Harrison Zhang
Applied Data Mining
May 7, 2022

Data Mining to Find Patterns of Long COVID in Electronic Health Records

GitHub Repo: https://github.com/harrisonzhang1/data_mining_final

Introduction

As vaccines and pharmaceutical drugs have become increasingly available in the United States for preventing and treating severe COVID-19 disease that may result in death, the focus has instead shifted to understanding the emerging disease that is known as long COVID. As defined by the CDC, long COVID is the name given to long-term effects of being infected by the virus, and it may include a wide range of ongoing health problems that last for weeks or even months. Examples of long COVID that have been reported in emerging studies include fatigue, inability to exercise, persistent coughing, chest pain, and difficulty thinking or concentrating. Long COVID bears such a large burden on patients that national policymakers have formally recognized it as a disability under the Americans with Disabilities Act (ADA). To this day, there is no established standard treatment strategy for long COVID due to limited understanding of the disease.

As this is an emerging disease, there is a general consensus that much work needs to be done to learn more about long COVID as public health scientists and healthcare providers seek to develop treatment policies for the disease. Since electronic health records (EHRs)—which have information recorded by physicians during routine interactions with patients—contain a wealth of information covering a wide range of patient populations and would not require significant data collection efforts, there is interest in examining whether existing EHR data can be leveraged to find disease patterns in post-COVID patients to better understand the wide spectrum of conditions that emerge after an initial infection. This can be realized if EHR data were linked to COVID-19 testing results for example. If successful, the identification of patterns of conditions that emerge after initial infection in COVID-19 patients would arm public health scientists and healthcare providers with increased knowledge about this newly emerging disease

and help inform the development of treatment strategies for this debilitating disease. Even more, it would demonstrate the feasibility of using existing EHR data to study long COVID.

More specifically, this project aims to link EHR data with COVID-19 testing results to identify patients with COVID-19 and then find meaning clusters of these patients. The goal is to group together patients which have similar groups of newly diagnosed conditions after initial infection, and then study the subpopulations of patients within clusters to better understand what groups of conditions are represented within COVID-19 patients to better understand the different forms of long COVID captured in the EHR data. Given the understanding that COVID-19 is a disease that primarily involves the lungs, it is very likely that some clusters reflect many respiratory conditions that persist after initial infections.

Description of the Data and Data Sources

EHR data representative of patients from the Mass General Brigham (MGB) healthcare system were available for this analysis. The EHR data were gathered as part of routine healthcare provider-patient interactions across the 14 hospitals that make up the MGB healthcare system and, the EHR data is stored in a centralized clinical data warehouse in the form of SQL tables. Data were extracted from SQL databases using SQL queries for downstream data wrangling and analysis in RStudio. The data is professionally managed by the MGB Research Patient Data Registry team, and they are responsible for the integrity of the data.

The available EHR data contains information such as historical records of hospital admissions and phenotype information in the form of diagnosis codes. Each patient can be identified by several different types of unique identifiers, such as a unique MRN code or a unique patient number code. The diagnosis codes in the EHR data are entered into a patient's EHR for routine billing purposes and provide phenotypic information on what diseases or conditions a patient has. For example, there are diagnosis codes for "Celiac disease" or "shortness of breath." The diagnosis code name and its calendar date of observation were available.

The MGB healthcare system is also a designated COVID-19 testing center and records the date and results of COVID-19 tests that were either performed or processed at their laboratory facilities. Using the relevant unique patient identifiers, COVID-19 testing data can be linked to routinely collected EHR data for each patient in order to identify patients who have

been exposed to the virus and provides a complementing feature that enables us to identify who the COVID-19 patients are and when they were infected. Then, we can look for patterns in the diagnosis codes that were observed after initial infection in these patients to learn more about what types of conditions arise after infection.

Data Wrangling and Preprocessing

To clean the COVID-19 testing data, patients with a positive COVID-19 test result were first identified. If there were multiple positive test results reported on different calendar dates for patients, the earliest recorded calendar date with a positive result was used, and the others were discarded. To clean the EHR diagnosis code data, repeated observations of the same diagnosis code within a patient's EHR on the same date were removed. After both of these pre-processing steps, the EHR data was linked to the COVID-19 positive test results by the patient number unique identifier. The date of the positive COVID-19 test result (which approximately represents the date of infection for the patient) was then set as the index date, otherwise known as day 0. By comparing the index date and the date of observation of a diagnosis code, a "days_since_infection" variable was constructed, which is defined as the time difference (in units of days) between the infection date and the observation of a diagnosis code. If a diagnosis code was observed two weeks after the initial infection date and another was observed a month after the initial infection date, the corresponding "days_since_infection" values would be day 14 and day 30 respectively. New onset codes, defined as a diagnosis code which was first observed after the initial infection date and never previously observed in the patient's EHR before the initial infection, were identified. All subsequent analyses used new onset diagnosis codes that were first observed after infection with COVID-19 to decrease the likelihood that existing conditions would be mistakenly associated with a long COVID.

Additional Feature Engineering

The CDC's website has also suggested that patients who initially had more severe forms of disease from COVID-19 may be at more risk for developing long COVID. To engineer a variable which reflects initial disease severity of COVID-19 that can be used in downstream clustering analyses, I engineered a COVID-19 hospitalization flag. This flag is a binary variable which indicates whether or not the patient was hospitalized during the initial infection of

COVID-19. I hypothesize that this variable will reflect the patient's initial disease severity as patients hospitalized during their initial infection with COVID-19 were most likely more in need of urgent medical attention. The feature was engineered as follows: hospital admissions data pertinent to the MGB healthcare system were obtained and linked to the COVID-19 testing dataset by mapping between different patient identifier systems. If there was a record of hospital admission up to seven days before the date of the first positive COVID-19 test or up to seven days after, then the patient would be counted as a "COVID-19 hospitalized patient." I selected this time window recognizing that there needs to be some flexibility in the overlap of the hospital admissions date and the COVID-19 test result date because of typical healthcare systems delays in processing, obtaining, and recording observations. Ultimately, 18,588 patients out of 69,563 patients (27%) were marked as "COVID-19 hospitalized patients." This variable would be used in downstream clustering methods to mine the data.

Exploration of the Processed Data

After data linkage, cleaning, and preprocessing, the study patient population was 69,563 COVID-19 patients whose infections were confirmed with a laboratory COVID-19 test result. The cleaned data frame containing new onset diagnosis codes had 2.56 million rows, reflecting the number of diagnosis codes observed from March 10, 2020 to March 14, 2022 across the patient population. There were five columns: patient identifier, diagnosis code, diagnosis code name, date of observation, and days observed from initial infection. Only diagnoses codes first observed within 90 days of COVID-19 infection in a patient's EHR were included as these possibly could be considered a form of long COVID while diagnosis codes first observed at farther dates from infection may be irrelevant. That is, if a diagnosis code was first observed within 90 days of COVID-19 infection and then again at 180 days, both observations would be included in the data. This is in accordance with the WHO's definition of long COVID. The median and average number of total diagnosis codes recorded per patient was 13 and 36.78 respectively, and there is a strong right skew in the distribution of diagnosis codes per patient. This indicates that most patients in the dataset have sufficient datapoints and the completeness of the data is acceptable. The median and average days from admission that diagnosis codes were observed was 113 and 139.8 days from initial infection respectively, indicating that many patients were recorded with diagnosis codes persistently. The mean and average number of

unique visits to their doctor upon/after initial infection was 4 and 7.91 respectively, and more than one third of patients had visited their doctor at least twice. The number of unique visits was estimated by calculating the number of unique calendar dates where diagnosis codes were observed. Across all patients, there were 1,766 unique diagnosis codes reflecting 1,766 unique diseases, and those pertaining to cardiovascular (15%), respiratory (12%), and metabolic (11%) diseases had the highest prevalence in the study population.

To gain a better understanding of a single patient's EHR in the produced dataset, I briefly describe an example patient. The patient was newly diagnosed with respiratory insufficiency; malaise and fatigue; renal failure, and chronic kidney disease in the first month from their infection, consistent with respiratory distress syndrome associated with initial infection of COVID-19. In the second month from their infection, the same patient was again noted with malaise and fatigue and chronic kidney disease as they were in the first month, with additional new diagnoses of generalized anxiety disorder and disturbances of skin sensation. In the fourth month from infection, the patient was again noted with malaise and fatigue; generalized anxiety disorder; and disturbances of skin sensation. The persistence of these new onset diseases in months 1 and 2 into month 4 suggestion these persistent symptoms may be long-lasting, consistent with reports of long COVID.

Data Mining with K-means Clustering Algorithm

To mine for patterns of long COVID and explore relationships between different features in the diagnosis codes data, I used the K-means clustering algorithm in an attempt to group patients together based on underlying patterns in their diagnosis codes recorded in their EHRs. Specifically, I am clustering only based on new onset diagnosis code data I constructed previously and the COVID-19 hospitalization flag I engineered. To identify the most meaningful features for clustering and reduce the importance of highly prevalent features, I created a TF-IDF matrix with patients in the rows and features in the columns. Here, I treated each patient's EHR as a document, and each diagnosis code as a term/word. I defined TF as the log-transformed frequency of each feature for each patient, and I defined IDF as $\log\left(\frac{N}{n_t}\right)$ where N is the total number of patients and n_t is the number of patients with the feature t. The final dimension of my TF-IDF matrix was 69,563 rows by 1,767 columns.

Then, to identify the most informative features for downstream clustering, I identified the maximum TF-IDF value across all patients for each feature. Among these values, I then identified the largest TF-IDF value among diagnosis codes used for very routine healthcare processes that are observed frequently in EHRs, such ordering of a screening test, laboratory tests, and presence of minor injuries or burns. In Figure 1, I show the distribution of the maximum TF-IDF values for each feature, and the red line represents the maximum TF-IDF value among the routine diagnosis codes that are frequently recorded in EHRs as a part of the healthcare delivery process. The features resting to the left of the red line are most likely not very informative as they have lower TF-IDF values than the red line value, and these features were subsequently removed in the downstream K-means clustering. There were 291 features that remained and were used in clustering.

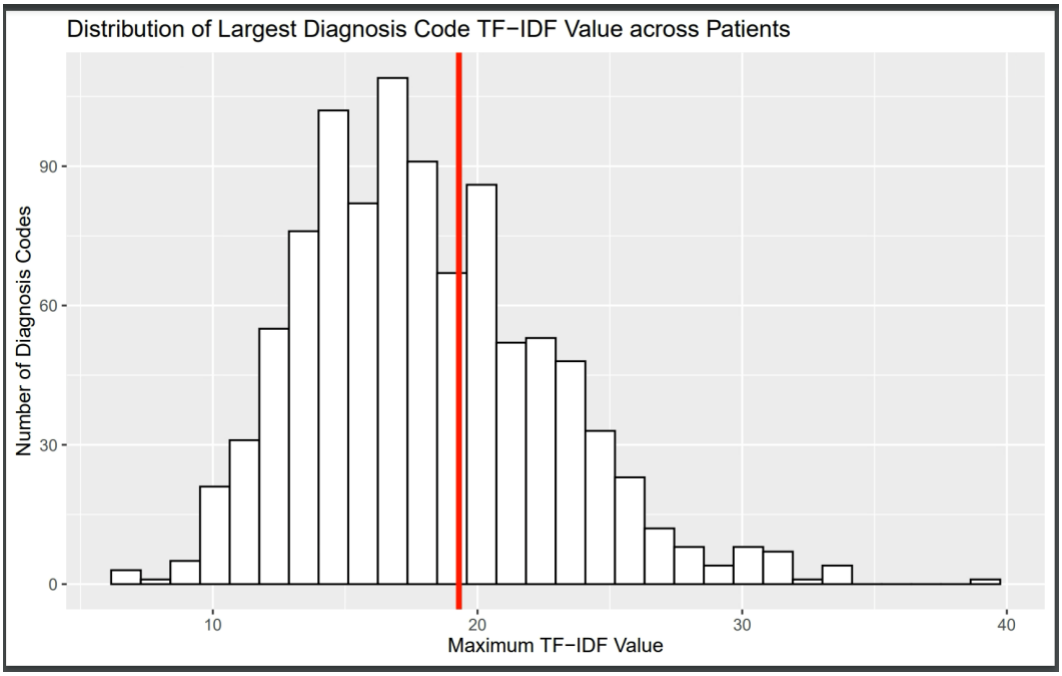


Figure 1: Distribution of Max TF-IDF Value in 1,767 Features

The K-means algorithm was then run with various possible cluster numbers ranging from 1 to 20. To identify the appropriate number of clusters that best fits the data and minimizes the differences within clusters while maximizing the differences between clusters, I show the ratio of between-cluster sum of squares against the total within-cluster sum of squares in Figure 2. I chose the largest number of clusters such that the ratio does not jump, and this would be the

number of clusters right before any jumps in the ratio value. Looking at Figure 2, it seems that there not an obvious big jump in the ratio value until we get to about K=17 clusters. However, interpretation of 17 different clusters would be quite difficult, so I looked for a smaller number of clusters, arriving at K=5 clusters. As seen in Figure 2, there is a first jump when going from K=5 to K=6 clusters, and K=5 clusters is much more interpretable than K=17 clusters.

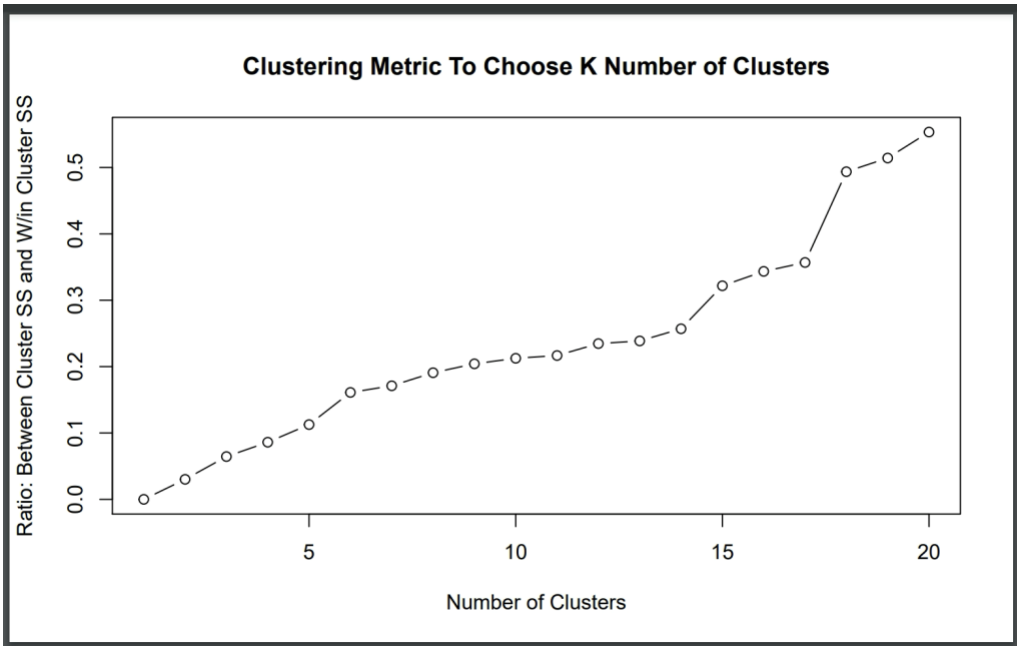


Figure 2: Ratio of between-cluster sum of squares against the total within-cluster sum of squares for $K = 1, \ldots, 20$ cluster values considered.

Validation and Interpretation of Results

To investigate whether the clustering results are due to chance (such as from data points which are outliers) or likely a real pattern, I used bootstrap resampling and refit the K-means clustering algorithm to the resampled datasets. I obtained 50 bootstrap samples by randomly sampling the original dataset with replacement and subsequently refit the K-means algorithm to each bootstrapped sample separately. To then assess whether or not the results are likely a real pattern, I identified the 50 features with the highest TF-IDF weight in each cluster, mapped these diagnosis codes to relevant body systems to understand what diseased body systems are being represented in each cluster of patients, and then graphed the proportion of the top 50 features that belong to each body system in a heatmap. I compared the original clustering results (Figure 3)

with heatmaps generated by clustering bootstrapped samples to assess if patterns were due to outliers or chance. In the majority of bootstrapped samples, the relative composition of clusters remains relatively consistent (the clusters 1, 2, 3, 4, and 5 are named randomly for each run of the algorithm, so I could not directly compare across all cluster 5's). We see this by comparing Figure 3 (original clustering results) with Supplementary Figure 1, 2, and 3 (heatmaps of bootstrapped samples), which shows that relative cluster composition in terms of effected body systems remains similar. However, there were a few samples which showed some variation in the clustering results, and this can be due to the random sampling of outliers during the bootstrapping (see Supplementary Figure 4). If there weas high uncertainty in the resampling results, this would dampen the conclusions one can draw from this analysis as it shows the results may be due to chance.

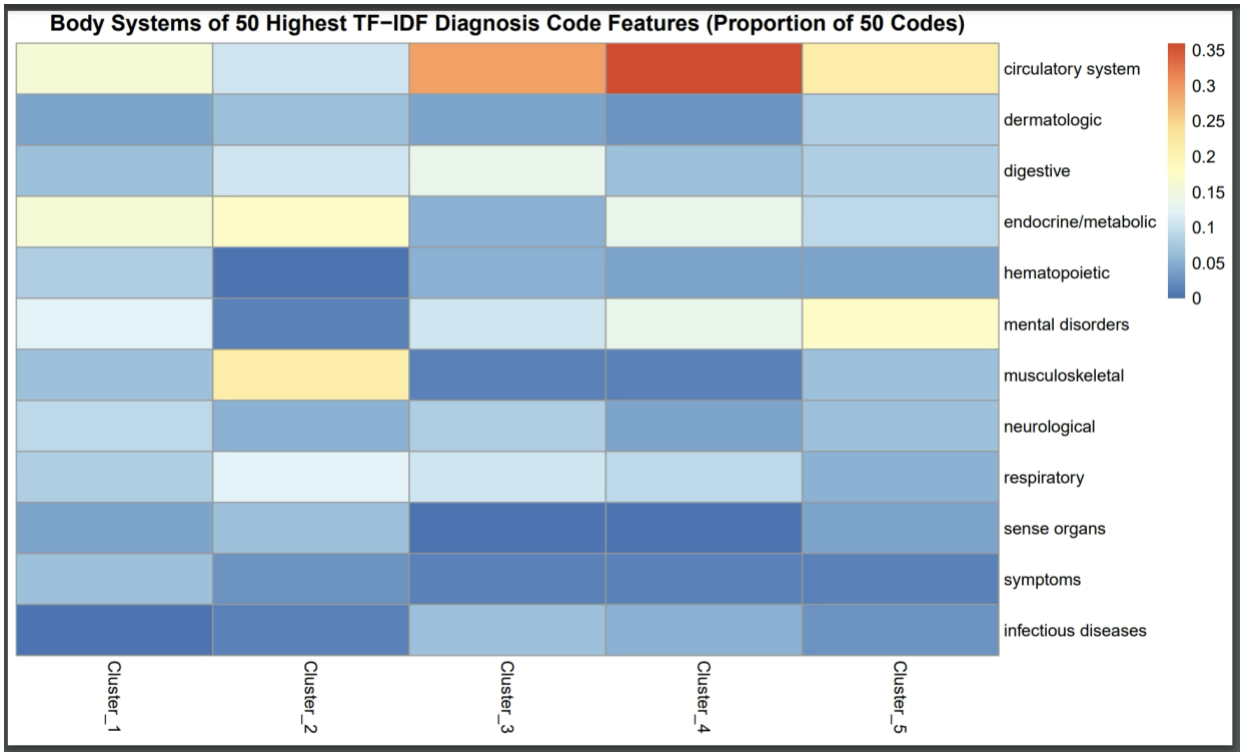


Figure 3: Summary of patient clusters showing the diseased body systems represented by diagnosis codes with the highest TF-IDF value within each cluster. The legend shows the proportion of the top 50 diagnosis codes with the highest TF-IDF value that belongs to each body system.

After validating that the observed results are less likely due to chance and more likely instead reflect real patterns in the EHR data, I examined Figure 3 better to learn how the K-means algorithm was grouping these patients. Based on Figure 3, it seems that the K-means algorithm was grouping patients based on the types of diseases they developed after infection as seen by the unique patterns of effected body systems between each cluster. Cluster 1 and cluster 4 seem very similar as there is an enrichment of persistent circulatory system diagnosis codes (such as congestive heart failure and arrythmias), endocrine/metabolic diagnosis codes (such as type 2 diabetes and obesity), and mental disorders (such as depression and generalized anxiety disorder). However, a greater proportion of the most informative diagnosis codes pertained to the circulatory system in cluster 4, and these diseases are often fatal, suggesting that patients in cluster 4 might require more long-term medical attention. This hypothesis was supported by the fact that 45% of cluster 4 patients were classified as COVID-19 hospitalized in contrast to 15% in cluster 1. The composition of cluster 2 is very interesting as there is the enrichment of musculoskeletal system related diagnosis codes (such as muscle and joint pains), which is not readily represented in the other clusters. This may represent a subpopulation of patients who have developed long-lasting musculoskeletal conditions after infection in accordance with emerging evidence. In cluster 3, about 48% of patients were COVID-19 hospitalized, and this subpopulation of patients differentiates itself from the others by its enrichment of digestive diagnosis codes (such as abdominal pain and persistent diarrhea). Finally, cluster 5 contains a subpopulation of patients with an enrichment of mental disorders (such as cognitive disfunction, brain fog, and generalized confusion), which has also been reported in other studies. Therefore, the K-means algorithm was able to identify distinct subpopulations of COVID-19 patients that are differentiated by newly developed conditions effecting different types of body systems. Further, the COVID-19 hospitalization flag that was engineered seems to also be meaningful in the clustering process as it was non-randomly assigned to clusters: clusters 1, 2, and 5 all had around 10-20% hospitalization rate while clusters 3 and 4 reflected around 47% hospitalization rate. Since clusters 3 and 4 were also the most enriched for circulatory system disorders, which require medical attention because of their life-threatening nature, the non-random distribution intuitively makes sense.

Conclusion

In this analysis, I linked MGB healthcare system EHR data with COVID-19 testing data to first identify COVID-19 patients and then cluster these patients using their post-infection diagnosis codes to identify patterns of potential long COVID. Using K=5 clusters, I was able to identify 5 subpopulations of patients with different combinations of new onset diseases that developed after initial infection. For example, the K-means algorithm identified clusters of COVID-19 patients that were enriched for mental disorders and another cluster with musculoskeletal disorders. The algorithm additionally identified subpopulations of patients enriched with circulatory/cardiovascular diseases, and almost half of these patients were initially hospitalized with COVID-19. The results were largely replicated when the clustering algorithm was fit to different samples of the data, suggesting the patterns were not likely observed by chance and may instead reflect real patterns in the data. Thus, these results underscore the multifaceted diseases that COVID-19 patients can potentially develop long after initial infection and suggest that public health scientists and healthcare providers must recognize that long COVID covers a wide spectrum of disorders. Thus, these people should recruit experts from different subspecialties (psychiatry, cardiology, rheumatology, etc) to develop an all-encompassing set of treatment strategies for long COVID because of the possible wide spectrum of diseases represented. This work also demonstrated the feasibility in leveraging existing EHR data that are generated as a byproduct of routine healthcare visits to study long COVID. More work and data are needed to investigate whether or not similar subpopulations of COVID-19 patients exist in other healthcare systems to improve generalizability of results.

To obtain the current set of results, 3-4 iterations of analysis were attempted. In these iterations, I tried different ways of selecting features to use in clustering. At the beginning, I used all available features (1,700+), but the results did not intuitively make any sense. Thus, in later iterations, I used agnostic methods to identify the most informative features to be used in clustering. For example, this involved removing features that are very frequent across all patients. Due to the iterations of analysis involved, there is the potential for data-snooping for this project. However, the concern for data-snooping may be attenuated when considering that useless features were removed in logical, systematic, and unbiased fashions and additionally because clustering on random samples of the dataset yielded largely similar patterns.

Classmate Project Critique

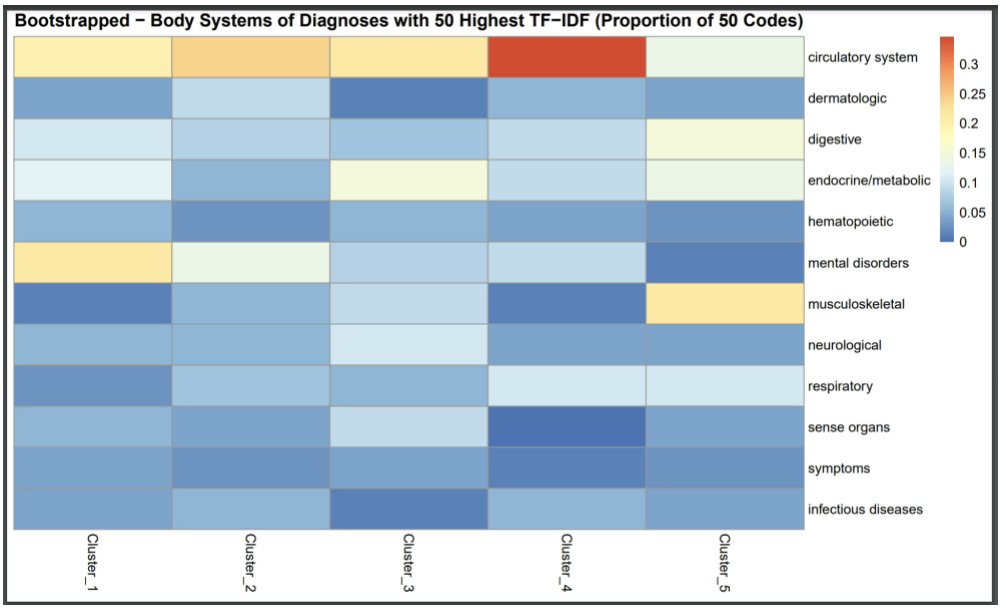
The initial motivation for tackling the project was to identify features that influence the success of start-ups so that early-stage investors can use this information to better evaluate potential investments and make informed investment decisions. This is clearly identified and outlined in the introduction of the project report.

The dataset used in this project was originally obtained from Kaggle, but the source/manager of the data is in actuality Crunchbase. It includes 11 different datasets which cover information ranging from acquisitions, funding rounds, investments, and number of people. The main dataset used in the analysis is the "Objects" data frame, which contains company-level features such as company product, company investment rounds, total funding obtained, and the current status of the company.

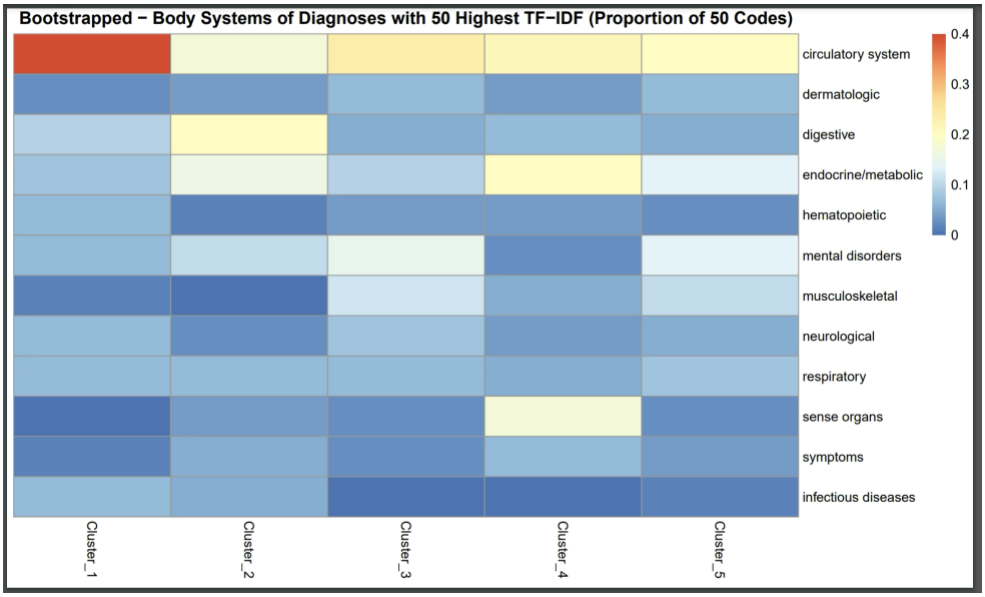
To mine for patterns of features associated with success of a startup, a LASSO penalized logistic regression model was fitted to the data after the feature matrix was scaled and centered. A 10-fold cross validation procedure was run to identify the optimal penalty value used in the penalization. The outcome variable was a success flag of the startup, which was defined by the author as an acquisition or IPO of the company. The outcome was regressed onto the company specific features mentioned in the previous paragraph, and the author was able to identify non-zero effect sizes pertaining to total funding used, the presence of a male CEO, and being in series B funding rounds. It is interesting that the number of funding rounds does not seem to demonstrate a significant effect on the success of the startup company!

I think that the project is well executed with clearly stated aims, a robust analytic approach, and interesting/meaning conclusions that will catch an early-stage investor's eye as they seek to make their next investment decisions. Something that I am interested in better understanding if there is any temporality that may exist in the data. For example, I would imagine that the early-stage investment culture has evolved in different ways over the past two decades, perhaps influenced by the highly visible success of Uber, AirBnB, Snapchat, etc in recent years. To operationalize this concept, one can engineer features for the company's calendar year of founding and the duration it has existed for.

Supplementary Appendix

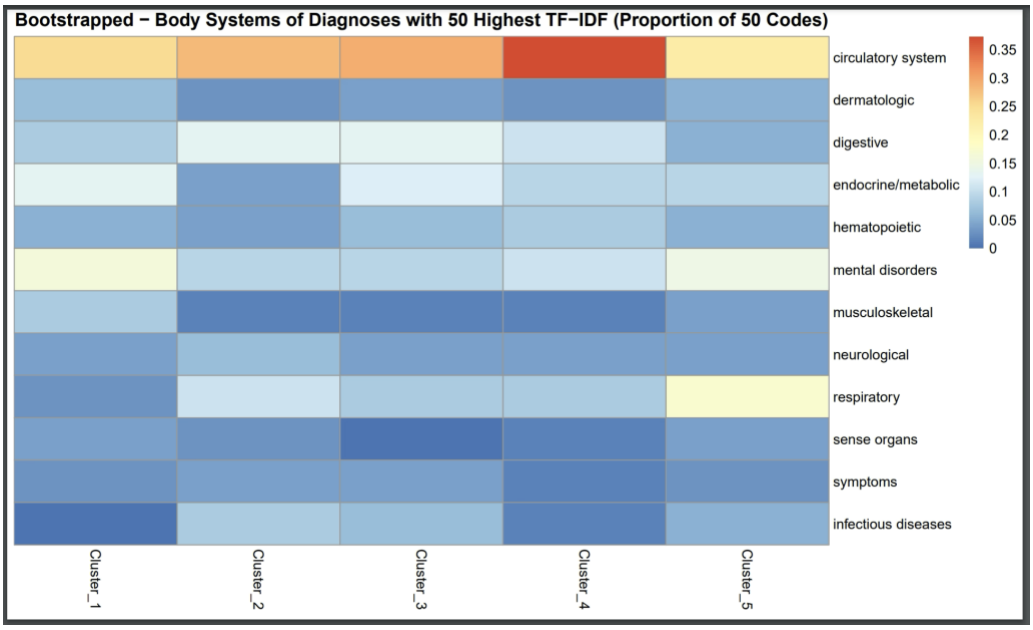


Supplementary Figure 1: Summary of bootstrap resampled patient clusters showing the diseased body systems represented by diagnosis codes with the highest TF-IDF value within each cluster. The legend shows the proportion of the top 50 diagnosis codes with the highest TF-IDF value that belongs to each body system.

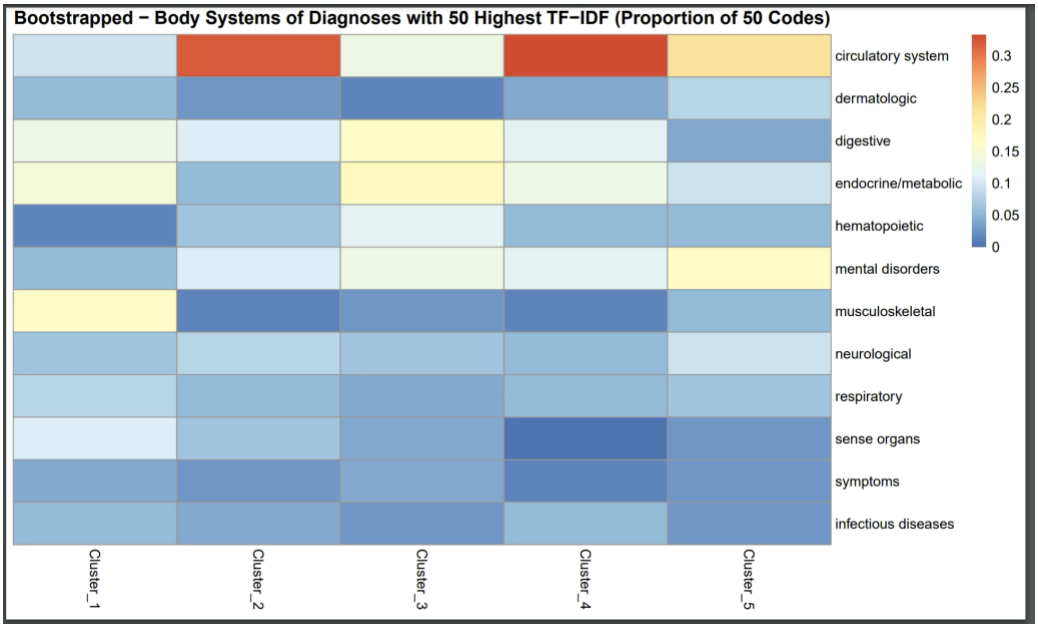


Supplementary Figure 2: Summary of bootstrap resampled patient clusters showing the diseased body systems represented by diagnosis codes with the highest TF-IDF value within each cluster.

The legend shows the proportion of the top 50 diagnosis codes with the highest TF-IDF value that belongs to each body system.



Supplementary Figure 3: Summary of bootstrap resampled patient clusters showing the diseased body systems represented by diagnosis codes with the highest TF-IDF value within each cluster. The legend shows the proportion of the top 50 diagnosis codes with the highest TF-IDF value that belongs to each body system.



Supplementary Figure 4: Summary of bootstrap resampled patient clusters showing the diseased body systems represented by diagnosis codes with the highest TF-IDF value within each cluster. The legend shows the proportion of the top 50 diagnosis codes with the highest TF-IDF value that belongs to each body system.