Classification

The problem:

We predict if a person will relapse in 6 months or not. Note that this is distinct from the problem of predicting whether or not someone will relapse at all, since some people relapse after 6 months. We chose six months in part because our data shows two groups of relapsers: those who relapse before 6 months and those who relapse after 6 months. Those who do not relapse are counted as not relapsing in 6 months, since they do not relapse at all.

This is also useful because 6 months gives an indication for a practioner if someone is likely to relapse soon, and so may require extra treatment now, or is likely to relapse far later or not at all, in which case the course of action required may be different.

Modeling approach:

Interpretability is important here. We wanted a model that was easily interpretable, as we're curious about the neuroscience behind this prediction problem.

The baseline to which to compare our model was the performance of a naive classifier that just classifies everyone as the most likely category. In our data set most people (25/39) do not relapse in 6 months. Therefore, if we predicted that no one would relapse in 6 months, our error rate (using 0-1 loss) would be $14/39 \approx 36\%$. In this case, every misclassified case would be a false negative, and so the false negative rate is also 36%.

For this problem, a false negative (i.e., predicting someone will not relapse in 6 months when they actually do) is worse than a false positive (i.e., predicting someone will relapse in 6 months when they actually don't). A false negative might lead to someone missing out on necessary immediate treatment. We decided that would be worse that someone getting immediate treatment he/she may not have needed. Therefore, we wanted to keep the number of false negatives low, even if it affected the overall estimated test error.

Best model:

Our best performing model was a logistic regression model with only one covariate: `Y` ~ `nacc_neutral_beta`.

When selecting a model, we only considered models with covariates that had no NA's. This is because we did not want to reduce our data set any further, as well as because we found that the features with NA's did not appear strongly correlated with relapse within 6 months. None of the fMRI covariates had any NA's, so we only ended up removing behavioral/demographic covariates from consideration.

To evaluate our models, we conducted 10-fold cross-validation and compared the cross-validated test error. Our objective was always 0-1 loss. We ran Lasso logistic regression because we wanted to reduce the number of covariates, as well as to see which covariates were important predictors of relapse within 6 months. We also conducted k-neartest neighbors with various values of k. We also fit logistic regression models with varying combinations of covariates (including the covariates selecting by the Lasso). However, none of our models had as low of a cross-validated test error a logistic regression model with `nacc_neutral_beta` as the sole covariate.

[insert hershel's paragraph about nacc and betas]

When using a threshold of .5, the cross-validated test error for this model is .210. The false negative rate is .145. This is therefore an improvement over the specified baseline, in terms of total misclassification rate and false negative rate. We used LOOVCV to compute this test error, after using 10-fold cross validation to compare across models. This is because when selecting a model we wanted to reduce variance, but didn't care much about bias. However, when estimating the test error we cared more about reducing bias. Our data set is also very small, and so LOOCV was feasible.

Our model is very simple. It therefore will have higher variance than a more complex model, and lower bias. This isn't necessarily ideal, since we don't really want a model that is sensitive to small changes in the

data. However, for our other models to have an estimated test error near the one of our simple model, we had to include both beta and signal coefficients, which sacrifices interpretability and is frowned upon in the neuroscience field.

Our model uses a threshold of .5–we classified those with predicted responses greater than .5 as relapsing in 6 months. We thought a threshold of .5 wouldn't be ideal in this case, because we were wanted to reduce false negative rate, even if it meant increasing total error rate slightly. However, changing the threshold enough such that the false negative rate went down a reasonable amount ment driving the total error rate up quite a bit. We decide the reduction in false negative rate was not worth the cost in total error, and so kept the threshold at .5