Investigating Links Between the Immune System and the Brain from Medical Claims and Laboratory Tests

Guhan Venkataraman
Department of Biomedical Informatics
Stanford University
Stanford, CA
guhan@stanford.edu

Tymor Hamamsy
Department of Biomedical Informatics
Stanford University
Stanford, CA
tymor@stanford.edu

Alex Chu
Biophysics Program
Stanford University
Stanford, CA
alexechu@stanford.edu

Keywords—immune system, neurological, brain, laboratory, claims, ICD-9, Optum, principal components analysis, logistic regression

I. Introduction and Related Work

The blood-brain barrier (BBB) is a system of tight cell-cell junctions that regulates the permeability of the neurovascular interface[1]. In the decades after its discovery, it was held by physiologists as a largely impenetrable barrier separating the brain from the circulatory system, and in particular, the immune system. Experiments showed that the BBB prohibited the passage of most ions, macromolecules, drugs, and neurotransmitters[2]. In recent years, though, this view has gradually transitioned to an understanding of the BBB as a regulatory interface between the central nervous and immune system[3]. Several immunomodulatory interactions were discovered that cross the BBB and affect neural function as well as immunological state, and new links between the central nervous and immune systems continue to be uncovered. Earlier this year, it was shown that lymphatic vessels in the brain were captured by an NIH team using advanced MRI technology[4], a discovery that had first been conjectured by anatomists over 200 years ago[5].

It now seems appropriate and timely to leverage the modern availability of large biomedical datasets to investigate more deeply the effects that the immune system has on the brain. In this project, we use data mining and machine-learning to query and analyze electronic health records and medical claims data from Optum[6], a massive EHR/Claims dataset covering 63 million patients over a 10 year time frame, to find connections between the immune system and disorders of the brain. The input to our algorithm is blood lab test results and ICD-9 (disease diagnosis) codes for approximately 3 million patients. We use principal components analysis (PCA) and correlation analysis to develop the connection between immune and brain diseases. We then use logistic regression and random forest to output a mental, immune, infectious, or sense-organ disease profile, as well as the subtypes of mental disorder phenotype.

II. Related Work

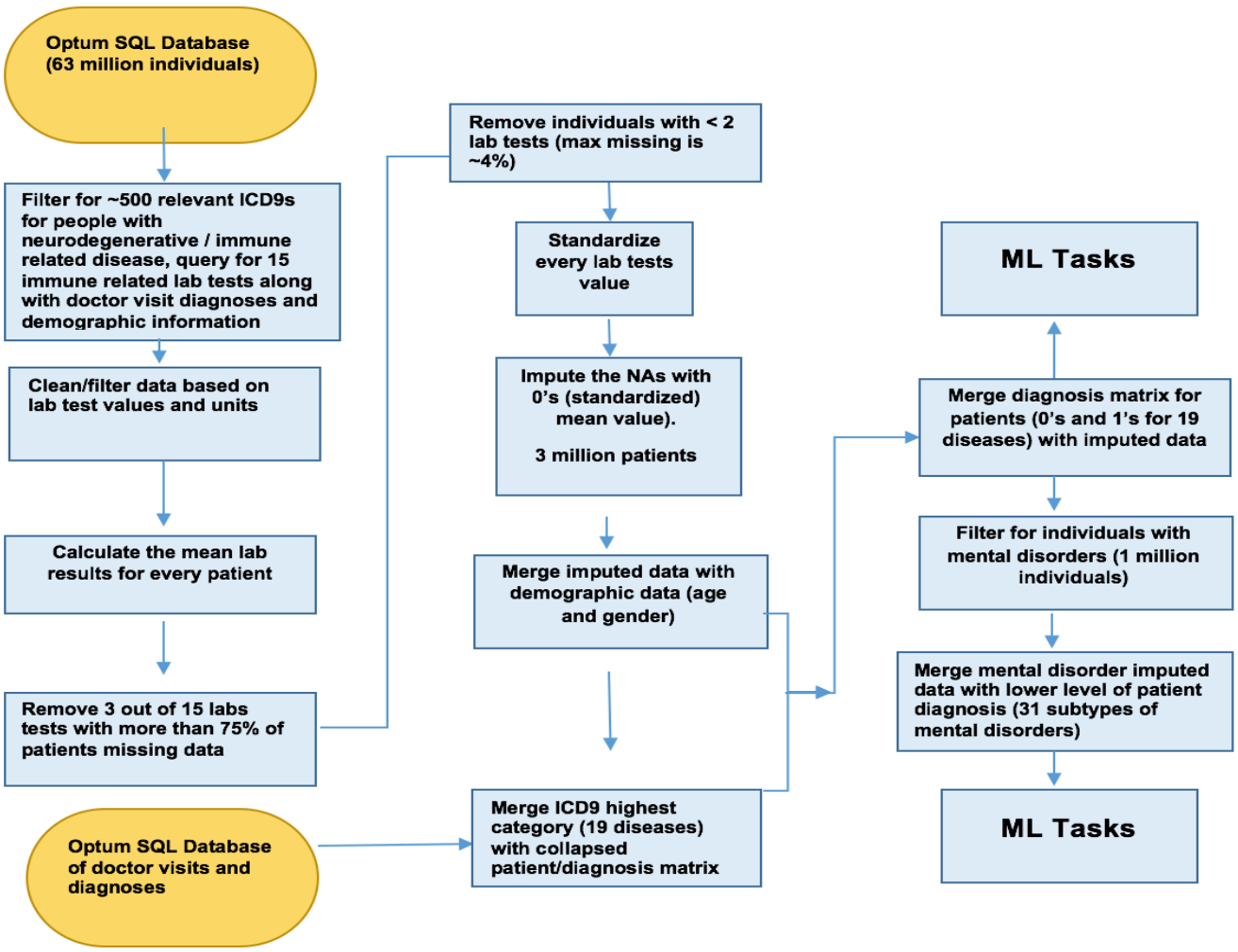
To our knowledge, no rigorous statistical analyses of the correlation between immune lab tests and the future onset of neurodegenerative disorders has been conducted. What few attempts exist at using machine learning to predict mental disorders have suffered from a variety of challenges, most glaringly the availability of large, quality datasets, and the effectiveness and speed of learning algorithms[7]. Genome-wide association studies have been conducted showing associations between many of the variants found associated with mental disorders (e.g. schizophrenia, bipolar disorder, and autism) and autoimmune related variants[8]; however, genetic-level profiles explain only a small proportion of these diseases. Furthermore, these studies used simple statistical correlation, without leveraging the complex models and unique strengths available with machine learning approaches. In another study, known as "Deep Patient," Miotto et al. developed an autoencoder to deterministically map patient electronic health records into a different feature space, and then use this broad representation to predict future health states and outcomes[9]. However, this more general approach, while useful for evaluating patients as a whole, is prone to suffering from noisy features when predicting a relatively specific single outcome, such as a relationship between immune profiles and mental or neurodegenerative disorders. We hope to use specific, targeted machine learning on individual- and systems-level data to answer the questions at hand.

III. Dataset and Features

The clinical data for this work comes from the Optum dataset, which is provided by Stanford's Population Health Services group and created by United Healthcare. Significant effort was required to acquire and preprocess our dataset. The Optum data is

[1] Neuron. Volume 57, Issue 2, 24 January 2008, Pages 178-201.
[2] Neurobiology of Disease. Volume 37, Issue 1, January 2010, Pages 13-25.
[3] 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.
[4] 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.
[5] Mascagni P, Bellini GB. 1816. Istoria Completa Dei Vasi Linfatici. Vol. II Florence: Presso Eusebio Pacini e Figlio.
[6] https://www.optum.com/solutions/prod-nav/integrated-claims-ehr-data.html
[7] Bzdok, D. and Meyer-Lindenberg, A. arXiv.
[8] 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.
[9] Miotto, R., et al. Scientific Reports. 6:26094. DOI: 10.1038/srep26094

FIGURE 1: PREPROCESSING WORKFLOW AND DATA EXAMPLE



X17	X18	X19	X1e3.uL_6690.2	X1e3.uL_704.7	X1e3.uL_711.2	X1e3.uL_731.0	X1e3.uL_742.7	X1e3.uL_751.8
0	0	0	0.045930535	-0.0392963175	-0.034957610	-0.014489594	-0.020638884	-1.820846e-02
0	1	0	-0.052973624	-0.0087256850	-0.006305739	-0.043737624	-0.030758329	-3.779154e-02
0	1	1	-0.007564796	-0.0392963175	-0.051330108	-0.038062633	-0.040877774	-2.115645e-02
0	1	0	0.095693634	-0.0392963175	-0.034957610	-0.019728047	-0.016591106	1.585181e-03
0	0	0	-0.023737803	-0.0392963175	-0.018585112	-0.013616518	-0.036829996	-4.052896e-02
0	1	0	-0.085941677	-0.0392963175	-0.051330108	-0.045920312	-0.053021108	-4.263467e-02

Figure 1. The complete process from start to finish (above), and a slice of the dataset (below). Not all examples or features are shown. "X*" refers to a top-level ICD-9 code; "X1e3.uL_704.7" refers to a lab test result measured in 1000 μL and a test code of 704.7. Age and gender not shown in this example.

contained within several SQL databases, so we developed SQL pipelines to access and extract the relevant data for our study (e.g. immune-related lab tests for patients who have been diagnosed with immune, mental, or neurodegenerative disorders). We manually curated a list of these disorders, querying the Optum database for 471 ICD-9 codes related to the diseases we are studying and 15 lab tests (blood tests done to assay immune function; see Appendices for details). We also extracted all of the doctor visits (including diagnoses) and demographic information (age, year of birth) for patients with these diseases. Once these data were obtained, we had to preprocess the data to prepare it for statistical analyses. Discrepancies in measurement units were removed, measurements were zero-mean centered and scaled, and missing values were imputed by replacing them with the zero-mean value. After examining patient time-series data and observing that little variation occurred over time for a majority of the cases, we also elected to collapse these longitudinal measurements into a single mean value independent of time of measurement.

Our final imputed data set included a total of n = 33 features: a binary variable matrix of the 19 top level ICD-9s, measurements for 12 lab tests (standardized), and values for 2 covariates (age and gender). After preprocessing, we had data on these 33 features for approximately 3 million individuals. Each individual represented a single training example. We opted not to use a cross

validation partition due to compute power restrictions, but we divided the dataset into a 70-30 split for training and test sets.

IV. Methods

In addition to assessing correlation between input variables, we applied unsupervised and supervised learning techniques to the data. Unsupervised learning is generally useful for teasing out complex relationships and patterns in the dataset that are hard to uncover manually, with little prior knowledge about the underlying structure of our data. We used principal components analysis (PCA) to attempt to recover this structure. Next, we applied logistic regression with Lasso regularization and Random Forest to assess the power of immune profiles (as described by lab tests and ICD-9 diagnoses) in predicting disease types.

A. Unsupervised Learning

PCA is a method for projecting a set of points in a high-dimensional space into a lower-dimensional subspace. It does so by computing the principal components of the data, or the basis vectors which will produce the greatest variance when the data is projected onto them. Expressed another way, there is some variance inherent in the dataset. The first principal component is the vector which preserves the most of this variance when we project all of the data onto it, expressed mathematically as:

$$argmax(u)\frac{1}{m}\sum_{i=1}^{m}(x^{(i)T}u)^2$$

The second principal component is similarly calculated to maximize the variance of the data projections onto it, with the restraint that it must be orthogonal to the first principal component. The same pattern follows for the third principal component, which must be orthogonal to the first two, and so on. We can choose the dimensionality of the subspace which we project the data into by selecting the number of principal components onto which we want to project the data. For this project, we have chosen to use the first two principal components of the data, which allows us to construct a two-dimensional visualization of the data.

B. Supervised Learning

Logistic regression is regarded as an excellent "off-the-shelf" classification algorithm that does not make too many strong assumptions about the data. It outputs a predicted value between 0 and 1 by applying the logistic, or sigmoid, function to a linear transformation of the data:

$$h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$$

where theta parameterizes the model and is the weight vector applied to the data vector, and the bias or intercept term is built into the theta vector. We learn optimal values of theta typically by stochastic gradient descent, which iterates through the training data and applies the update rule:

$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}$$

Here alpha is the learning rate. We use a Lasso-regularized version of logistic regression, which sets a budget or constraint on the weight vector parameter theta, such that the magnitude of theta must be less than the budget. This prevents excessive growth of the weight vector in attempting to fit the noise in the dataset, and induces the exclusion of features that are irrelevant to the model's prediction. Finally, multinomial or softmax logistic regression is a generalization of logistic regression to situations with more than two discrete output variables. We simply replace the sigmoid function with the softmax function:

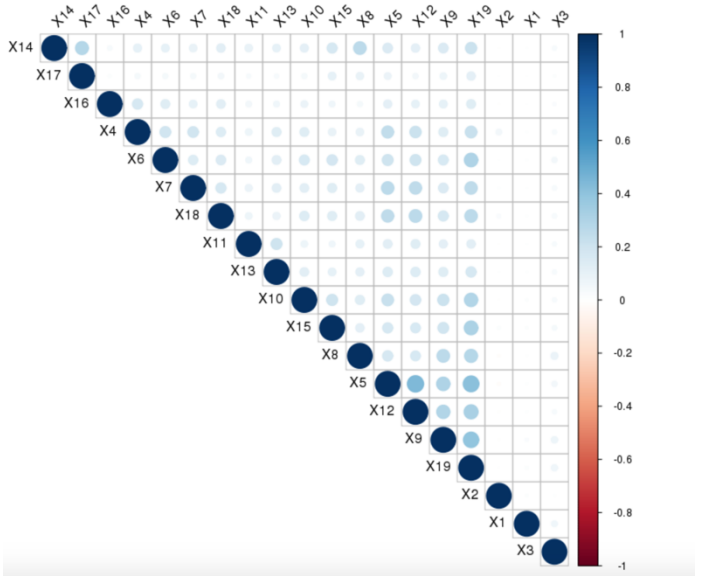
$$h_{\theta j}(x) = \frac{\exp(\theta_j^T x)}{\sum_{k=1}^{K}\exp(\theta_k^T x)}$$

The other supervised learning algorithm which we utilized is Random Forest. This is a bootstrap aggregating variation of the decision tree learning algorithm that is intended to preserve the benefits of decision trees, such as interpretability and robustness against noisy features, while solving their primary weakness, a tendency to overfit the data. Random forest works by repeatedly training decision trees on randomly sampled subsets of the data, using a randomly sampled subset of the features. These changes help minimize the variance of the model and reduce correlation between features, respectively.

V. Results and Discussion

With data acquired and cleaned, we first performed comorbidity analysis, investigating the co-occurrence of diseases as well as top-level ICD-9 categories (Fig. 2).

Figure 2: Disease Co-occurrence



Infectious And Parasitic Diseases	X13
Neoplasms	X16
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	X12
Diseases Of The Blood And Blood-Forming Organs	X4
Mental Disorders	X15
Diseases Of The Nervous System And Sense Organs	X9
Diseases Of The Circulatory System	X5
Diseases Of The Respiratory System	X10
Diseases Of The Digestive System	X6
Diseases Of The Genitourinary System	X7
Complications Of Pregnancy, Childbirth, And The Puerperium	X2
Diseases Of The Skin And Subcutaneous Tissue	X11
Diseases Of The Musculoskeletal System And Connective Tissue	X8
Congenital Anomalies	X3
Certain Conditions Originating In The Perinatal Period	X1
Symptoms, Signs, And Ill-Defined Conditions	X19
Injury And Poisoning	X14
Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services	X18
Supplementary Classification Of External Causes Of Injury And Poisoning	X17

Figure 2. Pairwise comorbidity analysis between top-level ICD-9 diagnoses. Disease diagnoses such as X15 and X12, which occur together with some frequency, can be predictive of each other.

As can be seen in Figure 2, there is considerable comorbidity of mental disorders with infectious and immune diseases. We then examined the correlation between the ICD-9 disease diagnosis codes we were interested in and immune-related lab tests and found significant correlation (as seen in Figure 3), albeit weak, between the top-level ICD-9 code disease diagnoses and the immune-related lab tests. These analyses turned out to be relevant in the context of interpreting our machine learning results later on. By deducing some of the interrelationships between the features in our dataset, and between disease diagnoses (comorbidity analysis), this provided some degree of interpretability for our later machine learning results. That is, we were then able to compare the features that turned out to generate greater predictive power with the features that had some correlation with each other to corroborate that these key factors worked together to generate predictive power.

FIGURE 3: DISEASE CO-OCCURRENCE

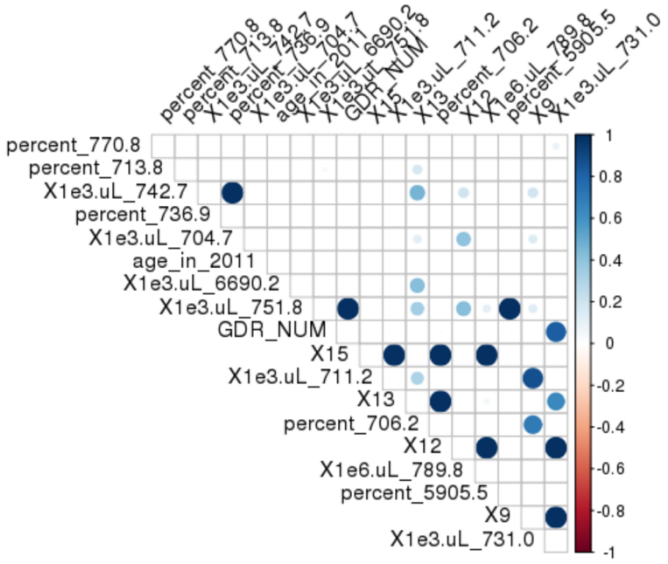
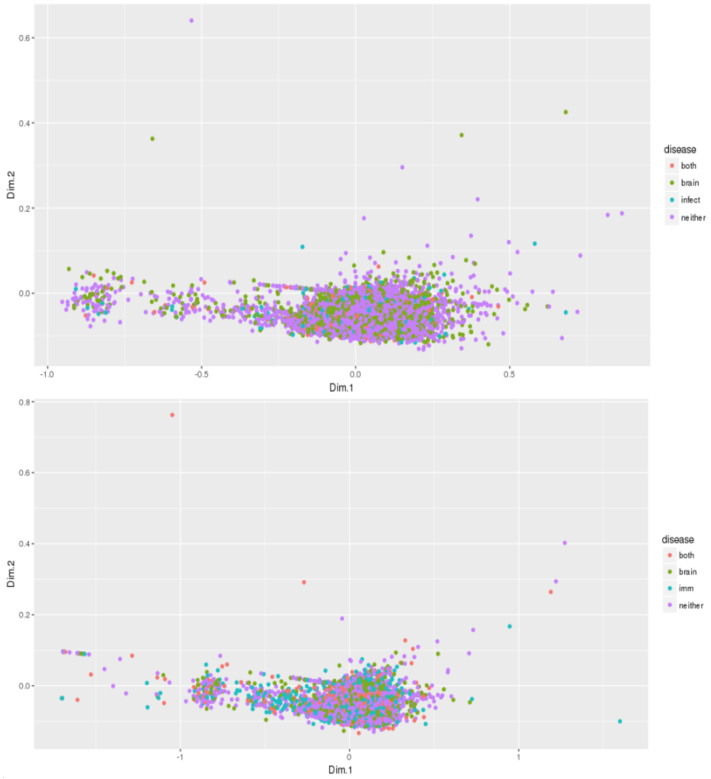


Figure 3. p-value matrix of correlation between lab tests and relevant ICD-9 disease categories. White signifies a p-value of 0; thus, mental disorders are significantly associated with many immune-related lab tests. However, actual correlation signal (effect size) is low (correlation values between .04 and .1)

Next, we explored 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 that these "representative" immune profiles can be recovered using unsupervised machine learning methods. We postulated that there would be some structure within the lab test data that renders individuals with neurodegenerative diseases and individuals with immune diseases separable in this Optum dataset. We explored this notion using PCA on our data. In a similar analysis, we also computed correlation matrices between features to see if any features were correlated with each other. We observed some clustering in the data; notably a large, central cluster and a smaller side cluster, and correlations between some of the features, which is expected, since our data projects down effectively with PCA.

FIGURE 4: DIMENSIONALITY REDUCTION



PC component	eigenvalue	% varianced explained	cumulative % variance explained
comp 1	5.373810942	44.78175785	44.78176
comp 2	4.091630559	34.09692132	78.87868
comp 3	0.838010973	6.98342478	85.8621
comp 4	0.59594825	4.96623542	90.82834
comp 5	0.387682836	3.2306903	94.05903
comp 6	0.343040356	2.85866963	96.9177
comp 7	0.151839451	1.26532876	98.18303
comp 8	0.104847652	0.87373043	99.05676
comp 9	0.068521947	0.57101623	99.62777
comp 10	0.02560307	0.21335892	99.84113
comp 11	0.013677028	0.11397523	99.95511
comp 12	0.005386935	0.04489112	100

Figure 4. Data projected onto principal components 1 and 2 (top) and the percentage of variance in the dataset explained by each principal component (bottom).

Finally, we conducted supervised learning analyses to explore how predictive our data was in finding the presence of neurodegenerative, immune, infectious, and sensory diseases. That is, we wanted to whether or not we could predict the presence of these diseases given top-level ICD-9 codes and lab results for all patients. 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 using Lasso regularization, taking all other features in our original patient-feature matrix as our input features and the one feature of interest as output, in each of the four cases. For example, in assessing our ability to predict immune diseases, we left out the feature corresponding to immune diseases ("X13") when training, and then tried to predict it using the trained model (Figure 5). Interestingly, training separate logistic regression models on only lab tests and only ICD-9 diagnoses in isolation revealed that the majority of the predictive power attained in the combined model was explained by the ICD-9 diagnoses (results not shown). This could perhaps indicate that immune lab work may not be directly related to mental diseases, but that comorbidities would be a more effective predictor. We did not employ cross validation in any of these models due to compute

power restrictions; training a single model alone took several hours or days.

FIGURE 5: SUPERVISED LEARNING RESULTS

Predict Immune Diseases
Confusion Matrix and Statistics

Prediction	Reference 0	1
0	471883	153352
1	83299	187986

Balanced Accuracy : 0.7003

Predict Mental Disorders
Confusion Matrix and Statistics

Prediction	Reference 0	1
0	530892	216186
1	61757	87685

Balanced Accuracy : 0.5922

Predict Infection Diseases
Confusion Matrix and Statistics

Prediction	Reference 0	1
0	835631	57480
1	1681	1728

Balanced Accuracy : 0.51359

Predict Nervous System And Sense Organs
Confusion Matrix and Statistics

Prediction	Reference 0	1
0	637622	183998
1	28784	46116

Balanced Accuracy : 0.5786

Random Forest Predict Mental with only labs

Prediction	Reference 0	1
0	18734	8858
1	1048	1244

Balanced Accuracy : 0.5351
P-Value [Acc > NIR] : 0.008336

Random Forest Predict Mental with all features

Prediction	Reference 0	1
0	17838	5618
1	1944	4484

Balanced Accuracy : 0.6728
P-Value [Acc > NIR] : < 2.2e-16

Figure 5. Machine learning models achieve predictive power when trained on our dataset. (Top) Using our patient-feature matrix as the design matrix, we used logistic regression to predict different columns of the matrix (immune, mental, infectious, and nervous system disease diagnoses each) using the rest of the matrix as input data. We achieved only modest performance for most of the tests, but did obtain predictive power for immune diseases. (Bottom) Using a random forest model, we were able to generate much improved accuracy when trying to predict mental disorders. Note that the lab test data contributed only weakly, if at all, to the prediction of mental disorders.

We achieved relatively poor predictive power on all of the disease predictions except for immune diseases. To assess whether this was due to a poor choice of model, we tried predicting mental disorders again, but using a different algorithm, random forest, which is more robust to overfitting and is more effective for working with categorical data inputs. Due to limitations on compute power, we ran the algorithm on a random subset of the data. Promising results show already as we had hoped, as random forest achieved better results overall on the testing examples than logistic regression (Fig. 5). Interestingly, we also split the training data and tried to predict solely on lab test data or solely on diagnosis data, but found that our predictive power stemmed primarily from the diagnosis data, in a pattern that was alluded to in our correlation analysis but not fully reflected as in the machine learning model. This may point to a weak or nonexistent connection between the immune state of a patient and the likelihood of developing neurodegenerative diseases. We hope to continue this analysis moving forward by training and testing on the full dataset to achieve even greater predictive power, as well as utilizing other supervised machine learning algorithms.

VI. CONCLUSION

In this project, we have uncovered connections between immune-related lab tests, immune and infectious diseases, and neurodegenerative diseases, but PCA was unable to detect distinct immune profiles between individuals that had differing neurodegenerative diseases. We have discovered that the predictive power of comorbidities to label patients with different types of disease using simple logistic regression models is relatively high for immune diseases, even after the simplest of feature engineering. This connection points to promising avenues for future work in developing predictive models of disease, although there is still work left to be done. Our compute power limitations prevented us from trying more complex machine learning algorithms on our data, such as support vector machines and neural networks, which have had positive outcomes in biomedical informatics recently; these, combined with more intelligent feature engineering, would most likely result in significant boosts in predictive power.

We acknowledge that these tests are just scratching the surface of what can be done with this data. For example, we would like to take a more unsupervised approach to discovering lab tests that are predictive of disease type while reengineering our pipeline to account for the incredible depth of longitudinal data Optum has. 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.

ACKNOWLEDGMENTS

The authors would like to acknowledge the CS 229 instructors and teaching assistants for guidance and constructive comments, and the members of the Khatri Lab at Stanford for constructive comments and assistance with data acquisition.