Jonathan Ramos
COMP 4441 Final Pape
Predicting Alzheimer's Disease Diagnoses

An Alzheimer's Disease (AD) diagnosis can only be confirmed through post-mortem analyses of neurofibrillary tangles and plaque deposits, and so it is necessary to identify alternative prospective biomarkers that can aid in the classification of patients as demented or non-demented. It has been known that changes in premorbid brain volume have been associated with AD neuropathology (Silbert et al 2003), and because brain and intracranial volume are easily assessed by MRI, whole brain volume is naturally a promising candidate biomarker (Kehoe et al 2014). In order to determine whether normalized whole brain volume (nWBV) estimated via MRI is a suitable biomarker for AD diagnosis, I will first assess differences in the development of nWBV in normally aging adults vs demented adults, and then lastly, I will evaluate how well linear discriminant analysis and random forests are able to classify patients as being either demented or nondemented based on these data.

MRI data was collected at the Knight Alzheimer Disease Research Center (ADRC) at Washington University as part of OASIS: Open Access Series of Imaging Studies (https://www.oasis-brains.org/) and for this project I will only be examining the longitudinal study, accessible via kaggle. In the longitudinal data set, normalized nWBV as well as other potentially useful metrics were measured in 150 participants aged 60-96. Participants were scanned 2 or more times with each visit separated by at least one year. Of the 150 participants, 72 were characterized as nondemented for the duration of the study, 64 were characterized as demented throughout the study and 14 were characterized as non-demented during the initial visit but were later characterized as demented by the end of the study. It is important to note that in this data set, the label "nondemented" means the participant maintained a clinical dementia rating (CDR) of 0 for the duration of the study, the label "demented" means the participant maintained a CDR of 0.5 of greater for the duration of the study, and the label "converted" means the participant entered the study with a CDR of 0, but ended the study with a CDR of 0.5 or higher (that is, a higher CDR indicates more severe dementia).

Ultimately, CDR is going to be the class that is to be predicted with our classification models, so the first portion of exploratory analyses focuses on identifying any relationships between CDR and other variables. From the correlation heat map and pair plot in Fig 1A and Fig 1B, CDR is negatively correlated with nWBV and mini mental state examination (MMSE). CDR is negatively correlated with MMSE because MMSE is the in-person examination that is a part of the CDR scoring. The negative correlation between nWBV and CDR, further explored in the violin plot visualizations (Fig 1D, panel 3), suggests that a higher cdr rating tends to be associated with a smaller nWBV, which supports the idea that nWBV decreases with AD progression. Lastly the correlation heatmap also suggests nWBV is also negatively correlated with age (Fig 1A, B). This suggests that there is a natural decrease in nWBV that occurs as part of the normal aging process.

Next, in order to further explore the relationship between CDR and nWBV, I used a one-way ANOVA to test whether there is a difference in mean nWBV between participants who have a CDR of either 0, 0.5 or 1. Note that in order to not overrepresent certain participants (some participants returned more than others during the study) data will only be taken from each participant's initial visit. From Fig 2 B, the p-value of the ANOVA is 0.00012 and so we may reject the null hypothesis and conclude that at least one of the means is different. From the post-hoc Welch's t-tests, all 3 p-values are less than 0.016667 (Bonferroni corrected p value for 3 comparisons) and so we may reject all three null hypotheses and conclude that the mean nWBV between participants whose CDR is 0, 0.5 or 1 are all different from each other (Fig 2B). The scatter plot and regression line in Fig 2A shows that nWBV is negatively correlated with Age (Pearson's r, p=1.18e-13) and so we may conclude that there is a relationship between nWBV and age; however, the spread of CDR across Ages (indicated by color) suggests that even though nWBV naturally decreases with age, it is not necessarily that case that CDR also increases with age. Taken together with the box plots from Figure 2B, this indicates that the differences in nWBV based on CDR are not just due to age, but suggests that they may instead be associated with disease progression.

Next, rather than just looking at nWBV at a single timepoint, I wanted to assess whether or not demented vs nondemented participants have a different nWBV rate of change, or atrophy rate. That is, do demented participants lose brain volume faster than non demented participants? Because this is a longitudinal dataset, there are multiple nWBV measurements of each patient over at least one year. To estimate the atrophy rate of each patient, I took the linear regression of nWBV over time for each patient, and then considered the slope of the line as an estimate for the rate of change of nWBV, which is taken to be the estimated atrophy rate. Fig 3 shows the nWBV measurements of a single representative participant over about 1200 days with the regression line plotted over it. After repeating this process for all 150 patients, I used a one-way ANOVA to determine if there is a difference in mean atrophy rate amongst any of the CDR groups. From the ANOVA in Fig 4B, p=0.02407 and so we reject the null hypothesis and conclude that at least one of the means is different. From the post hoc Welch's t-tests, the only difference in atrophy rate is found between Demented and Nondemented (Fig 4B Welch's t-test, p= 0.009001). This suggests that those who are labeled as Converted, which we may consider as being in the early stages of AD, fall somewhere in between. Lastly, the scatter plot and regression line of atrophy rate over age shows that atrophy rate is not related to age (Fig 4 A, Pearson's r, p=0.2419). Altogether, this means that we may conclude that demented participants have a higher atrophy rate than non demented participants (sign is negative to indicate loss of brain mass) and that this difference in atrophy rate is not due to age.

Lastly, because disease outcomes were determined by CDR, I trained two models, linear discriminant (LDA) and random forest (RF), to classify participants' CDR based on atrophy rate, MR delay (the number of days between each visit), sex, age, education, socio-economic status, MMSE, estimated total intracranial volume and nWBV. Note that since I'm using the most up to date atrophy rate for all participants, that is, atrophy rate was estimated using all available nWBV data points, I will only be training/testing these models on data from the participants' last visit. LDA is type of supervised classifier that works similarly to PCA but instead of maximizing variance, LDA maximizes the separability of two classes by maximizing the function $\max_{v:\|v\|=1} \frac{(\mu_1-\mu_2)^2}{{s_1}^2+{s_2}^2}$ where $\mu_j = \boldsymbol{v}^T \boldsymbol{m}_j, \quad j =$

$1, 2\ for\ class\ 1, class\ 2\ etc.$ In essence, it finds a new direction vector, v, that is some linear combination of the features where, once the data is projected on to v, the squared difference of the means between two classes is maximized and their variances are minimized resulting in minimal overlap. In random forest, rather than training a single decision tree, which can be highly sensitive to training data and might fail to generalize, we train many random trees. First, data is bootstrapped (rows are randomly selected with replacement to generate many new sets), then each tree is trained on a randomly selected subset of features. The final output is the majority vote of all the trees, a process also known as "bagging."

Fig 5 summarizes the LDA model trained to classify CDR into 0, 0.5, 1 or 2. The ordination plot for the trained LDA model in Fig 5A,B shows that despite LDA finding maximal separation, there is still a decent amount of overlap. Ideally, if LDA was very effective, there would be no overlap between classes. It is interesting to note that classes of CDR that are more different from each other tend to be farther apart than adjacent classes, i.e. CDR 0 and CDR 2 do not share overlap whereas CDR 0 and CDR 0.5 share much more overlap. From the confusion matrices, the training accuracy was 68.57% and the testing accuracy was only 64.44%. The confusion matrices also show that adjacent CDR classes tended to get the most confused, that is, it is more likely to confuse CDR 0 and CDR 0.5, but not CDR 0 and CDR 2 (Fig 5C). The RF model did not fair any better with a training accuracy of 54.29%, a testing accuracy of 64.44% and an out of bag (OOB) of about 40% (Fig 6A,C). Again, from the RF confusion matrices, the most mistakes were made between adjacent CDR classes. Features axes plotted on the ordination plot as well as the mean decrease GINI plots indicate that both models consider nWBV and atrophy rate to be important features but the RF model weighs MMSE and MR Delay more heavily than the LDA model (Fig 5A, Fig 6B).

Because both models were unbalanced and performed poorly, I thought to reframe the question. Rather than trying to classify subtle differences in the level of dementia, it might still be informative to classify whether a participant had any dementia at all or no dementia. This means that at the end of the

study, “converted” participants are considered as “demented”. Specifically, all CDR scores equal to 0 were considered nondemented (Class 2) and any non-zero CDR was set equal to 1 and considered demented (Class 1) where demented was treated as the positive class. This also made the groups more balanced with 72 nondemented and 78 demented. From Fig 7A, despite there still being some overlap, the density plot across the first (and only) LD component from the binary LDA model shows improved separability between classes than the ordination plot from the multiclass LDA model (Fig 5A). From the confusion matrices for the binary LDA model, if we consider demented as the positive class, sensitivity for training and testing was 60.00% and 74.07% respectively (Fig 7C). The random forest also improved; it had an estimated out of bag error rate of 26.67% with a training and testing sensitivity of 60.0% and 81.41% respectively (Fig 7 B,D). Overall, the RF model had slightly better sensitivity (or recall) across testing and the LDA model had both high specificity and high precision (30/33 = 90.90% for training, 20/20 = 100% for testing Fig 7C). In cases such diagnostics it is often better to prioritize recall or sensitivity because it can be better to be a bit overly cautious and have more false positives than to incorrectly categorize a diseased patient as not having the disease. This means if we were to strictly prefer higher sensitivity, we would prefer the RF model. One other approach to take into account both precision and recall however, is to calculate the F1 which is given by $\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2(Precision)(Recall)}{Precision+Recall}$. On the testing set the binary LDA model has an F1 of 0.851 and the binary RF model has an F1 of 0.815, and so we may consider the binary LDA model to be preferred.

Fig 8 shows that the LDA model also had a larger area under the receiver operating characteristics (AUROC) which is an indication that LDA is the preferred model. In the ROC plot, the diagonal represents predictions being made as if only by random chance and if a model performs better than random chance, we expect the curve to lie above the diagonal. The area under this curve is a summary of the overall sensitivity and specificity of a model at various thresholds and so a model whose AUROC approaches 1 is considered better; a perfect model that accurately classified both all true positives and all

true negatives has an AUROC of exactly 1. Given that the AUROC for the binary LDA model is 0.957 and the AUROC for the binary RF model is 0.893, LDA is the preferred model (Fig 8).

In conclusion, all three of the initial exploratory questions are addressed. First, mean nWBV does vary by CDR score: at the initial visit, mean nWBV in those who have a CDR of 0 is higher than the mean nWBV of those who have a CDR of 1. Second, after the final visit, those who were considered demented had a higher baseline atrophy rate than those who were considered nondemented; this was shown to be independent of age. Lastly, both LDA and RF performed poorly when trying to classify 4 different levels of dementia, but improved when under a binary model, only classifying whether or not a participant had any dementia at all. Overall, the binary LDA model had similar sensitivity to the binary RF model, however, the binary LDA model also had a higher F1 and a larger AUROC than the RF model, so the binary LDA is the preferred model. All together these studies show that nWBV is a promising biomarker that can be helpful in the diagnostic process of AD; however, it is important to keep in mind that AD is a complex disease with many other factors. Although we cannot fully rely on these models to predict an AD diagnosis, frequent MRIs can still provide some useful insights into AD progression.

References:

1. Kehoe EG, McNulty JP, Mullins PG, Bokde AL. Advances in MRI biomarkers for the diagnosis of Alzheimer's disease. Biomark Med. 2014;8(9):1151-69. doi: 10.2217/bmm.14.42. PMID: 25402585.

2. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci. 2007 Sep;19(9):1498-507. doi: 10.1162/jocn.2007.19.9.1498. PMID: 17714011.

3. Silbert LC, Quinn JF, Moore MM, Corbridge E, Ball MJ, Murdoch G, Sexton G, Kaye JA. Changes in premorbid brain volume predict Alzheimer's disease pathology. Neurology. 2003 Aug 26;61(4):487-92. doi: 10.1212/01.wnl.0000079053.77227.14. PMID: 12939422.