ADNI Meeting Notes

10/3/2024

* Variable importance – finish this. Get preliminary data.
  + Marshall made a separate module for doing this analysis. Can look at different variables and see what gives us information.
  + Training random forest models on pieces of the data. Looking at some of the datasets that would be helpful for the prediction. Random forest models tend to be pretty resilient to noise.
* Cluster analysis to try to regroup things. Put CN and SMC, EMCI and LMCI, and AD.
  + Use the mental battery.
  + Use a model to figure out which data is useful, and then use that for clustering.
  + Compare how close the clustering is to the already assigned groups.
* Look at all of the ways that they list the MPRAGE scans, are they the same?

9/19/2024

* Meet with Marshall to learn how to do Random Forest stuff with the Neuropathology Data. (Meeting on Tuesday 9/24)
  + CRAN: randomForest package
  + Need to fit the randomForest model
  + Check out the randomForest CRAN documentation.
* Need to look into how to do Machine Learning with Time Series model stuff.
* Questions to Address:
  + More information about Accelerated Sagittal. How compatible are Accelerated Sagittal and MP-RAGE?
  + Does the Tesla difference matter for structural scans?
  + What PET scans would be useful? How can we analyze this?
  + Size variability? Scanner variability and replacement?
  + Look for a recently published ADNI MRI paper and see how they are handling these issues.
  + Regions that are affected by Alzheimer’s.
  + Blood Biomarker info from Dallan.
  + ADNI sets used for Training and Validation purposes?

9/5/2024

* Spencer is working on getting more of the data. Worse results with the more recent data, may be between ADNI 1 and ADNI 2.
  + Accelerated Sagittal? What is it? Is it different from MP-RAGE?
    - Accelerated sagittal is a general method used to decrease the time it takes to get a T1-structural scan (there are multiple methods of doing this).
    - ADNI-GO and ADNI 2 use the GRAPPA method (SENSE from another manufacturer) of doing an Accelerated Sagittal scan. (Jack et al., 2010). ADNI 1used a typical MPRAGE sequence (Jack et al., 2008).
    - ADNI 2 should have both MPRAGE scans and Accelerated Sagittal scans, which in theory should be decently similar and, with both, should provide a way to see if we can train the model to be able to interpret both types of scans (Falkovskiy et al., 2016). This will be important as ADNI 3 and ADNI 4 started using accelerated sagittal scans instead of MPRAGE scans.
      * The papers cited here are in the ADNI Box folder Research Papers.
  + Are there changes in the scanning parameters between the different cohorts? Did they use different Tesla? Lots of variability in the sizes of the images going beyond ADNI 1.
    - There are some changes in the scanning parameters. One of the most concerning is that sometimes scanners were replaced, at which point ADNI suggests that the longitudinal data is not compatible between the two scanners (<https://adni.loni.usc.edu/data-samples/adni-data/neuroimaging/mri/>)
    - ADNI 1 seemed to use 1.5T scanners and 3T scanners, but everything beyond ADNI-GO used a 3T scanner (<https://adni.loni.usc.edu/data-samples/adni-data/neuroimaging/mri/>)
    - I don’t yet know why there is such variability in the sizes of the images. I may need to meet with Spencer to figure that out.
  + Look for a recently published ADNI MRI paper and see how they are handling these issues.
  + Are there any ADNI sets that are used for training or validation purposes?
    - One study was doing research with participants not in ADNI and used a small subset of ADNI individuals as a validation set (Zhu et al., 2022).
  + Regions that are affected by Alzheimer’s.
  + Blood Biomarkers – anything with Tau (either in blood or CSF) would be correlated. Amyloid is interesting, but doesn’t correlate with cognitive function super well. (Dallin – Box – Biomarkerdata.excel) Insulin? How much data is there? How often was this data collected?
    - Emailed Dallan.
* Morgan – Finding data to fit to a Random Forest Model
  + Use the UPENNBIOMK\_MASTER (ADNI -> Labeled Data Files -> UPENNBIOMK\_MASTER)
  + Take Diagnosis in Column of Labels (DX SUM), then look at training a random forest model. Will need to look at variable importance.
  + Dr. Rhodes will send me instructions/starter code to get started.
* Jake – Next Steps
  + How can we take the existing methods, and new methods, and apply it to data in a way that is publishable? Whether it’s focused on methodology or the AD problem, what do we need to do to get this multi-dimensional domain problem into an embedding something that is useful?

8/21/2024

* No meeting at this time next week.
* Jake is sending out a WhenIsGood to get the new meeting time for the semester.

8/14/2024

* Marshall:
  + Looking at the models that Marshall ran.
  + We might need some sort of metric to measure how healthy the individual is, some sort of status, to track across their disease progression. A large part of visit data is scores from an AD test, which they took very consistently across the board.
  + We need to get a trend, because the patient is going to get worse. What if we used the overall ADS score as an indicator of time lag? This may help us to better find the trend to predict progression and where someone is in the disease progression. The more variables that we have, the more data points we need and the more helpful data we need to learn. Have been including the different subscores in the test, but there isn’t a reliable scale within these subscores (the range varies dramatically).
  + Use total 13 as a way to help predict the progression.
  + DTW may outperform Euclidian distance, so haven’t given up on DTW yet, but would take lots of recoding.
* Spencer:
  + Able to rerun something and it turned out much better. Some things separated and some things didn’t.
  + We can start to look at which features and portions seem to be more relevant.
  + Examining the slices to figure out where they are and looking for patterns between early and late ones. These are still the MPRAGE. The orientation should all be the same.
  + Need to look at the VisCodes and how those relate to the MRI visits.

8/7/2024

* Questions for Dawson
  + How can we take the Q#Score data and make a comparison out of it? Can we extrapolate the data and fill in the missing variables between visit?
  + How does the disease progress? Exponentially? Curve? Log?

7/31/2024

* Marshall
  + Has graphics for some of the tabular data.
  + Been working on getting the manifold alignment running on his computer. So now we can run stuff for ourselves on this project. Working with Adam – (18 patients included so far) Generated a similarity matrix, graph with different data points linked to different domains, manifold alignment with Timeless and Visit Domains (this generated a heat map of similarities of the patients that were included), representing webs
  + Jake – manifold alignment is looking at features that have some sort of correlation that we may not fully understand. Cross-embedding classification accuracy – if you take all of the points from the data domains and try to predict the labels from the points of the other domains, can I take the low dimensional representations of those points and predict what the other labels should be. Other metric – given the ground truth, given that we know we have the same person represented in two domains, which points are closer than the actual true representation. The closer it is to zero, the better the alignment. One problem we are facing is that we have a lot of missing data, so given that we have this missing data, can we predict the labels of observation from a different set of features. Eventually we want to be looking at the diagnosis. It’s not something that you do in practice, but just a check to see how well the alignment model is actually doing. Eventually, we will take the images and align them with the timeless data and the longitudinal data in like a three domain problem (eventually), and pairwise for now. Heat map – two squares represent distance between patients within a given domain. Other rows represent distance between profile data and longitudinal data.

7/17/2024

* Morgan:
  + \*\*Send email about next week’s meeting.
* Dallin:
  + Probabilities of retaining or types of learning (essentially different types of memory). This is in a CSV for Marshall.
  + Biomarkers – UPEN Plasma. Master file (gives for 6000 RIDs) - raw amyloid beta 1-40 & 1-42 (ratio or raw values? Master file just gives beta for 1-42), z scores, tau-raw, ptau, tau181. This was CSF. Adding to the list of things to add.
  + Meeting with Marshall to integrate data.
  + Going to continue going through the UPEN stuff. Masterfile is huge, with a baseline and a 12-month follow-up.
* Marshall:
  + Met up with Adam (Dr. Rhodes’s grad student). Took the two tables and ran through the manifold alignment. Data isn’t super useful yet. Ran data with Timeless variables and visit variables. Had the profile/timeless variables in one domain and the visit variables in the other domain, instead of including images at this stage. Not complete yet, but was a good way to test the manifold alignment to see if it works.
  + Going to keep working on a way to run it and have it spit out results in a way that makes sense.
  + We think the paper is about the method.
  + Trying to learn how to run the manifold and understand the manifold. Pretty close to being able to run the manifold alignment.
* Dawson:
  + Will be important to get an initial paper out. Forces us to be organized.

7/10/2024

* Jake not here at this time next week.
* Spencer:
  + Preliminary 2D results. Froze network and got 80%-90%.
  + Balance between the classes: seems like an even distribution, but Spencer is going to look into it.
  + Working on the 3D CNN. Takes a long time to run and took a while to get it to work. 55% accuracy. Small data and only ran for 2 epochs.
    - Look at increasing the batch size.
    - Issue with the batch size. Need to look into why it won’t let you use the batch size. Batch size shouldn’t be tied to the architecture whatsoever.
  + Going to meet and see if they can fix it.
  + Wanting to see if we get better results from the 3D CNN. Still working with the 2D images.
  + GitHub – Medical Net. Using Resnet 50 3D.
    - Maybe use the Resnet 10 to see if we get similar results and is faster.
    - Resnet 50 is a pretty consistent baseline though.
  + Look at num\_worker to see if that changes how it runs.
  + Work on extracting the last linear layer and apply some dimensionality reduction to it.
  + Once we run it:
    - Are there clusters based on the class level? Are there position of slices that tend to group together?
    - Do the groupings for an individual patient end up close together or are they more based on location of the brain?
    - SKLearn – t-SNE, PHATE, ISOMAP, MDS – these are all just versions of dimensionality reduction tools.
      * OP = MDS()
      * Op.fit\_transform(X)
      * 2048 channels – that many features by 4000 observations.
      * This will produce a 2D array that you can put into a scatter plot. Hopefully it has some sort of pattern with different groupings. Probably see multiple groupings in the same class scattered throughout.
    - We can take two different embeddings and run different manifold alignment algorithm on it to see which points are linked and if there is anything we can do with it.
* Marshall:
  + Timeless variables, visit variable, event variables (random times in the timeline, like head trauma).
  + Add variable to the analysis

6/17/2024 (Monday meeting with Wednesday being a holiday)

* Marshall:
  + Met with Adam, lots of very technical parts of the code.
    - Dr. Rhodes: He has his own naming conventions.
* Dr. Rhodes:
  + Time series data. Manifold alignment needs two different data domains. Want to take data with time components and treat as own domain. Everything static treat as domain. Neuropathology as a third domain (that way we don’t have missing variables throughout the data).
  + How many anchors do you need to have to make the manifold work? The more the better, but no numerical threshold. Anytime there is an RID value, that will help us provide a link and an anchor. Datapoints that don’t show up for everyone can be a different domain with anchors throughout there.
  + Is it possible to anchor the same entry in one dataset to multiple datasets in another? Yes.
  + Will need to determine distance between observations. Usually in Euclidian distances, but we will probably need to consider Dynamic Time Warping (DTW). This will allow us to find the difference to generate similarities and will better account for the time. Will kind of account for the starting points being different. Will be important to determine how to measure similarity between patients.
* Morgan:
  + Run a regression model.
    - Response variable(s): NIA-AA Alzheimer’s disease neuropathologic change; Whole Brain Weight
  + Last week info:
    - Neuropathology potentially interesting variables:
      * NPWBRWT [L] a. Whole brain weight (if half brain, multiply weight by two) (9999=unknown)
        + We might also want NPWBRF, which identifies if the weight above is given from a Fresh or Fixed specimen.
      * Some of the atrophy variables: Cerebral Cortex atrophy (NPGRCCA) and Hippocampus atrophy (NPGRHA)
      * NPTAN [T] a. Tau antibody (Check One)
      * NPABAN [V] b. Amyloid beta antibody (Check One)
      * NPTDPAN [Z] d. TDP-43 antibody (Check One)
      * NPTHAL [AH] a. Thal phase for amyloid plaques by immunohisto-chemistry (IHC) (A score – Check One)
      * NPBRAAK [AI] b. Braak stage for neurofibrillary degeneration (B score – Check One)
      * NPNEUR [AJ] c. Cerad score for density of neocortical neuritic plaque (plaques with argrophilic dystrophic neurites, with or without dense amyloid cores). Score without respect to age or diagnosis. (C score – Check One)
      * NPADNC [AK] d. NIA-AA Alzheimer’s disease neuropathologic change (ADNC) (Check One)
      * NPDIFF [AL] e1. CERAD semi-quantitative score for diffuse plaques (plaques with non-compact amyloid and no apparent dystrophic neurites). Score from the neocortical field with the highest plaque density and without respect to age or diagnosis. (Check One)
      * NPAMY [AM] e2. Cerebral amyloid angiopathy (Check One)
      * NPARTER [BS] e. Arteriolosclerosis? (Check One) (Assess in subcortical white or gray matter)
      * ??NPPATH [BU] g. Other pathologic changes related to ischemic or vascular disease not previously specified?
        + If this variable is included, questions 12g1-12g12 would probably need to be analyzed as well.
      * --While NPPDXP [DY] p. AD-related genes (dominantly inherited); do not include APOE or other polymorphisms or genetic risk factors. Would be interesting, all of the collections indicate Missing or Not Collected.
      * --Similarly, the autopsy data would be interesting, but only one had a full autopsy performed at this stage.

6/12/2024

* Morgan:
  + Last week info:
    - They do have blood and CSF info.
    - Potential spreadsheets of interest:
      * Biomarker Samples [ADNI1, GO, 2, 3, 4]
      * CruchagaLab CSF metabolic Feature Info (and corresponding data spreadsheets)
      * ADMC UCSD Untargeted Metabolomics with LC MSMS Baseline Methods [ADNIGO/2] (and corresponding plates)
    - In the Neuropathology Report, there is a lot of information about the condition of the blood system in the brain. As well as other things that were causing disease or caused death.
  + Dawson: This stuff is complicated, so we would have to figure out how to aggregate it and what it means. Could be a separate analysis to make heads or tails out of this information. We’ll work on figuring out which obvious ones to include.
  + Jake: Selection: look at existing feature selection methods, like what you would do in a regression. Then, go with different selections of more or less important variables, and then doing all of the combinations.
  + \*Make a list of the variables in the Neuropathology report that would be useful. Put them in the Interesting Timeless Variables spreadsheet?
* Marshall:
  + Setting up meeting with Adam, Dallin, Spencer, and Jake(?) to walk through the manifold alignment. Help explain the structure of the code and figure out what needs to be fed into the code.
  + Putting together the not time-bound variables. Looking through the files that we have and where it is.
  + Going to keep going through the subjects that are giving problems for the code and making sure the data is in good shape.
  + How many domains can the manifold alignment do?
    - Jake: Right now it is set-up to handle two at a time, but it will be able to handle as many as we want. Adam is still working on that.
  + Family history?
    - Jake: Condense the data into some sort of aggregate score would probably be easier than trying to create a new object of a family tree. Maybe start with focusing on the parents. The things below will allow us to aggregate to a single aggregate score.
    - Dallin: Parent data: alive, current age, dementia or AD?, age at diagnosis. Keeps things simple by just focusing on the parents because those numbers are consistent across the different patients.
    - Will be important to capture AD and dementia (lots of misdiagnosis).
  + Visit Codes: Each cohort has a code for each visit of theirs data is being encoded for. Based on this, the data can be grouped together by visit code. We can then use that in the time series to match up people’s time frames.
* Dallin:
  + PTID – first three numbers correspond to the site, but this is a different number than the site ID. This is consistent. Last 4 numbers are the patient’s RID.
  + Medication history – can't find the what the medication is. Lots of info, but not the medications. Does have the reason it was prescribed. ADNI1 didn’t record the medication info. It started with ADNI GO.
    - Dawson: We could grab the quantity of medications, but that would still be very rough. The reason it was prescribed would be helpful, even if we can’t find the name of the medication.
    - Jake: Could group it based on the medication ID. Then, we would know the impact of that one medication and how many people are taking it, but we wouldn’t know what that medication is.
    - Marshall: Potentially need to aggregate the information.
* Spencer:
  + Two problems:
    - Primary – neural network 3D CNN isn’t working super well. Maybe go back to flattening the matrices, but that sounds like not a good idea and you lose a lot of the spatial dimensional data. Using other networks that are just 3D based. Nothing super obvious, but could be worth looking into. Paper with CNNs for medical imaging could provide solutions.
      * Jake: All sound like promising options. Maybe add another network that is just focused on the third dimension and would totally untrained and starting from scratch. This might be worse than starting with 2D images. Might be easiest initially to just look at the 2D images and see if anything useful is coming out of that.
      * There are so many more 2D images though, like 12,000x170. We need to start with a subset and see if it is capable of learning from these images. This would be to test the retraining.
      * Try the 2D first because it will be easier to set up. Then, if this doesn’t work, go back to looking at the 3D.

6/5/2024

* Adam joined today (instrumental with the manifold alignment).
* Spencer:
  + Lots of very technical neural network with imaging stuff. Flattening data from 3D to 2D. Looking at Pytorch to do it. Maybe being able to do an aggregation function. There is a premade way to do this in Tensorflow, so Pytorch should be able to do it.
  + Trying it with the 2D images first, to see how it works. Then going to try it in the 3D to 2D.
  + Tried it with a generic car image, and that seemed to work. Passed car through and got image net classes, output was sports car (got it).
  + Going to write a training loop. Have a separate training and validation step. See what kind of prediction you can get on the validation set for AD, MCI, etc. Start with the ResNet 50.
  + Look at index of same individual with the number of scan and/or the time between the scans. Then, we can use the vector embeddings and see if there are any sorts of trends or clustering.
  + Any other labels that we want to include? Don’t know right now and won’t need to figure that out at this stage. Mostly, this will be for more exploratory purposes. We do want to look at the DICOM files and see if there is anything there that we need to note.
* Jake
  + Going to look into the 3D to 2D convolution.
  + Being able to link the different types of data based on time-point will be important and useful for running the alignment.
  + Since a lot of the data is tabular data, that’s useful and we should be able to include a lot of it together or one at a time. We have lots of options.
  + Will be important to link tests and demographic information with the different time points.
  + We can do a clustering on the timestamps? Neural network? Should be within about a month, which shouldn’t be too much of an issue.
* Morgan:
  + What blood biomarkers do they have?
  + What blood biomarkers do we want to use?
  + What codes are they using for the medical history?
  + What inflammation markers do we have?
  + Do they have CSF info?
  + What amyloid-beta things do we have? Tau?
* Marshall
  + Python script to label datasets to browse through them. Running tables through the program to identify useful variables. Script that merges the data and applies the protocols. Not time sensitive variable analysis (genetic and demographic data would be good for this), but there is a lot of data that is timescale based.
  + Going to demo next week the tools that are being made. Looking at a lot of timescale stuff.
  + Going to look at the demographic information and link them with the assessments. Using the raw date and time for each item (biomarker test, mental exam, etc.) will be a good place to start. Then we can figure out from there how to group it. Create a dataframe where the information is there and tagged with a date. Later we can figure out how to aggregate those timescales in a more meaningful way.
  + Going to look for all time-sensitive variables and we’ll pair those with the images.

5/29/2024

* Marshall:
  + Manifold alignment explanation.
  + Talked to Jake yesterday. Want to start with the genetic data because genetic data doesn’t change over time. Then, trying to use this as a simple way to start training the model. Could be a control comparison.
  + Adam (who helped Jake build the manifold alignment) is willing to come and do stuff with this part of the project.
  + Jake will be around for the whole month of June.
* Dawson:
  + We need to decide which variables we want to use and compile them.
  + Polygenic Hazard Scores will be important and the APOE alleles (and we want all of the genetic composition, not just heterozygous or homozygous and for all three alleles).
  + Other data to think about: we are going to need to figure out all of the medical data (especially medical history) and the demographic data. How is it grouped? What do we have? We won’t need this just yet, but it definitely will be important. How will we be able to maximize the use of that data?
    - This may get to be the next thing we add because we don’t have to worry about the time course measurements. These values are stable and unchanging.
* Dallin:
  + Going to take a closer look at the variables to make sure we know what we have so we can choose.
  + Having a hard time knowing where to put effort.
  + Also going to start looking at the medical and demographic data.
* Spencer:
  + Working with the images. We are really close to being able to feed the images into the neural network. Some of it may be able to be fed into the manifold alignment and there are other things that we can do with it.

5/15/2024 (Lance Erikson joined this meeting)

* Marshall:
  + Data management details.
  + Data interface to changing the excel files to represent the variables that we don’t completely understand. Functional as of now. Going to continue to update the information.
  + Talking through the dataset and how we can manipulate it to be useful. Working on communicating understanding for what the data dictionary (codebook) is able to communicate.
  + Suggestion –
    - We should pick an analysis and try to figure out something to run as a preliminary analysis.
* Catching up Dr. Erikson.
  + Need to get him a login with ADNI.
  + Need to get him on Box.
  + Need to get him in the email chains and included in meetings and discussions.
  + Suggestion –
    - We need to figure out which datasets are applicable and which ones we should use.
    - We need to make sure that we are not merging all of the data. It’s not effective. Ex. pulling all the data for different parasites. We run the risk of losing significance.
    - We need to have Dr. Rhodes’ input on organization, especially for the purposes of verification. How we are running the analyses and which files are running are important for being able to replicate the data. Dr. Rhodes wanted to make sure that what we are doing is usable to our colleagues.

5/8/2024

* Marshall:
  + Background research on the biomarkers. Organizing datasets with biomarkers
  + Downloaded and scraped the data dictionary. Catalog is pretty good, but not perfect. Working on a script to take study data files with encoded variables and edits files with the properties of each variable underneath. Thus, any file put in the files of interest folder will have the tags and meanings in the data.
  + Come up with a more generic python script that would do the third party data base things listed under clay.
* Clay:
  + Looking into third party data base providers that will combine the datasets that we need. Would be rebuilding our own database outside of ADNI.
* Dallin:
  + Background research on the APoE and the polygenic hazard score. PHS run through the Cox proportional hazard model. Lower score is better for you and will take longer for the disease to develop. Combined genetic data with incidence rate. SNPs primarily used. About 1854 SNPs that are significant. Small effect size assumed, but not sure. They use a really low significance threshold (p < 10e-5) to do false discovery rates and make sure that we aren’t getting things with a false positive or false negative with a Type I error.
  + Couldn’t find data about parental imprinting.
  + Started looking through demographic data. We really want specific occupation, race/ethnicity, any demographic data that we can get a hold of.
* Spencer:
  + Uploaded the DICOM files to Box.
* Morgan:
  + Going to run the DICOM files Spencer uploaded through the dicom2nifti code and generate NIFTI files.
* Jake will be gone next week on Wednesday, so we may need to move the meeting to earlier in the week. Dawson can only make Wednesday. We’ll go ahead and meet Tuesday, then have a brief meeting without Jake on Wednesday to fill Dawson in.

5/1/2024

* Spencer:
  + Imaging
  + Looked through the database, figured out setup, suggested to work with MP-RAGE scans.
  + When you read in the scan into Python, it essentially becomes a pixel matrix.
  + \*\*\*NifTI files are 3D. Convert the DICOM to NifTI and see what they can do with Python.
  + \*\*\*Not using MP-RAGE after ADNI 1? If so, what are they using instead?
    - ADNIGO – MPRAGE
    - ADNI2 – MPRAGE & MPRAGE SENSE 2
    - ADNI3 – Accelerated Sagittal MPRAGE
    - ADNI4 –Accelerated Sagittal MPRAGE (MSV21)
      * Not every ADNI4 individual has this scan, but many of them do.
  + Lots of questions about where to go from here.
  + BIDS formatting for file organization when we get there?
* Dallin:
  + Genetics
  + Looking at how they generate the polygenic hazard score.
    - (Stepwise Cox proportional regression model) Cox regression model based on previous patients. Probably did a genome-wide association study looking at variables without outcome and effect of each variable. Each individual patients with variants and them summed the scores of those variants.
    - Dr. Hedges has a paper on how to calculate the PHS.
  + Will look into the APoE scores.
  + Drop files into the ADNI Box Folder along with questions.
* Marshall:
  + File organization & Biomarkers
  + Ask Sophia about a file that she was going to send.
  + Looked into what we might want to consider for the biomarkers.
    - We have some Tau measures that would be good. Dataset has hyperphosphorylated tau 181. 217 may be strongly linked, but it doesn’t look like we have this data.
  + Regular Expressions? Data frame? – Trying to manage the data dictionary.
  + Maybe input primary files of interest.
* Morgan:
  + \*\*\*Does APoE have parental imprinting?
  + \*\*\*Does ADNI report the post-mortem data? When is AD diagnosed?