Dear Dr. Kline,

[Brandon, Please use first person plural throughout. This is not quite a scientific paper but should feel that way.]

I transferred all of the data you gave us into MATLAB, where I've done the work up till now. I created a 447 by 45 matrix (rows represent patients, while a column represents a predictor and its value for each patient). This matrix includes BMI, ICV, age, gender, ApoE status (the binary you gave us) and everything from column H to column AS in the "Curated\_2" spreadsheet. These constitute our potential "predictors".

I also made a 447 by 1 vector that represents the 3-step ordinal diagnosis (NL, MCl, and MCl-AD). We tried the 2-step diagnosis (have Alz or not), but it did not prove as fruitful, because almost everyone has Alz and therefore just predicting “yes” was almost always correct..

Based on MATLAB's software, I then wrote a code for this data. The most important MATLAB program I used was the "Regression Tree Fit" program, which takes your predictors (biomarkers, age, gender, etc.) and the 'result' values associated with each patient (the 3-step diagnosis) and creates a decision tree from it. I can further explain precisely how the program works tomorrow, but briefly, the program is a form ensemble learning which takes the data given and creates a decision tree that minimizes mean-squared error. The decision tree was used to predict values (i.e. 1, 2 or 3 or fractional values in between which were then rounded) for data input. Finally, I tested how well the tree did by comparing the predictions with the true values; I will explain how I did this and the statistical tools I used in a few paragraphs.

We did three types of trees. The Full Tree took 100% of the data (i.e. used all 447 patients) and tested the tree on those same 447 patients. All results for the full trees were averaged over 10,000 trials to reassure us of our results.

Then, we had two types of Leave-Out Trees. One of these took out one patient at a time, created a tree from the remaining 446 patients, and tested that tree on the patient. This process was repeated until the end of the patients and, again, the predictions for each patient was compared with the true values for those patients.

Finally, the other leave-out tree was created on 90% of the data, chosen at random. The resulting tree was tested on the remaining 10%. This process was repeated 10,000 times to ensure robustness, and all statistics were averaged over all 10,000 trials.

The statistical measures we used are 1) precision and 2) recall. Just to recap, precision refers to the percentage of the model’s predictions which correctly identified a class. Recall measures the percentage of a class which the model correctly identified. I found these values for each class (1,2,3) and made a weighted average to find overall p&r. The weighting is based on the number of patients in each class.

In addition to numerical computations of precision and recall we show the confusion matrix, which refers to a 2x2 matrix consisting of 4 categories: true positives, true negatives, false positives and false negatives. I compared the trees' predictions to the actual values (i.e. if the prediction was in a given class, but the true value was outside that class, we had a false positive).

We also used Predictor Importance, which estimates the predicting power of a given predictor/variable (i.e. age, BMI) by adding up changes in mean squared error caused by nodes (i.e. branch splits) on a given predictor, and divides this value by the total number of nodes in the tree. Again, when I measured the importance of the predictors, I summed it up over all runs and averaged.

[Brandon, you should also identify importance by looking at the variables that are high up in the tree.]

The results are given below.

RESULTS

**Full Tree** (using all data to create tree and testing on same data)

Class 1:

True Positives = 40

True Negatives = 401

False Positives = 2

False Negatives = 4

Precision = 95.2%

Recall = 90.9%

Class 2:

True Positives = 260

True Negatives = 171

False Positives = 7

False Negatives = 9

Precision = 97.4%

Recall = 96.7%

Class 3:

True Positives = 131

True Negatives = 306

False Positives = 7

False Negatives = 3

Precision = 94.9%

Recall = 97.8%

WEIGHTED TRAINING SETPRECISION: 96.4%

WEIGHTED TRAINING SET RECALL: 96.4%

**Patient-wise Tree** (creating tree with 446 patients and testing on remaining patient, one patient at a time)

Class 1:

True Positives = 12

True Negatives = 379

False Positives = 24

False Negatives = 32

Precision = 33.3%

Recall = 27.3%

Class 2:

True Positives = 178

True Negatives = 75

False Positives = 103

False Negatives = 91

Precision = 63.4%

Recall = 66.2%

Class 3:

True Positives = 49

True Negatives = 232

False Positives = 81

False Negatives = 85

Precision = 37.7%

Recall = 36.6%

WEIGHTED OUT OF SAMPLE PRECISION: 52.7%

WEIGHTED OUT OF SAMPLE RECALL: 53.5%

**Leave Out Tree** (randomly choosing 90% of data to create tree, then testing tree on other 10% of data)

Class 1:

True Positives = 1.6 +/- 1.2

True Negatives = 38.1 +/- 2.4

False Positives = 2.4 +/- 1.6

False Negatives = 2.8 +/- 1.6

Precision = 40.2% +/- 28%

Recall = 36.6% +/- 25.6%

Class 2:

True Positives = 16.9 +/- 3.2

True Negatives = 8.2 +/- 2.5

False Positives = 9.7 +/- 2.8

False Negatives = 10.2 +/- 3.0

Precision = 63.5% +/- 9.1%

Recall = 62.4% +/- 9.9%

Class 3:

True Positives = 5.5 +/- 2.1

True Negatives = 22.7 +/- 3.3

False Positives = 8.8 +/- 2.8

False Negatives = 8.0 +/- 2.6

Precision = 38.7% +/- 12.9%

Recall = 41.0% +/- 13.9%

PRECISION: 53.8% +/- 7.4%

RECALL: 53.5% +/- 7.5%

We also ran another test to assure ourselves that our results were better than random guessing. This test involved assigning 1s, 2s and 3s at random to the patients (although maintaining the same numbers of each class) and checking how often this test did better than the leave-out predictions in terms of precision and recall. We found that it did better than our leave-out tree about 16-17% of the time, which is the p-value that the machine learning is doing no better than chance. This is reasonably good considering the preponderance of class 2 values.

Important predictors:

The importance of predictors was measured per predictor, averaged over all 10,000 trials for the Patientwise Tree. The predictors below are all the predictors with an importance value that is at least 1 standard deviation greater than the mean importance of all predictors (over all trials). Others did not make this cutoff.

+4.1 std's: Right Cortex divided by ICV

+3.1 std's: APOE.1

+1.6 std's: APOAII.1

+1.4 std's: LEPTIN.1

+1.3 std's: BMI

Finally, the decision tree for the full tree is shown in the PDF.