90% of the data and then make our test on the last 10%. Otherwise, we are polluting our testing set.] was then tested, filtering out some of the predictors based on their importance values from the full tree (1 std below the mean to 2 std above the mean, in steps of 0.2). The following precisions and recalls were obtained over this iteration.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold (# std’s above mean importance) | # variables considered | Precision Full Tree | Recall Full Tree | Precision Leave-out-one | Recall Leave-out-one |
| -1.0 std | 54 | 0.9675 | 0.9672 | 0.5306 | 0.5279 |
| -0.8 std | 54 | 0.9675 | 0.9672 | 0.5306 | 0.5279 |
| -0.6 std | 27 | 0.9675 | 0.9672 | 0.6493 | 0.6492 |
| -0.4 std | 26 | 0.9675 | 0.9672 | 0.6556 | 0.6557 |
| -0.2 std | 23 | 0.9607 | 0.9607 | 0.6350 | 0.6328 |
| 0.0 std | 18 | .09576 | 0.9574 | 0.6637 | 0.6623 |
| 0.2 std | 14 | 0.9484 | 0.9475 | 0.6200 | 0.6197 |
| 0.4 std | 14 | 0.9484 | 0.9475 | 0.6200 | 0.6197 |
| 0.6 std | 12 | 0.9103 | 0.9082 | 0.6135 | 0.6131 |
| 0.8 std | 10 | 0.9273 | 0.9246 | 0.6583 | 0.6590 |
| 1.0 std | 9 | 0.9264 | 0.9246 | 0.5928 | 0.5934 |
| 1.2 std | 7 | 0.9084 | 0.9049 | 0.6152 | 0.6164 |
| 1.4 std | 7 | 0.9084 | 0.9049 | 0.6152 | 0.6164 |
| 1.6 std | 5 | 0.9025 | 0.9016 | 0.6014 | 0.6033 |
| 1.8 std | 4 | 0.8756 | 0.8754 | 0.5258 | 0.5311 |
| 2.0 std | 4 | 0.8756 | 0.8754 | 0.5258 | 0.5311 |

*Table 6.* Precision and recall as a function of threshold importance.

The two underlined trials were local maxima for precision and recall on the leave-out-one tree. These trials were selected for further resolution of predicting [prediction] capacity. In order to do this, the threshold importance values associated with these maxima were used to filter out unimportant variables, and trees were made based on the remaining variables. The accuracy profile, shown below, of each of these trees was assessed.

*Table 7a-g.* Tree statistics. Full tree confusion matrices (a-c), leave-out-one confusion matrices (d-f), and precision/recall for each tree (g), using threshold importance of 0.0 stds above the mean.

|  |  |  |
| --- | --- | --- |
| Trans. ‘2’ | Positive | Negative |
| True | 155 | 137 |
| False | 5 | 8 |

|  |  |  |
| --- | --- | --- |
| All | Positive | Negative |
| True | 292 | 292 |
| False | 13 | 13 |

|  |  |  |
| --- | --- | --- |
| Trans. ‘2’ | Positive | Negative |
| True | 108 | 94 |
| False | 48 | 55 |

|  |  |  |
| --- | --- | --- |
| All | Positive | Negative |
| True | 202 | 202 |
| False | 103 | 103 |

|  |  |  |
| --- | --- | --- |
| Stable ‘1’ | Positive | Negative |
| True | 137 | 155 |
| False | 8 | 5 |

|  |  |  |
| --- | --- | --- |
| Stable ‘1’ | Positive | Negative |
| True | 94 | 108 |
| False | 55 | 48 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mild stable ‘1’ | Mild transition ‘2’ | Total |
| Number of patients | 142 | 163 | 305 |
| Frequency | 0.466 | 0.534 | 1.00 |
| Full Tree | Precision | 94.5% | 96.9% | 95.8% |
| Recall | 96.5% | 95.1% | 95.7% |
| Leave-out-one | Precision | 63.1% | 69.2% | 66.4% |
| Recall | 66.2% | 66.3% | 66.2% |

*Table 8a-g.* Tree statistics. Full tree confusion matrices (a-c), leave-out-one confusion matrices (d-f), and precision/recall for each tree (g), using threshold importance of 0.8 stds above the mean.

|  |  |  |
| --- | --- | --- |
| Trans. ‘2’ | Positive | Negative |
| True | 158 | 124 |
| False | 18 | 5 |

|  |  |  |
| --- | --- | --- |
| All | Positive | Negative |
| True | 282 | 282 |
| False | 23 | 23 |

|  |  |  |
| --- | --- | --- |
| All | Positive | Negative |
| True | 201 | 201 |
| False | 104 | 104 |

|  |  |  |
| --- | --- | --- |
| Stable ‘1’ | Positive | Negative |
| True | 124 | 158 |
| False | 5 | 18 |

*Table 9.* Random permutation experiment, number of instances of

|  |  |  |
| --- | --- | --- |
| Stable ‘1’ | Positive | Negative |
| True | 87 | 114 |
| False | 49 | 55 |

|  |  |  |
| --- | --- | --- |
| Trans. ‘2’ | Positive | Negative |
| True | 114 | 87 |
| False | 55 | 49 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mild stable ‘1’ | Mild transition ‘2’ | Total |
| Number of patients | 142 | 163 | 305 |
| Frequency | 0.466 | 0.534 | 1.00 |
| Full Tree | Precision | 96.1% | 89.8% | 92.7% |
| Recall | 87.3% | 96.9% | 92.5% |
| Leave-out-one | Precision | 64.0% | 67.5% | 65.8% |
| Recall | 61.3% | 69.9% | 65.9% |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Permutation Values | # instances better than full tree | # instances better than leave-out-one  |
| Precision | 50.2% +/- 2.9% | 0/1000 | 0/1000 |
| Recall | 50.2% +/- 2.9% | 0/1000 | 0/1000 |

**Discussion**

The full tree lost a few percentage points of accuracy on its precision and recall, as compared to when minor degree of accuracy (precision down to 92.7% from 96.8% and recall down to 92.5% from 96.7%). There were two leave-out-one trees that scored particularly well, and resolution of these trees showed favorable results. The leave-out-one tree gained as much as 13% in precision and recall, as compared to with no importance filtering. The two leave-out-one trees do better than the random permutation experiment and the single-value predictions every time and by a larger margin than initial tests showed.

With these results, there is a greater degree of reliability from these trees. Overfitting remains a concern, but with the leave-out-one trees doing well, the predicting [prediction] power is more clear [what does “more clear” mean?]. Between the two leave-out-one trees, the one with a lower threshold (importance > 0.0 std’s) did slightly better on the full tree, but this tree also had more variables. Fewer variables suggests a lesser possibility of overfitting, so the tree with the higher threshold (importance > 0.8 std’s) is likely of greater future interest.

Finally, the importance values of the predictors on the full trees are given below. As explained, only predictors scoring above a threshold importance on the full tree were used to rebuild the tree. Then, the predictors that are shown are those with an importance +1 std on the rebuilt tree. [Ok, so I hope that my corrections to the writing suggest other corrections as you reread the entire paper. More active voice. More thought in choice of superlatives. The writing is generally good, but it could still use improvement. As far as work is concerned. In order to avoid polluting the testing set, please use both your standard deviation approach and the approach of simply identifying the top 5 predictors on the full tree of 90% of the data and then testing it on the remaining 10%.]

Full tree, importance threshold = 0.0 std’s

Right Hippocampus volume/ICV +2.7 std’s

BDNF, a blood plasma biomarker: +2.7 std’s

Left Hippocampus volume/ICV +2.5 std’s

Full tree, importance threshold = 0.8 std’s

Right Hippocampus volume/ICV +1.4 std’s

BDNF +1.3 std’s

BMI (weight/(height^2)) +1.3 std’s

The effect of these predictors is complicated by the branching nature of the tree. It is not always clear whether a predictor correlates positively or negatively with the target. A simplistic idea is simply to correlate the predictors with the target to determine the relationship. This was done below, and it is shown that all predictors correlate negatively with the target. These results therefore show that there is greater likelihood of being considered a type ‘2’ transitional for patients whose measurements are lower (lower BMI, smaller hippocampus volume, etc.).

BMI -0.049

BDNF -0.076

Left Hippocampus/ICV -0.042

Right Hippocampus/ICV -0.009