Capstone Project: ADNI Data

Brian Collica, Ben Searchinger, Ryan Roggenkemper, James Koo

2/16/2021

Data

Access and Acquisition

The main source of our data will be DICOM format files of PET and MRI brain scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. For the purposes of this assignment, the data is coming from the Australian Imaging, Biomarkers, and Lifestyle (AIBL) database while we wait for final access approval from ADNI. Both databases are provided courtesy of the Laboratory of Neuro Imaging at the University of Southern California. It is unclear at this moment whether or not AIBL data will play a part in the final analysis, although this is a possibility we are exploring.

Exploratory Data Analysis

Downloading and Reading DICOM files

Much of the data is in DICOM format, a standardized file and metadata formatting system used in medical imaging. Other formats are also available such as NiFTI and ANALYZE, although we plan to restrict ourselves to the DICOM format.

There are many existing open-source software libraries in R and Python for reading and processing DICOM files. One such R package is `radtools` which is available on the Neuroconductor repository and also on GitHub. We have been able to successfully download PET brain scan images from AIBL and process them into R using the `radtools` package.

Below is an example work flow where we read in the entire 3D image from a directory containing 90 separate DICOM files, one for each image slice. Using the functions in `radtools`, we are able to inspect the dimensions of the data along with the metadata attributes. We can also transform the slices into a three dimensional array and inspect the results. `radtools` also facilitates viewing the actual image slices.

```r
# Install Package
source("https://neuroconductor.org/neurocLite.R")
neuro_install('radtools', release = "stable", release_repo = "github")
```

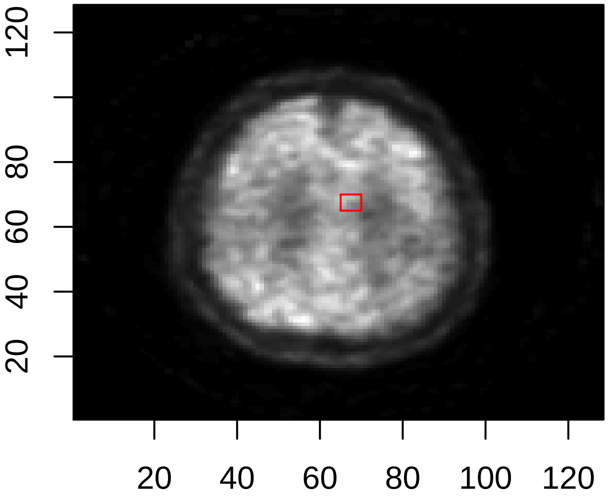
```r
library(radtools)

# Path to image directory
img_path <- "AIBL/10/summed.img__RSRCH_RAMLA3D-SUV/2006-10-17_13_53_08.0/I153055/"

# Read in 90 slices of PET image
PET <- read_dicom(img_path)

# Inspect image dimensions and number of slices
img_dimensions(PET)
num_slices(PET)
```

```
## [1] 128 128  90
## [1] 90
```

```
img_array <- img_data_to_mat(PET)
dim(img_array)
img_array[65:70, 65:70, 65]
```

```
## [1] 128 128  90
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]  535  569  625  670  668  612
## [2,]  517  540  594  635  646  616
## [3,]  486  494  495  547  634  613
## [4,]  459  435  432  531  630  609
## [5,]  423  404  408  506  571  554
## [6,]  359  373  380  434  488  494
```

This single 6 x 6 matrix corresponds to the region outlined below in the image representing slice 65.



The DICOM files' rich metadata is contained in the header. This conveys information about each scan such as the study date, manufacturer of the scanning device, slice thickness, and also patient-specific information like age and sex. We can transform this information into a data frame where each row corresponds to a different attribute. The first four columns contain information about each attribute, and the remaining 90 columns hold the values of each attribute for each of the 90 image slices.

```
dicom_head <- dicom_header_as_matrix(PET)
dicom_head <- as_tibble(dicom_head)
dicom_head[1:10, 1:4]
```

```
## # A tibble: 10 x 4
##    group element name                        code
##    <chr> <chr>   <chr>                       <chr>
##  1 0002  0000    GroupLength                 UL
##  2 0002  0001    FileMetaInformationVersion  OB
##  3 0002  0002    MediaStorageSOPClassUID     UI
##  4 0002  0003    MediaStorageSOPInstanceUID  UI
##  5 0002  0010    TransferSyntaxUID           UI
##  6 0002  0012    ImplementationClassUID      UI
##  7 0008  0000    GroupLength                 UL
##  8 0008  0008    ImageType                   CS
##  9 0008  0016    SOPClassUID                 UI
## 10 0008  0018    SOPInstanceUID              UI
```

```
dicom_head[1:10, 5:7]
```

```
## # A tibble: 10 x 3
##    slice_1                     slice_2                     slice_3
##    <chr>                       <chr>                       <chr>
##  1 "182"                       "182"                       "182"
##  2 "\u0001"                    "\u0001"                    "\u0001"
##  3 "1.2.840.10008.5.1.4.1.1~ "1.2.840.10008.5.1.4.1.1~ "1.2.840.10008.5.1.4.1.1~
##  4 "2.16.124.113543.6006.99~ "2.16.124.113543.6006.99~ "2.16.124.113543.6006.99~
##  5 "1.2.840.10008.1.2.1"       "1.2.840.10008.1.2.1"       "1.2.840.10008.1.2.1"
##  6 "2.16.124.113543.6006.99~ "2.16.124.113543.6006.99~ "2.16.124.113543.6006.99~
##  7 "400"                       "400"                       "400"
##  8 "ORIGINAL PRIMARY"          "ORIGINAL PRIMARY"          "ORIGINAL PRIMARY"
##  9 "1.2.840.10008.5.1.4.1.1~ "1.2.840.10008.5.1.4.1.1~ "1.2.840.10008.5.1.4.1.1~
## 10 "2.16.124.113543.6006.99~ "2.16.124.113543.6006.99~ "2.16.124.113543.6006.99~
```

Some of this metadata is specific to the type of study being conducted, so we expect the headers to be different for the ADNI image collections. In particular, we are interested in finding out if any other information, such as cognitive test scores, is encoded into the DICOM headers for any of the ADNI images available. This will be something to look into once we finally get access to the actual ADNI image data.

Project Outline and Analysis Plan

State of Current Research

Many papers have been published addressing the challenges associated with Alzheimer's research and using neuroimaging data. There is an overall lack of consensus regarding standardization of metrics and also which data and attributes are of importance. Additionally, there is much heterogeneity in the analysis approach employed by researchers with everything ranging from multiple linear regression models to deep learning with 3D convolutional neural networks.

Below are a few publications of interest:

- *Tau pathology in cognitively normal older adults*
  - https://doi.org/10.1016/j.dadm.2019.07.007
- *Robust automated computational approach for classifying frontotemporal neurodegeneration: Multimodal/multicenter neuroimaging*
  - https://doi.org/10.1016/j.dadm.2019.06.002
- *Practical algorithms for amyloid β probability in subjective or mild cognitive impairment*
  - https://doi.org/10.1016/j.dadm.2019.09.001
- *Added value of amyloid PET in individualized risk predictions for MCI patients*
  - https://doi.org/10.1016/j.dadm.2019.04.011
- *Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects*
  - http://dx.doi.org/10.1016/j.neuroimage.2014.10.002
- *Deep learning detection of informative features in tau PET for Alzheimer's disease classification*
  - https://doi.org/10.1186/s12859-020-03848-0

Question of Interest

The specific question of interest being addressed in this project is: can we combine the approaches and metrics from several recent studies into a general, multi-modal framework for identifying and tracking Alzheimer's Disease?

Analysis Approach

We plan to approach this question by first identifying a few main studies of interest. Then we plan to compare the various analytic approaches, including how each study processed their images, in order to come up with a method that can generalize across many platforms, image types, and patient characteristics.

Since this is a multi-modal approach, we hope to utilize various statistical methods where appropriate including, but not limited to, linear and logistic regression, support vector machine classifiers, and convolutional neural networks. Given the high dimensional nature of the data, we also plan to utilize dimension reduction techniques such as regularization or principal components analysis in order to facilitate noise reduction and aide in feature selection.

Breaking the overall analysis into different steps means that we can utilize different software for various tasks where appropriate. For example, one part of the model can be fit in R while another part can be fit in Python. We also plan to leverage parallel processing when fitting and validating our model. To facilitate this, we're seeking access to the campus Savio computing cluster and have also reached out to AWS for access to free research computing credits.

Project Roadmap

The project can be broken up into the following main steps:

Step Number	Due Date	Description
1	2/16	Preliminary Plan, EDA, and Project Description
2	2/23	Define the exact data to be used for analysis and have it downloaded
3	3/2	Have pipelines built for pre-processing images in R or Python with FreeSurfer & EDA Presentation Due
4	3/16	Have all images processed and ready to be fit into various models
5	3/30	Evaluate fit of preliminary model presentation due
6	4/13	Finish and evaluate secondary model fit
7	4/27	Have final results ready to go and finish details of presentation
8	5/4	Final presentations due

While each step has a specified target completion date, this is flexible depending on the unknown complications we will encounter, such as still not having access to the ADNI data despite numerous attempts to reach out to the appropriate parties in charge.