Investigating Links between the Immune System and the Brain from medical claims and lab tests

Alex Chu, Tymor Hamamsy, Guhan Venkataraman
Stanford University, CS 229 Machine Learning Term Project, Andrew Ng & Dan Boneh

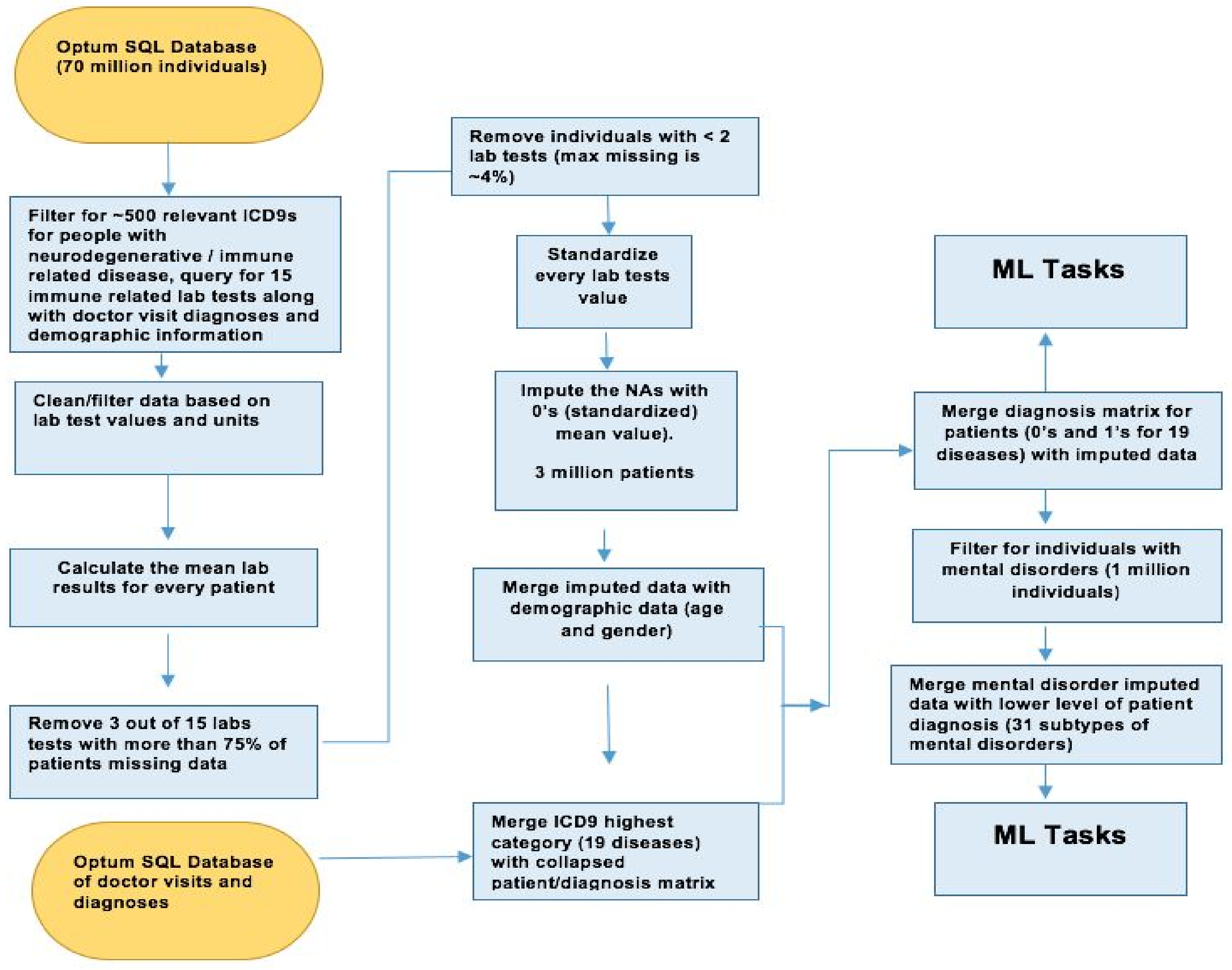
Overview

In this project, we used machine-learning to query and analyze electronic health records and medical claims data for connections between the immune system and the brain. Until very recently, the blood-brain barrier was thought to completely separate the brain from the rest of the body's immune system[1]. The first anatomist conjectured that there were lymphatic vessels going to the brain over 200 years ago[2]. However, it was not until this year that these lymphatic vessels in the brain were captured by an NIH team using advanced MRI technology[3]. In Genome Wide Association Studies (GWAS), many of the variants found to be associated with brain disease (i.e. Schizophrenia, Bipolar disorder, Autism) are also autoimmune-related variants[4]. There has already been significant literature about the genetic overlap of brain and immune diseases. Here we connect disease diagnoses and blood work to explore the co-occurrence as well as the clinical association of these diseases using a massive EHR/Claims dataset, Optum.

Methods

Data

For this study, we queried the Optum database. Stanford's Population Health Services provides a SQL server that contains the Optum data, which has administrative health claims covering a nine year period (2007-2015) and includes 15-18 million annual-covered lives for a total of roughly 63 million unique lives. After the pipeline described below, we focused on records containing 471 relevant ICD-9 codes (related to the immune and neurological diseases we are studying) and 12 lab tests (related to immune function).



Optum SQL Database (70 million individuals)

Filter for ~500 relevant ICD9s for people with neurodegenerative / immune related disease, query for 15 immune related lab tests along with doctor visit diagnoses and demographic information

Clean/filter data based on lab test values and units

Calculate the mean lab results for every patient

Remove 3 out of 15 labs tests with more than 75% of patients missing data

Remove individuals with < 2 lab tests (max missing is ~4%)

Standardize every lab tests value

Impute the NAs with 0's (standardized) mean value).

3 million patients

Merge imputed data with demographic data (age and gender)

Optum SQL Database of doctor visits and diagnoses

Merge ICD9 highest category (19 diseases) with collapsed patient/diagnosis matrix

Merge diagnosis matrix for patients (0's and 1's for 19 diseases) with imputed data

ML Tasks

Filter for individuals with mental disorders (1 million individuals)

Merge mental disorder imputed data with lower level of patient diagnosis (31 subtypes of mental disorders)

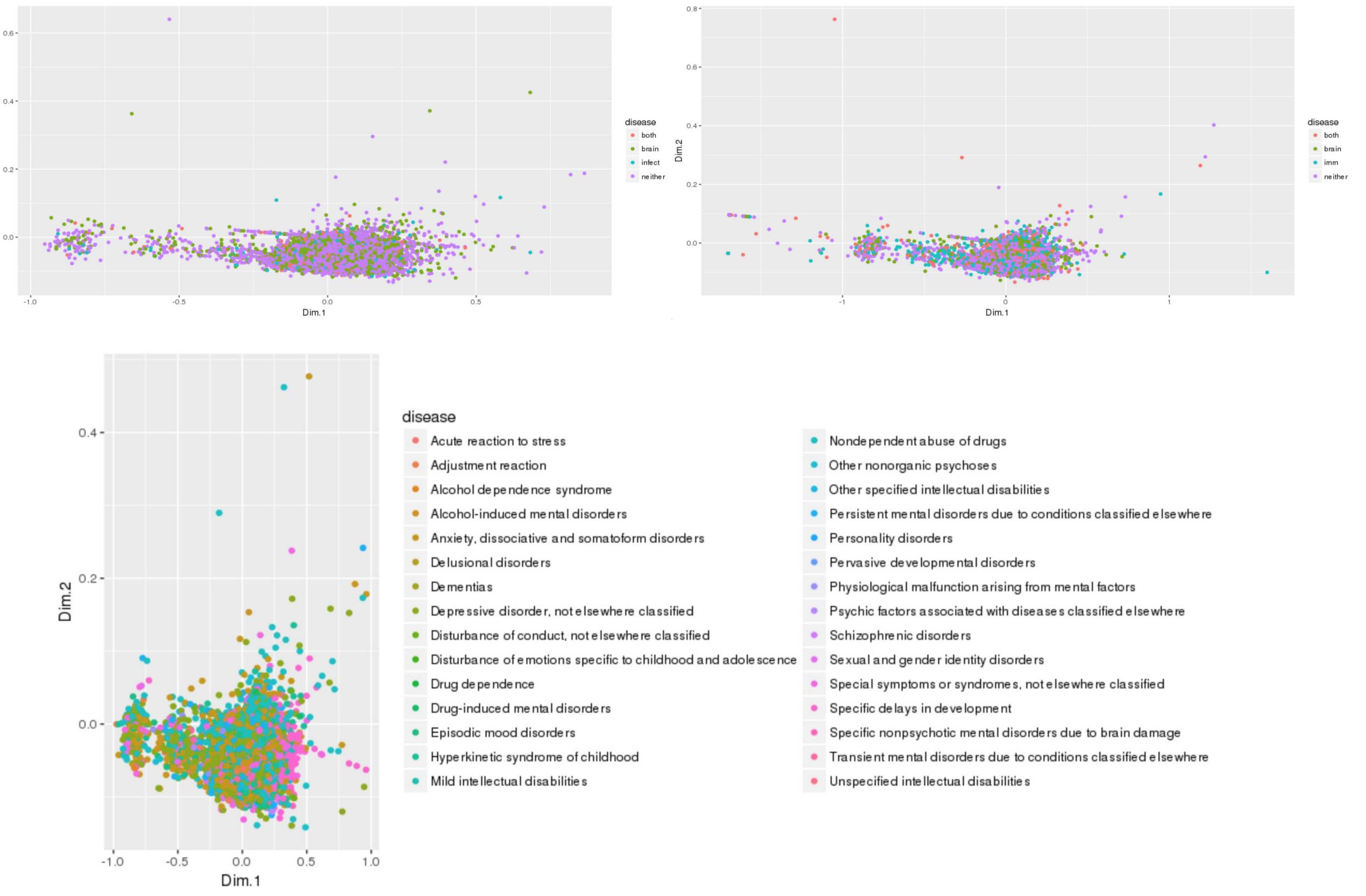
ML Tasks

Fig. 1. Preprocessing Pipeline Followed to Acquire and Clean Data.
A multitude of steps were involved to manipulate the data into the state it is in.

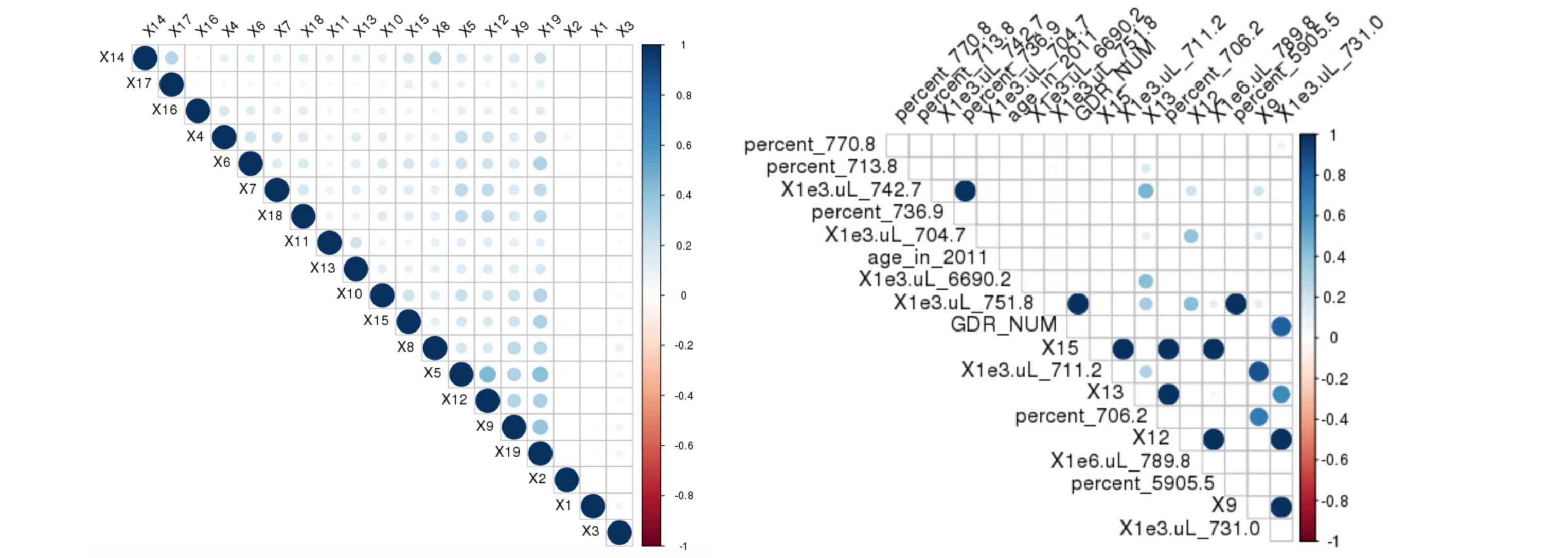
Models and Motivation:

Unsupervised Learning

We first wanted to explore whether or not the immune profiles of the patients' 12 immune-related lab tests were informative in finding what kind of diseases the patients were diagnosed with. To accomplish this, we hypothesized that there are distinct types of immune profiles associated with each type of disease (neurodegenerative, immune, infectious), and these "representative" immune profiles can be recovered using unsupervised machine learning methods. We postulated that there would be some structure within the data that renders individuals with neurodegenerative diseases and individuals with immune diseases somehow separable in this Optum dataset. We explored this notion using PCA on our data (the patient-feature matrix as described in the Preprocessing section). We also computed correlation matrices between features.



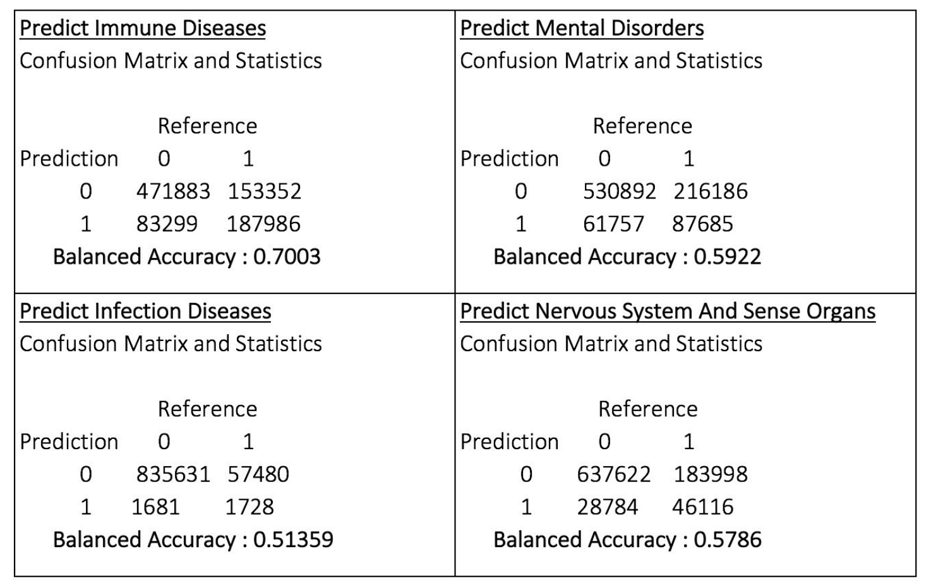
Figs. 2 (top left), 3 (top right), and 4 (bottom). Dimension 1 and 2 PCA plots comparing patients (2) brain and immune disorders, (3) brain and infectious disorders, (4) overall disease categories. There appears to be minimal clustering.



Figs. 4 (left) and 5 (right). (4) Comorbidity of disease. Correlation matrix of 19 top-level ICD-9 categories. Displays significant correlation between Endocrine, Nutritional And Metabolic Diseases, Immunity Disorders, Mental Disorders, Diseases Of The Nervous System And Sense Organs, and Infectious And Parasitic Diseases. **(5) p-value matrix between lab tests and relevant ICD-9 disease categories.** White signifies a p-value of 0; thus, many diseases are significantly associated with disease categories and each other, but offer little in the way of correlation.

Supervised Learning

We also wanted to explore how predictive our data was in finding presence of neurodegenerative, immune, infectious, and sensory diseases. That is, we wanted to see how well top-level ICD-9 codes and lab results can predict the presence of these categories of disease. Since each of these rough categories (neurodegenerative, immune, infectious, and nervous system diseases) map to top-level ICD-9 codes, we trained four logistic regression models, taking all other features in our patient-feature matrix as our features and the one feature of interest as outcome, in each of the four cases.



Predict Immune Diseases
Confusion Matrix and Statistics

Prediction	Reference 0	Reference 1
0	471883	153352
1	83299	187986

Balanced Accuracy : 0.7003

Predict Mental Disorders
Confusion Matrix and Statistics

Prediction	Reference 0	Reference 1
0	530892	216186
1	61757	87685

Balanced Accuracy : 0.5922

Predict Infection Diseases
Confusion Matrix and Statistics

Prediction	Reference 0	Reference 1
0	835631	57480
1	1681	1728

Balanced Accuracy : 0.51359

Predict Nervous System And Sense Organs
Confusion Matrix and Statistics

Prediction	Reference 0	Reference 1
0	637622	183998
1	28784	46116

Balanced Accuracy : 0.5786

We also tested the predicted power of multinomial logistic regression to predict the presence of different neurodegenerative diseases (31 subtypes); however, we were generally unable to predict disease subtypes with sufficient accuracy.

Conclusions and Future Directions:

In this research, we have found a significant connection between immune-related lab tests, immune and infectious diseases, and neurodegenerative diseases. However, there is still work left to be done before we have a robust predictive model for neurodegenerative disease. In addition, PCA was unable to detect distinct immune profiles between individuals that had differing neurodegenerative diseases.

Future experiments would further utilize the Optum data by using all lab tests rather than a strict subset of immune-relevant tests. We would take a more unsupervised approach to discovering lab tests that are predictive of disease type while reengineering our pipeline to account for longitudinal data. On the modeling front, we would like to explore multiple-output neural networks for diagnosis prediction, with one neuron per top-level ICD-9 code. We acknowledge that these tests are just scratching the surface of what can be done with this data. Lab test data in large medical claims datasets are largely untapped; given their tremendous potential for biomedical discovery in precision medicine, we would like explore more connections between them and disease.

References:
[1] Banks, W.a. "The blood-Brain barrier in neuroimmunology: Tales of separation and assimilation." *Brain, Behavior, and Immunity*, vol. 44, 2015, pp. 1–8., doi:10.1016/j.bbi.2014.08.007. [2] Mascagni P, Bellini GB. 1816. Istoria Completa Dei Vasi Linfatici. Vol. II Florence: Presso Eusebio Pacini e Figlio. p. 195. [3] Absinta, Martina, et al. "Human and nonhuman primate meninges harbor lymphatic vessels that can be visualized noninvasively by MRI." *ELife*, vol. 6, Mar. 2017, doi:10.7554/elife.29738. [4] Pouget, Jennie G, et al. "Genome-Wide Association Studies Suggest Limited Immune Gene Enrichment in Schizophrenia Compared to Five Autoimmune Diseases." 2015, doi:10.1101/030411. [5] Yokoyama, Jennifer S. et al. "Association Between Genetic Traits for Immune-Mediated Diseases and Alzheimer Disease." JAMA neurology 73.6 (2016): 691–697. PMC. Web. 20 Oct. 2017.