COMORBIDITY CLUSTERS IN CLINICAL CONDITIONS: AN ANALYSIS OF ELECTRONIC HEALTH RECORD DATA

by
Ting He

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Sciences

Baltimore, Maryland

August 2019

ABSTRACT

Background and Objective: Comorbidity is defined as other conditions present alongside a major condition at the same time. Knowledge of existing comorbidities in study participants may help to guide their disease assessment and management. The objectives of this research were to understand comorbidity patterns and to assess the performance of a range of clustering methods applied to study participant comorbidity profiles in order to stratify study participant's by disease severity.

Methods: We selected study participants who had diabetic retinopathy, glaucoma, or chronic kidney disease from an Electronic MEdical Records & GEnomics dataset (our data source). Then we then created a "gold standard" categorization of study participants into disease severity groups using International Classification of Diseases, Ninth Revision, Clinical Modification codes (i.e., mild, moderate or severe disease). After that, we applied K-means, hierarchical and spectral clustering methods to see how well each performed to classify study participants into the correct severity group, considering two different data subsets. The first data subset considered all "EDC" diagnostic categories and the second data subset only considered selected EDCs that were considered relevant to the conditions.

Results: Our results show that there are no significant differences in the number of comorbidities among different severity levels for diabetic retinopathy (p = 0.8261) and glaucoma (p = 0.5748). However, there was a statistical difference among severity levels for chronic kidney disease (p = 0.0008). Also, we found that for diabetic retinopathy study participants, when using K-means and spectral clustering methods and taking all EDCs disease categories into consideration, it is possible to stratify study participants into three groups based on diagnostic category clustering which corresponds to their severity.

Conclusions: We found a statistical difference in the number of comorbidities present among patients categorized into different severity group for one condition (chronic kidney disease). But there are no significant differences in the number of comorbid conditions among different severity levels for diabetic retinopathy (p = 0.8261) and glaucoma (p = 0.5748). However, there is statistical difference among severity levels for chronic kidney disease (p = 0.0008497). When applying clustering approaches to all EDCs of study participants, we found that, two clustering approaches (K-means and spectral clustering) could be used to classify study participants with diabetic retinopathy into the correct severity group. Clustering approaches were not successful for other scenarios we explored. There were some limitations to this work due to a reliance on administrative data to categorize study participants into severity groups. Findings from this work, however, are promising start to exploring machine learning approaches to identify the severity of disease.

Thesis Advisors: Dr. Casey Overby Taylor Ph.D.; Dr. Jonathan Weiner, DrPH

ACKNOWLEDGMENT

I spent a wonderful 2 years at Hopkins. Thank you to Dr. Casey Overby Taylor for all of her help and guidance during this time. Under your mentorship, I started to understand the process of conducting good research: to think about a topic you love, to study the literature around that topic, to dig in the details, to not be afraid of all the difficulties, to come up with creative ways to overcome those difficulties, to work hard, to understand the process of writing paper, to not be afraid to ask questions in seminar, to not be afraid to talk with other people about your research anytime, to give a pubic talk, to give a poster presentation at international conference, to keep work-life balance, to go for my dreams and so on. Thanks for her hard working and all the giving. I can't express my gratitude enough to you.

Thanks to Dr. Jonathan Weiner's for his advice and help. I attended my first project team meetings with he and Dr.Taylor. They let me know how to cooperate with teammates, how to deal with a different opinion, and how to make sure everything stays on schedule. Also, I couldn't have completed this thesis without your help. Your thoughtful ideas, your clinical sense, and your research experience have helped me to complete this thesis. I appreciate your time to give me advice and to modify the thesis.

Thanks to Dr. Harold P. Lehmann for being available to me all the time. Your kindness and humor make us feel very close to everyone in the Health Sciences Informatics program. I appreciate all our long talks, long emails, and deep thoughts with you. Thanks to Ms. Kersti Winny and Ms. Stacey Szczypinski for all the help during my time here.

TABLE OF CONTENTS

1. INTRODUCTION AND BACKGROUND METHOD

1.1 Objectives

- To understand patterns of comorbidities in a range of clinical conditions.
- To determine which clustering method applied to comorbidity profiles is the most effective in stratifying study participants within a disease category into severity groups.

1.2 Motivation

Comorbidity is defined as a pre-existing medical condition of a study participant, or the presence of one or more medical conditions at the same time.[1] It is associated with higher mortality, increased disability, a decline in functional status and a lower quality of life.[2] In clinical practice, the evaluation of multimorbidity helps to shift physicians away from the old disease-based model to an individual-based perspective.[3] The comorbidities in a cluster may be related through a common etiology or mechanism, shared variance, or a common outcome, which can distinguish different study participants.[4]

The existing approaches to identify comorbidity patterns focus mainly on either using descriptive measures of comorbidity about the prevalence of coexisting conditions or addressing the prevalence of comorbidities based on a particular disease or a specific population.[5] Recently, clustering approaches have been implemented to identify clinically relevant comorbidity patterns and to predict health-related outcomes.[6,7] However, the application of clustering has been limited to specific diseases, populations or analytic approaches.

In this research, we try to address these issues. We aim to identify and describe the patterns of comorbidities among study participant cohorts from different health care settings and to assess the performance of different clustering methods to stratify study participants within a disease category into severity groups. We want to focus on understanding the patterns of comorbidities among study participants selected on the basis of a

single serious dominant condition and also determine which clustering method applied to comorbidity profiles can stratify the study participants by disease severity best.

This thesis contains five major parts. The first topic is introduction and background which concludes the objective of research, the motivation of this job, introduce about dataset and analysis method. The second topic is method. It mentions about the preprocessing of dataset and implementing clustering algorithms. The third part is result section which introduces the population characteristics, the found comorbidity pattern, comorbidity cluster and model evaluation. The following is the discussion part. In this part, it mentions about summary of research, implications limitation and potential future work. The last part is conclusion that will make a summary about all the results got.

1.3 Data and Resources

Our approach leverages the data from the eMERGE Network (Section 1.3.1). We selected study participants with one of three disease according to their ICD-9 codes. A brief description of each selected disease is provided in Sections 1.3.2, 1.3.3., and 1.3.4. We also used the Johns Hopkins ACG software to characterize the comorbidities of selected study participants described in Section 1.3.5.

1.3.1 eMERGE

Electronic MEdical Records & GEnomics (eMERGE) is a national network organized and funded by the National Human Genome Research Institute (NHGRI).[8] EMERGE study participant data sets include data extracted from the electronic medical record (EHR) and also corresponding genome information. The Network includes 11 geographically distinct groups in total.[9]

The eMERGE network has three phases: phase I initiating from March 2007, phase II starting from August 2011, and it is currently undergoing phase III.[9] For phase I, the eMERGE network had goals to use electronic health record (EHR) data for precise phenotyping, conducting genome-wide association studies (GWAS)

based on strict rules of selecting phenotypes, and exploring the ethical, legal and social implications.[10] With phase II, the eMERGE network had one of its major changes, adding several pediatric sites: Children's Hospital of Philadelphia, Cincinnati Children's Hospital, and Boston Children's Hospital.[10] The main focus in phase III is genomic medicine implementation.[11] For example, scholars studied in establish methods for transmit the genetic test results from laboratories to healthcare provider , which can provide better electronic clinical decision support for physician.[12]

The de-identified electronic health record (EHR) phenotype data and diagnostic information used in this research were collected through this network. To answer our research questions, we used de-identified study participant EHR data to select three phenotypes (i.e., diseases) which can be diagnosed as falling into three "severity" levels using ICD-9-CM codes. These included: diabetic retinopathy, glaucoma and chronic kidney disease. The specific ICD 9 codes that represent each different severity level within each condition are outlined below in Table 1.

Table 1: Three Selected Phenotypes with Different Severity Levels

ICD-9-CM Code	Disease & Severity Level
362.04	Mild nonproliferative diabetic retinopathy
362.05	Moderate nonproliferative diabetic retinopathy
362.06	Severe nonproliferative diabetic retinopathy
365.71	Mild glaucoma
365.72	Moderate glaucoma
365.73	Severe glaucoma

585.2	Chronic kidney disease, stage 2 (mild)
585.3	Chronic kidney disease, stage 3 (moderate)
585.4	Chronic kidney disease, stage 4 (severe)

1.3.2 Diabetic Retinopathy

Diabetic retinopathy is a disease that affects blood vessels in the retina. Around 40% to 45% of Americans diagnosed with diabetes have some stage of diabetic retinopathy.[13] Although nearly half of diabetes study participants have this condition, it is hard to detect at the early stages of disease since we might not see the clear symptoms. However, as the condition progresses, its symptoms can include spots or dark strings floating, blurred vision, fluctuating vision, impaired color vision and so on.[14]

1.3.3 Glaucoma

The leading cause of irreversible vision, especially for people over 60, is glaucoma. It affects more than 70 million people worldwide, with approximately 10% of people being bilaterally blind.[15] Early diagnosis can be very challenging because there is no single perfect reference standard for establishing the diagnosis of glaucoma.

1.3.4 Chronic Kidney Disease

Chronic kidney disease will influence people's appetite, sleeping quality and cause cramping, swollen feet, itchy skin, and so on.[16] It includes all the conditions that will damage our kidneys. People can get chronic kidney disease at any age, but when people have diabetes and high blood pressure, they have a much higher probability of developing kidney disease.[17]

1.3.5 Johns Hopkins Adjusted Clinical Groups Software

The Johns Hopkins Adjusted Clinical Groups (ACG) System is a case-mix adjustment system that can be used to categorize all ICD codes that a person may be assigned over a period of time. It measures a study participant's risk by different factors based on collected measures such as medical services, medication prescriptions, and diagnoses.

The ACG system includes various morbidity constructs including Expanded Diagnosis Clusters (EDCs). Tens of thousands of ICD codes are grouped into 282 EDCs representing broad categories representing the most common chronic and acute conditions.[18] For example, in a recent paper, researchers used Expanded Diagnosis Clusters (EDCs) and Rx-defined Morbidity Groups (RxMGs) – based on pharmacy codes - are used as comorbidity indicators to conduct risk adjustment data subsets, which can help health insurance companies to adjust for differences in study participant characteristics when estimating future health care resource cost.[19] Researchers also examined the medication utilization and annual health care costs among study participants with diabetes based on the Johns Hopkins ACG System. They used ACGs to examine and control for comorbidities other than those that are diabetes-related, using the comorbidity index to predict the resources consumed in the next year.[20]

In this research, EDCs, based on the Johns Hopkins ACG system software logic, were used to define comorbidities. We aimed to categorize cases with similar diseases or conditions. We wanted to use the EDCs as a comorbidity profile for each study participant to achieve our research objectives, to gain understanding of the patterns of comorbidities within each of the three selected clinical conditions and to determine which clustering method applied to comorbidity profiles is the most effective to stratify study participants into disease severity groups.

The ACG system assigns all ICD codes found in EHRs or administrative data into one or more of 282 EDCs. The EDCs can be used to categorize cases with similar diseases or conditions. By aggregating many hundreds or even thousands of unique ICD codes into one EDC, it can help to remove the difference in coding behaviors among different physicians.[21]

1.4 Analysis Method

1.4.1 Community Detection Algorithms

Cluster analysis identifies patterns by seeking to partition the dataset into distinct groups based on the similarities among subgroups. Within each group, the similarities value between observations are similar, while observations in different groups are quite different from each other. Cluster analyses has potential to help us to find patterns in the dataset of comorbidities that can distinguish one group from another. While beyond the scope of this thesis, related techniques to community detection algorithms such as principal component analysis[22] and t-SNE[23]may help to visualize high dimensional patterns.

Community detection algorithms include a very broad set of techniques to achieve clustering needs, which have different metrics and may be beneficial for different use cases. There are two broad characteristics of clustering: compactness and connectivity. The K-means method[24] is one example of compactness. The points that lie close to each other will fall in the same cluster and be compact around the cluster center, and the closeness can be measured by the distance between the observations. Spectral clustering[25] is an example of connectivity. Points that are connected or are immediately next to each other are put in the same cluster. For example, even if the distance between two points is small, if they are not connected, they will not be clustered together.[26]-Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters, which can fall into two types bottom-up (each observation starts in its own cluster then pairs of clusters will merge as one and move up the hierarchy) and top-down (all observations start in one general cluster, and splits are performed recursively) approaches.[27] K-means clustering, hierarchical clustering and

spectral clustering are the three methods we used in this research. Each approach is described in more detail in the following sections, including its strengths and limitations.

K-means

In K-means clustering, we seek to partition the observations into a pre-specified number of clusters. It tries to separate samples into n groups of equal variances, minimizing within-cluster sum-of-squares. It scales well to a large number of samples and has been used across a large range of application areas in many different fields.[26]

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (\|x_i - \mu_j\|^2)$$

The K-means algorithm divides a set of N samples (X) into K disjoint clusters (C), each described by the mean (μ_j) of the samples in the clusters. The means here are named the cluster centroids. This method aims to minimize the within-cluster sum-of-squares criteria. However, it suffers from various drawbacks[28]: It assumes that clusters are spherical about the cluster center, which is a strong assumption and may not always be relevant. In very high-dimensional spaces, Euclidean distances tend to become inflated, which will result in the curse of dimensionality problem.

Hierarchical Clustering

In hierarchical clustering, we do not know in advance how many clusters we want. It is a general family of clustering algorithms that builds nested clusters by merging or splitting them successively, and it will end up with a tree-like visual representation of the observations. The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. We used ward hierarchical clustering in this work, which aims to minimize the sum of squared differences within all clusters. It is the same mechanism as with K-means but tackled with an agglomerative hierarchical approach.

This method can also scale to a large number of samples when it is used jointly with a connectivity matrix but is computationally expensive when no connectivity constraints are added between samples because it considers at each step all the possible merges.[28]

Spectral Clustering

Spectral clustering treats the data points as nodes of a graph. The nodes are then mapped to a low-dimensional space that can be easily segregated to form clusters. This method doesn't make a strong assumption on the statistics of the clusters. It can help to get a better result than if one used the K-means method, with the cluster not being spherical about the cluster center.

The disadvantage of this method is that it will increase the time complexity quite a bit since eigenvalues and eigenvectors need to be computed and then we have to do the clustering on these vectors.

Table 2 A, "Comparison of three clustering algorithms" shows a brief comparison among K-means, ward-hierarchical clustering, and spectral clustering.

Table 2: A Comparison of Three Clustering Algorithms [26]

Method Name	Parameters	Scalability	Use case	Geometry (metric used)
K-Means	Number of clusters	Very large number of samples. Medium number of clusters	General-purpose, even cluster size, flat geometry, not too many cluster	Distance between points
Ward hierarchical clustering	Number of clusters or distance threshold	Large number of samples and number of clusters	Many clusters, possibly connectivity constrains	Distances between points
Spectral clustering	Number of clusters	Medium number of samples and small number of clusters	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. neatest-neighbor graph)

2. METHOD

2.1 Preprocessing of Dataset

2.1.1 Inclusion & Exclusion Criteria

We selected our population of interest based on Figure 1 "Inclusion & Exclusion Criteria." The starting eMERGE dataset contained 23,453,405 observations from 81,202 unique study participants. This represents study participants had the records for 289 times on average during the study. In the first stage, we selected the population who had one or more in scope diseases. This narrowed our sample to 123,636 observations, or records. Among this sample, 718 study participants with diabetic retinopathy according to the eMERGE EHR phenotype[29] 983 had glaucoma according to the eMERGE EHR phenotype[30] and 2,699 had chronic kidney disease according to the eMERGE HER phenotype[31] As the next step, we selected the final sample that only included study participants that had ICD codes which could determine their severity levels: mild, moderate, or severe (see Table 1). In total, we included 24,091 observations, each potentially including a range of ICD codes, which met the inclusion requirements. Among this sample, we had 87 individuals with diabetic retinopathy, 329 study participants with glaucoma, and 1,549 study participants with chronic kidney disease. We obtained all ICD codes assigned by the study site, not necessarily limited to the in-scope conditions.

In order to define mutually exclusive groups of study participants, we categorized them according to the following rules to create a gold standard grouping:

- When multiple codes indicating disease severity were recorded for a study participant, we categorized the participant according to the most severe condition
- When more than one of the in-scope conditions existed for a study participant, the most recent diagnosis was used for that study participant.

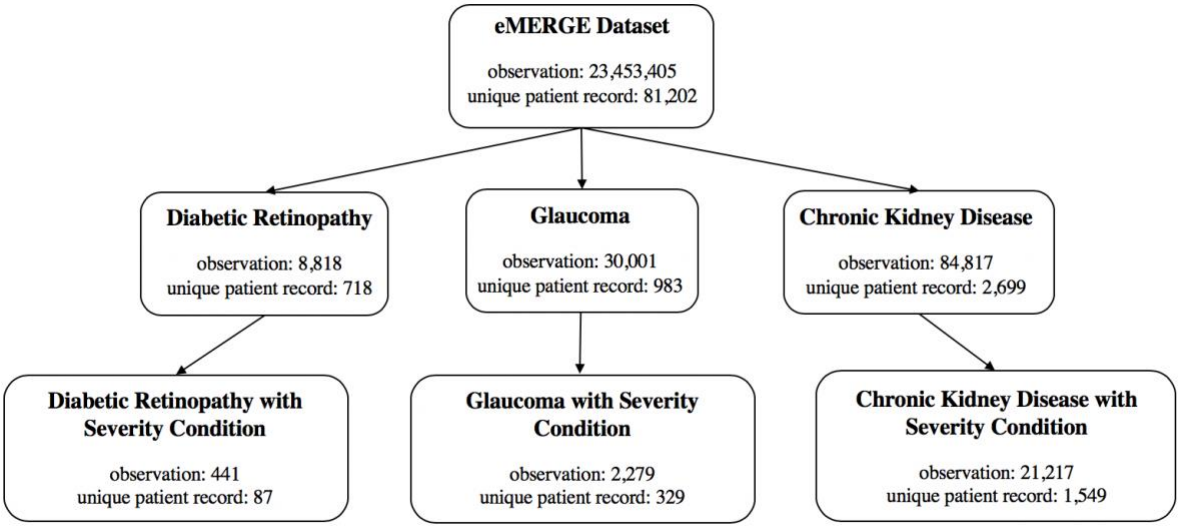

Fig 1. Inclusion & Exclusion Criteria

2.1.2 Gold Standard: Disease Severity Level Grouping

A gold standard grouping of study participants by disease severity level was created according to the rules described in the previous section. In summary, for those study participants with more than two severity levels for one disease, we placed the participant into the more severe condition group. For study participants who had multiple in-scope diseases, we placed the participant into the groups according to their most recently diagnosed disease. Figure 2 represents the results after applying the above rules for diabetic retinopathy study participants. Figure 3 represents outcome for glaucoma study participants and Figure 4 represents chronic kidney disease study participants.

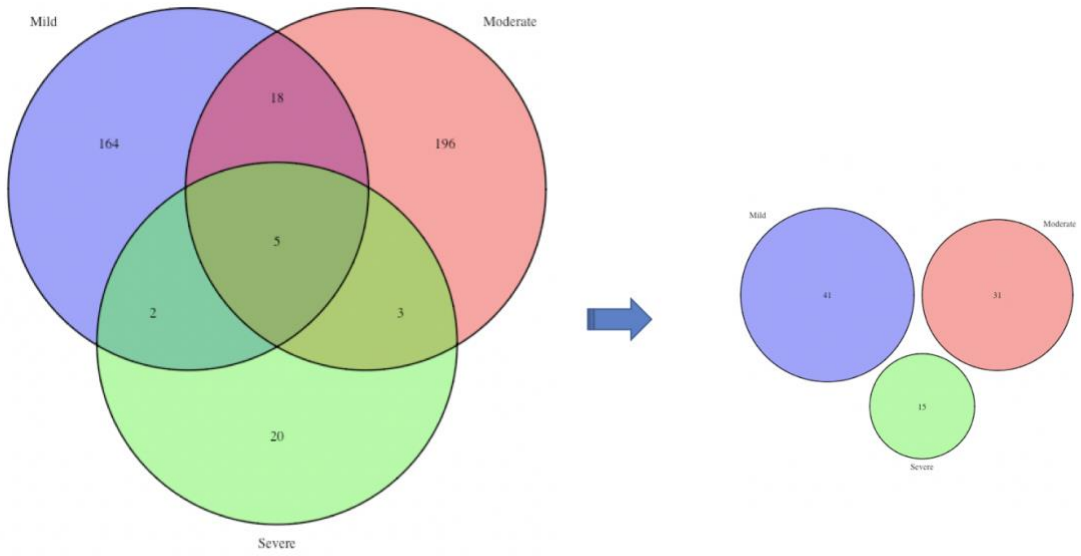

Fig 2. Venn plots for people having diabetic retinopathy with different severity levels

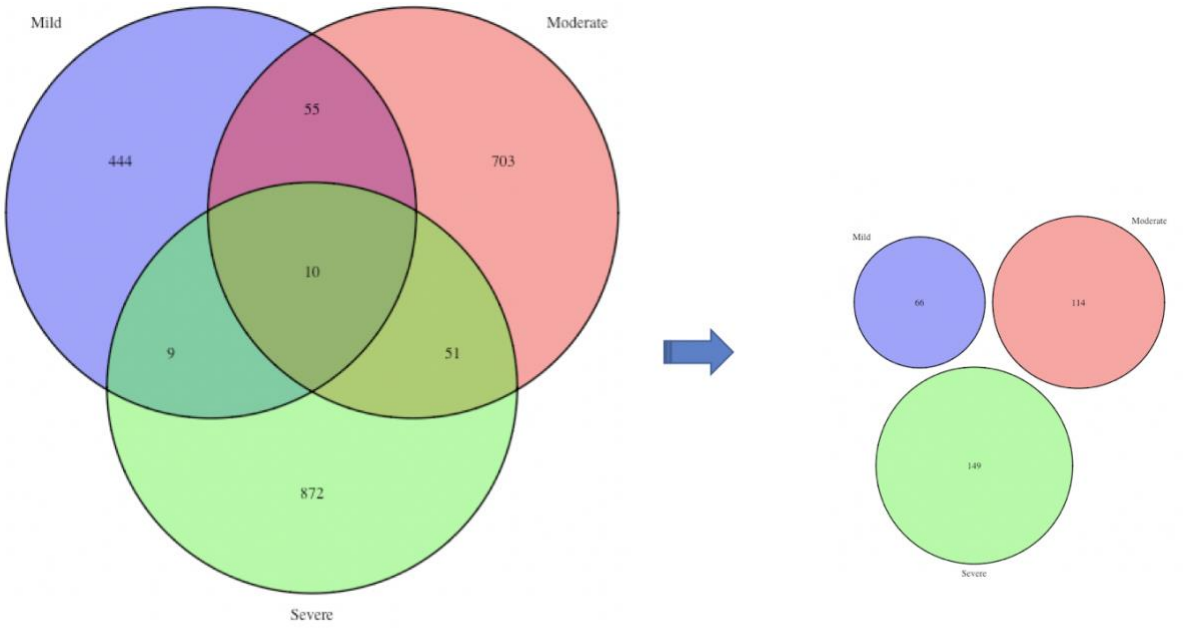

Fig 3. Venn plots for people having glaucoma with different severity levels

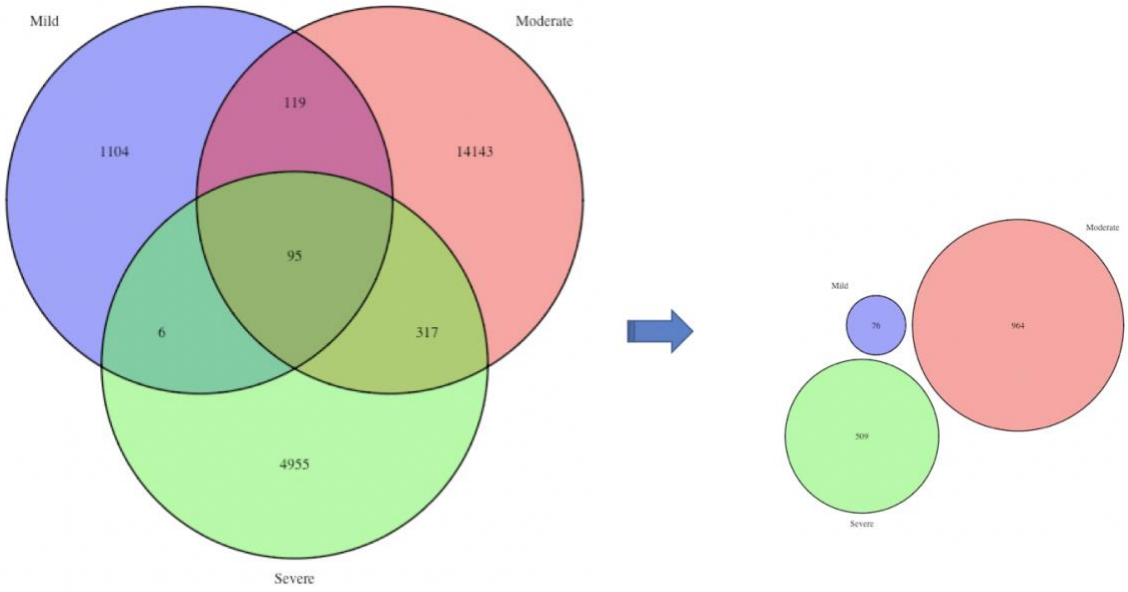

Fig 4. Venn plots for people having chronic kidney disease with different severity levels

2.2 Implement Clustering Algorithms

In the introduction and background sections, we described and compared three different clustering algorithms. In this section, we implement those clustering methods with two data subsets of comorbidities.

2.2.1. Clustering Applied to All EDCs

The initial data subset takes all 282 EDCs into consideration with three different clustering methods. Given that there were many comorbidities that may be irrelevant to the condition, we also considered only the comorbidities that were relevant to the condition (see Section 2.2.2).

2.2.2 Clustering Applied to Relevant EDCs

This data subset only considered the EDCs which are related to the disease of the study participant. Medical background experts (a medical student and Jonathan Weiner, a professor in health policy and management department) helped us to identify the clinically related EDCs for each in scope disease.

Medical experts identified the clinically related EDCs for us. We selected the diseases relevant EDCs from the whole 282 EDCs variable list and took them into our model, and we filtered out the EDC codes specifically for diabetic retinopathy, glaucoma and chronic kidney disease.

1) For diabetic retinopathy, the relevant EDCs were EYE01 ophthalmic signs and symptoms, EYE02 blindness, EYE03 retinal disorders (excluding diabetic retinopathy), EYE04 disorders of the eyelid and lacrimal duct, EYE05 refractive errors, EYE06 cataract, aphasia, EYE07 conjunctivitis, keratitis, EYE09 infections of eyelid, EYE10 foreign body in eye, EYE11 strabismus, amblyopia, EYE12 traumatic injuries of eye, EYE14 eye, other disorders, EYE15 age-related macular degeneration.
2) For glaucoma, the relevant EDCs were GUR02 undescended testes, GUR03 hypospadias, other penile anomalies, GUR04 prostatic hypertrophy, GUR05 stricture of urethra, GUR06 urinary symptoms, GUR07 other male genital disease, GUR08 urinary tract infections, GUR09 renal calculi, GUR10 prostatitis, GUR11 incontinence, GUR12 genitourinary disorders, other.
3) For chronic kidney disease, the relevant EDCs were REN01 chronic renal failure, REN02 fluid/electrotype disturbances, REN03 acute renal failure, REN04 nephritis, nephrosis, REN05 renal disorders, other, REN06 ESRD

2.3 Evaluation by Data Subset

Our objectives were to understand patterns of comorbidities in a range of clinical conditions and to determine which clustering method was the most effective to stratify study participants by the severity of the condition. To understand patterns of comorbidities, we drew tables for the top ten most frequent comorbid conditions for three severity levels and then conducted a three-proportion z test to assess difference in mean among averaged proportions for a least one group. To determine which clustering methods were the most effective, we compared the clustering results with the gold standard disease severity levels (see Section 2.1.2). The best result is that after we used clustering method, it can divide the participants into 3 groups which is corresponding to their severity levels.

3. RESULT

3.1 Sample Characteristics

Based on inclusion and exclusion criteria, we selected our study sample. Table 3 provides the characteristics of the study sample. 87 individuals had diabetic retinopathy according to the eMERGE EHR phenotype and the diseases severity ICD-9-CM code. 329 individuals had glaucoma with a severity ICD-9-CM code, and 1549 individuals had chronic kidney disease with a severity ICD-9-CM code. Eight hospitals (Geisinger, Kaiser Permanente, Marshfield Clinic, Mayo Clinic, Mount Sinai Hospital, Northwestern Memorial Hospital and Vanderbilt University Medical Center) had records in which study participants had one of the three diseases of interest with the severity level. Vanderbilt University Medical Center had the largest population of study participants with diabetic retinopathy and a severity ICD-9-CM code (34%), and of study participants with chronic kidney disease and a severity ICD-9-CM code (34%). The majority of people with glaucoma having a severity ICD-9-CM code came from Marshfield Clinic (48%). For all three diseases, the majority of the study participants were over 50. A moderate number of study participants had diabetic retinopathy or chronic kidney disease between the ages of 20 and 50; however, no one at these ages had glaucoma based in our dataset.

Table 3: Characteristics of the Study Population

Disease	Diabetic retinopathy with severity condition	Glaucoma with severity condition	Chronic kidney disease with severity condition
Gender			
Female	49 (56%)	212 (64%)	737 (48%)
Study Site			
Geisinger	-	-	255 (16%)
Kaiser Permanente	8 (9%)	120 (37%)	55 (4%)
Marshfield Clinic	12 (14%)	157 (48%)	283 (18%)
Mayo Clinic	9 (10%)	4 (1%)	103 (7%)
Mount Sinai Hospital	-	-	253 (16%)
Northwestern Memorial Hospital	28 (32%)	-	66 (4%)
Vanderbilt University Medical Center	30 (35%)	48 (14%)	534 (35%)
Diagnosed Age Group			
20-30	1 (1%)	-	2 (0%)
30-40	2 (2%)	-	11 (1%)

40-50	4 (5%)	-	43 (3%)
50-60	18 (21%)	13 (4%)	172 (11%)
60-70	29 (33%)	40 (12%)	361 (23%)
70-80	18 (21%)	85 (27%)	486 (31%)
80-90	15 (17%)	188 (57%)	474 (31%)

3.2 Assessment of Comorbidities

We filtered out non-disease conditions, such as surgical aftercare, administrative concerns and non-specific laboratory abnormalities, and preventive care. Also, we removed EYE13 diabetic retinopathy while considering study participants with diabetic retinopathy and EYE08 glaucoma when considering patients with glaucoma. Because there was no direct EDC code for chronic kidney disease, no EDCs were excluded for this group. The following table (Table 4) shows the top ten comorbid conditions among study participants in each disease group explored in this work.

Table 4: Top Ten EDC Condition Markers among eMERGE EHR Selected Phenotypes

Diabetic Retinopathy (n=87)	**Glaucoma (n=329)**	**Chronic Kidney Disease (n=1549)**
Eye, other disorders (100%)	Musculoskeletal signs and symptoms (96%)	Hypertension, w/o major complications (99%)
Hypertension, w/o major complications (98%)	Cataract, aphasia (95%)	Disorders of lipid metabolism (96%)
Disorders of lipid metabolism (95%)	Refractive errors (92%)	Musculoskeletal signs and symptoms (89%)
Musculoskeletal signs and symptoms (90%)	Cardiovascular signs and symptoms (88%)	Chronic renal failure (83%)
Cataract, aphasia (86%)	Benign and unspecific neoplasm (84%)	Hypertension, with major complications (83%)
Cardiovascular signs and symptoms (83%)	Low back pain (84%)	Fluid/electrolyte disturbances (83%)
Respiratory signs and symptoms (80%)	Musculoskeletal disorders, other (83%)	Cardiovascular signs and symptoms (82%)
Chest Pain (79%)	Hypertension, w/o major complications (82%)	Iron deficiency, other deficiency anemias (82%)
Peripheral neuropathy, neuritis (77%)	Disorders of lipid metabolism (81%)	Urinary symptoms (76%)
Neurologic disorders, other (76%)	Eye, other disorders (80%)	Chest pain (74%)

We also created tables to compare the difference in comorbid conditions for mild, moderate and severe disease according to ICD-9-CM codes. The results are shown in Tables 5 to 7. Among mild, moderate,

severe levels 4 of the top 10 comorbidities differ for diabetic retinopathy, 4 of the top 10 differ for glaucoma, and 3 of the top 10 differ for chronic kidney disease. We found there are no significant differences in the number of comorbidities among different severity levels for diabetic retinopathy ($p = 0.8261$) and glaucoma ($p = 0.5748$). However, there is statistical difference among severity levels for chronic kidney disease ($p = 0.0008497$). Table 5 shows the detail comparisons for the three diseases at different severity levels.

Table 5: Top Ten EDC Condition Markers for Diabetic Retinopathy among Different Severity Conditions.

Diabetic Retinopathy, Mild Condition (n=41)	**Diabetic Retinopathy, Moderate Condition (n=31)**	**Diabetic Retinopathy, Severe Condition (n=15)**
Hypertension, w/o major complication (100%)	Eye, other disorders (100%)	Disorders of lipid metabolism (100%)
Eye, other disorders (100%)	Hypertension, w/o major complication (94%)	Hypertension, w/o major complication (100%)
Disorders of lipid metabolism (98%)	Disorders of lipid metabolism (90%)	Eye, other disorders (100%)
Musculoskeletal signs and stains (90%)	Cataract, aphasia (87%)	Musculoskeletal signs and stains (100%)
Chest pain (85%)	**Retinal disorders (excluding diabetic retinopathy) (84%)**	Cardiovascular signs and symptoms (93%)
Respiratory signs and symptoms (85%)	Musculoskeletal signs and stains (84%)	Cataract, aphasia (93%)
Cardiovascular signs and symptoms (83%)	**Chest pain (81%)**	**Gastrointestinal signs and symptoms (87%)**
Cataract, aphasia (83%)	**Fever (81%)**	**Neurologic disorders, other (87%)**

Gastrointestinal signs and symptoms (80%)	**Peripheral neuropathy, neuritis (81%)**	**Ischemic heart disease (excluding acute myocardial infarction) (80%)**
Urinary tract infections (80%)	Cardiovascular signs and symptoms (77%)	**Cardiovascular disorders, other (80%)**

***bold text indicates that the comorbidity is unique for the severity level among the top ten comorbidities. For each comorbidity, we show the percentage participants within the severity group that have the comorbidity.**

Table 6: Top Ten EDC Condition Markers for Glaucoma among Different Severity Conditions

Glaucoma, Mild Condition (n=66)	**Glaucoma, Moderate Condition (n=114)**	**Glaucoma, Severe Condition (n=149)**
Musculoskeletal signs and symptoms (100%)	Musculoskeletal signs and symptoms (100%)	Cataract, aphasia (95%)
Cataract, aphasia (94%)	Refractive errors (96%)	Musculoskeletal signs and symptoms (93%)
Cardiovascular signs and symptoms (90%)	Cataract, aphasia (96%)	Refractive errors (88%)
Refractive errors (89%)	Cardiovascular signs and symptoms (89%)	Cardiovascular signs and symptoms (86%)
Musculoskeletal signs and symptoms (89%)	**Disorders of lipid metabolism (87%)**	**Hypertension, w/o major complications (82%)**
Musculoskeletal disorders, other (88%)	Benign and unspecified neoplasm (87%)	**Eye, other disorders (82%)**
Benign and unspecified neoplasm (86%)	**Hypertension, w/o major complications (85%)**	Benign and unspecified neoplasm (82%)

Low back pain (86%)	**Bursitis, synovitis, tenosynovitis (85%)**	Low back pain (82%)
Bursitis, synovitis, tenosynovitis (86%)	Low back pain (84%)	**Musculoskeletal disorders, other (82%)**
Disorders of lipid metabolism (83%)	**Acute sprains and strains (83%)**	**Gastrointestinal signs and symptoms (81%)**

***bold text indicates that the comorbidity is unique for the severity level among the top ten comorbidities. For each comorbidity, we show the percentage participants within the severity group that have the comorbidity.**

Table 7: Top Ten EDC Condition Markers for Chronic Kidney Disease among Different Severity Conditions

Chronic Kidney Disease, Mild Condition (n=76)	**Chronic Kidney Disease, Moderate Condition (n=964)**	**Chronic Kidney Disease, Severe Condition (n=509)**
Hypertension, w/o major complications (100%)	Hypertension, w/o major complications (99%)	Hypertension, w/o major complications (99%)
Disorders of lipid metabolism (96%)	Disorders of lipid metabolism (97%)	Disorders of lipid metabolism (94%)
Cardiac arrhythmia (87%)	Musculoskeletal signs and symptoms (90%)	**Fluid/electrolyte disturbances (94%)**
Respiratory signs and symptoms (83%)	Respiratory signs and symptoms (82%)	Hypertension, with major complications (93%)
Musculoskeletal signs and symptoms (80%)	Cardiovascular signs and symptoms (80%)	Iron deficiency, other deficiency anemias (92%)
Cardiovascular signs and symptoms (79%)	Hypertension, with major complications (78%)	Musculoskeletal signs and symptoms (90%)

Hypertension, with major complications (76%)	**Fluid/electrolyte disturbances (78%)**	Cardiovascular signs and symptoms (87%)
Iron deficiency, other deficiency anemias (76%)	Iron deficiency, other deficiency anemias (77%)	Respiratory signs and symptoms (86%)
Ischemic heart disease (excluding acute myocardial infarction) (75%)	**Urinary tract infections (73%)**	**Renal disorders, other (83%)**
Debility and undue fatigue (75%)	**Appendicitis (72%)**	**Urinary symptoms (81%)**

***bold text indicates that the comorbidity is unique for the severity level among the top ten comorbidities. For each comorbidity, we show the percentage participants within the severity group that have the comorbidity.**

3.3 Performance of Clustering Approaches to Group Study Participants by Disease Severity

In this section, I will describe the results for clustering methods applied to two different data subsets (1) all EDCs (See Section 2.2.1) and (2) relevant EDCs (See Section 2.2.2) after running three different clustering methods: (1) K-means, (2) spectral clustering, and (3) hierarchical clustering. I will also focus on the pattern of clustering methods and the outcome related to the severity conditions.

3.3.1 Clustering Applied to All EDCs

The initial data subset takes all 282 EDCs from the Johns Hopkins ACG software into consideration. We included them as variables in our three different clustering methods (K-means, spectral, and hierarchal clustering). For all the clustering methods, we prespecified that there would be three cluster groups. Tables 8-13 show the number and percentage of study participants from each gold standard severity level (Mild, Moderate, or Severe) that were represented in each of three clusters (1st Cluster Group, 2nd Cluster Group,

3rd Cluster Group). For example, considering the 1st Cluster Group for K-means in Table 1, there were 19 study participants with mild diabetic retinopathy identified in that cluster. A clustering method is considered effective in stratifying patients by severity level if the majority of study participants from each gold standard severity level were in a distinct cluster.

Diabetic Retinopathy:

The K-means method was effective in stratifying patients by severity level. In particular, we found that study participants from each gold standard severity level could be clearly divided into three different groups. In the first cluster, the majority of study participants (46.3%) have a mild condition, in the second cluster the majority of study participants (48.7%) have moderate conditions, and in the third cluster the majority of study participants (40%) have severe conditions. Similar to the K-means method, the spectral clustering method was effective in dividing the study participants from each gold standard severity level into distinct clusters. However, the hierarchical clustering method was unable to correctly group the majority of study participants into distinct clusters.

Table 8: Clustering Applied to All EDCs for Diabetic Retinopathy Study Participants

	Mild (N=41)	Moderate (N=31)	Severe (N=15)
K-means			
1st Cluster Group	**19 (46.3%)**	9 (29.0%)	5 (33.3%)
2nd Cluster Group	10 (24.4%)	10 (32.3%)	**6 (40%)**
3rd Cluster Group	12 (29.3%)	**12 (38.7%)**	4 (26.7%)
Spectral			
1st Cluster Group	**14 (34.1%)**	8 (25.8%)	5 (33.3%)
2nd Cluster Group	13 (31.8%)	10 (32.3%)	**6 (40%)**
3rd Cluster Group	14 (34.1%)	**13 (41.9%)**	4 (26.7%)

Hierarchical			
1st Cluster Group	14 (34.1%)	9 (29.1%)	3 (20%)
2nd Cluster Group	**20 (48.8%)**	**12 (38.7%)**	**11 (73.3%)**
3rd Cluster Group	7 (17.1%)	10 (32.2%)	1 (6.7%)

***Bolded text indicates which cluster has majority study participants from the gold standard severity level group**

Glaucoma:

When it considers the all EDCs for glaucoma study participants, the spectral clustering method and hierarchical clustering method were more effective in stratifying patients by severity than the K-means method. The K-means method did not divide the study participants into their goal standard severity groups, while the spectral clustering method and hierarchical clustering method was able to divide the study participants into two severity groups. When using the spectral clustering method, in the first cluster the majority of study participants have moderate (43.0%) and severe (42.3%) conditions and in the third cluster the majority of the study participants have a mild (34.8%) condition. After applying hierarchical clustering, in the first cluster the majority of study participants have mild (50.0%) and moderate (57.0%) conditions, and in the third cluster the majority of study participants have severe conditions (43.0%).

Table 9: Clustering Applied to All EDCs for Glaucoma Study Participants

	Mild (N=66)	Moderate (N=114)	Severe (N=149)
K-means			
1st Cluster Group	**28 (42.4%)**	**61 (53.5%)**	**72 (48.3%)**
2nd Cluster Group	14 (21.3%)	16 (14.1%)	31 (20.8%)
3rd Cluster Group	24 (36.3%)	37 (32.4%)	46 (30.9%)
Spectral			
1st Cluster Group	20 (30.3%)	**49 (43.0%)**	**63 (42.3%)**

2nd Cluster Group	23 (34.9%)	36 (31.6%)	44 (29.5%)
3rd Cluster Group	**23 (34.8%)**	29 (25.4%)	42 (28.2%)
Hierarchical			
1st Cluster Group	**33 (50.0%)**	**65 (57.0%)**	58 (38.9%)
2nd Cluster Group	7 (10.6%)	8 (7.0%)	27 (18.1%)
3rd Cluster Group	26 (39.4%)	41 (36.0%)	**64 (43.0%)**

***Bolded text indicates which cluster has majority study participants from the gold standard severity level group**

Chronic Kidney Disease:

When considering all EDCs for chronic kidney disease study participants, the K-means and hierarchical clustering methods were able to divide the study participants into two different severity groups, which was more effective than the spectral clustering method was. When considering spectral clustering method, the majority of mild (39.5%), moderate (39.5%) and severe (36.8%) study participants belong to the third cluster. For the K-means method, the majority of study participants with mild (38.1%) and moderate (43.4%) conditions belong to the second cluster, and in the third cluster the majority of study participants (47.6%) have severe conditions. The same situation occurred with the hierarchical clustering method. In the first cluster, most study participants have mild (48.7%) and moderate (46.2%) conditions, and in the third cluster, the majority of study participants have severe (47.5%) conditions.

Table 10: Clustering Applied to All EDCs for Chronic Kidney Disease Study Participants

	Mild (N=76)	Moderate (N=964)	Severe (N=509)
K-means			
1st Cluster Group	25 (32.9%)	314 (32.6%)	94 (18.4%)
2nd Cluster Group	**29 (38.1%)**	**418 (43.4%)**	173 (34.0%)

3rd Cluster Group	22 (29.0%)	232 (24.0%)	**242 (47.6%)**
Spectral			
1st Cluster Group	22 (28.9%)	220 (22.8%)	163 (32.0%)
2nd Cluster Group	24 (31.6%)	364 (37.7%)	159 (31.2%)
3rd Cluster Group	**30 (39.5%)**	**380 (39.5%)**	**187 (36.8%)**
Hierarchical			
1st Cluster Group	**37 (48.7%)**	**445 (46.2%)**	229 (45.0%)
2nd Cluster Group	6 (7.9%)	80 (8.3%)	38 (7.5%)
3rd Cluster Group	33 (43.4%)	439 (45.5%)	**242 (47.5%)**

***Bolded text indicates which cluster has majority study participants from the gold standard severity level group**

3.3.2 Clustering Applied to Relevant EDCs

Diabetic Retinopathy:

The following findings present the conditions of diabetic retinopathy study participants when including only relevant EDCs. The K-means and spectral clustering methods can separate study participants into two severity groups; only the moderate and severe levels cannot be divided. However, the hierarchical method performed much poorer. The third cluster group has the majority of mild (41.5%), moderate (41.9%), and severe (53.3%) conditions.

Table 11: Clustering Applied to Relevant EDCs for Diabetic Retinopathy Study Participants

	Mild (N=41)	Moderate (N=31)	Severe (N=15)
K-means			
1st Cluster Group	**15 (36.6%)**	10 (32.3%)	6 (40.0%)

2nd Cluster Group	11 (26.8%)	8 (25.8%)	3 (20.0%)
3rd Cluster Group	15 (36.6%)	**13 (41.9%)**	**6 (40.0%)**
Spectral			
1st Cluster Group	12 (29.3%)	9 (29.0%)	3 (20.0%)
2nd Cluster Group	**15 (36.6%)**	8 (25.8%)	5 (33.3%)
3rd Cluster Group	14 (34.1%)	**14 (45.2%)**	**7 (46.7%)**
Hierarchical			
1st Cluster Group	10 (24.4%)	6 (19.4%)	3 (20.0%)
2nd Cluster Group	14 (34.1%)	12 (38.7%)	4 (26.7%)
3rd Cluster Group	**17 (41.5%)**	**13 (41.9%)**	**8 (53.3%)**

***Bolded text indicates which cluster has majority study participants from the gold standard severity level group**

Glaucoma:

When including the relevant EDC for glaucoma study participants, only the K-means method can divide the study participants into different severity group. With the spectral clustering method and hierarchical clustering method it is difficult to divide the study participants according to their severity levels. When applying the K-means method, the second cluster group has the majority of mild (45.5%) and severe (43.6%) conditions. In the first cluster, the majority of study participants have moderate (36.0%) conditions.

Table 12: Clustering Applied to Relevant EDCs for Glaucoma Study Participants

	Mild (N=66)	Moderate (N=114)	Severe (N=149)
K-means			
1st Cluster Group	24 (36.4%)	**41 (36.0%)**	54 (36.2%)
2nd Cluster Group	**30 (45.5%)**	41 (36.0%)	**65 (43.6%)**

3rd Cluster Group	12 (18.1%)	32 (28.0%)	30 (20.2%)
Spectral			
1st Cluster Group	17 (25.8%)	37 (32.5%)	47 (31.5%)
2nd Cluster Group	17 (25.8%)	34 (29.8%)	44 (29.5%)
3rd Cluster Group	**32 (48.4%)**	**43 (37.7%)**	**58 (39.0%)**
Hierarchical			
1st Cluster Group	21 (31.8%)	47 (41.2%)	53 (35.6%)
2nd Cluster Group	16 (24.2%)	17 (14.9%)	26 (17.4%)
3rd Cluster Group	**29 (44.0%)**	**50 (43.9%)**	**70 (47.0%)**

***Bolded text indicates which cluster has majority study participants from the gold standard severity level group**

Chronic Kidney Disease:

When including the relevant EDC for chronic kidney disease study participants, the K-means and hierarchal clustering methods did a better job than the spectral clustering method, separating the study participants into two different severity levels. For the K-means method, the second cluster has most study participants displaying mild (52.6%) and moderate (41.7%) conditions; in the first cluster the majority of study participants have severe (36.7%) conditions. When applying the hierarchical clustering method, the second cluster has the majority of study participants displaying mild and severe conditions and the first cluster has the most study participants displaying moderate (48.0%) conditions.

Table 13: Clustering Applied to Relevant EDCs for Chronic Kidney Disease Study Participants

	Mild (N=76)	Moderate (N=964)	Severe (N=509)
K-means			
1st Cluster Group	12 (15.8%)	191 (19.8%)	**187 (36.7%)**

2nd Cluster Group	**40 (52.6%)**	**402 (41.7%)**	226 (44.5%)
3rd Cluster Group	24 (31.6%)	371 (38.5%)	96 (18.8%)
Spectral			
1st Cluster Group	16 (21.1%)	210 (21.8%)	161 (31.6%)
2nd Cluster Group	**40 (52.6%)**	**427 (44.3%)**	**223 (43.8%)**
3rd Cluster Group	20 (26.3%)	327 (33.9%)	125 (24.6%)
Hierarchical			
1st Cluster Group	32 (42.1%)	**463 (48.0%)**	144 (28.3%)
2nd Cluster Group	**37 (48.7%)**	350 (36.3%)	**199 (39.1%)**
3rd Cluster Group	7 (9.2%)	151 (15.7%)	166 (32.6%)

***Bolded text indicates which cluster has majority study participants from the gold standard severity level group**

3.4 Model Evaluation

We compared the clustering results with the gold standard, which aims to see whether the clustering method can divide the study participants into different severity groups. To help quantify this comparison we used purity. Given the gold standard severity level of a condition, purity is defined as the percentage of study participants in a clustering group from all study participants with the severity level. The higher the purity is for a clustering group; the more effective the clustering group is at explaining the severity of a condition. Based on the result from section 3.3, we further evaluate which clustering method can better stratify the study participants into their gold standard severity level groups. Tables 14-19 show the results from this evaluation.

3.4.1 All EDCs Data Subset

For diabetic retinopathy study participants, based on our research question, the K-means and spectral clustering methods were most effective at stratifying study participants into their gold standard severity level groups. To provide more detail, when using k-means, the majority of participants (46%) who had mild diabetic retinopathy were in the 1st cluster group; 39% participants with moderate diabetic retinopathy were in the 3rd cluster group and 40% participants with severe diabetic retinopathy were in the 2nd group.

Diabetic Retinopathy:

Table 14: Comparison of Clustering Approaches Applied to All EDCs for Diabetic Retinopathy Study Participants

Method		Mild	Moderate	Severe
K-means	Clustering Group	1	3	2
	Purity	0.46	0.39	0.4
Spectral	Clustering Group	1	3	2
Clustering	Purity	0.34	0.42	0.4
Hierarchical	Clustering Group	2	2	2
Clustering	Purity	0.49	0.39	0.73

Glaucoma:

For glaucoma study participants, the three clustering methods were not able to stratify study participants into their gold standard severity level groups. However, spectral clustering and hieratical clustering were able to divide the study participants into two severity groups.

Table 15: Comparison of Clustering Approaches Applied to All EDCs for Glaucoma Study Participants

Method		Mild	Moderate	Severe
K-means	Clustering Group	1	1	1
	Purity	0.42	0.56	0.48
Spectral	Clustering Group	3	1	1
Clustering	Purity	0.35	0.43	0.42
Hierarchical	Clustering Group	1	1	3
Clustering	Purity	0.50	0.57	0.43

Chronic Kidney Disease:

For chronic kidney disease, the three clustering methods were unable to stratify study participants into their gold standard severity level groups. However, the K-means and hierarchical clustering methods could divide the study participants into two severity groups.

Table 16: Comparison of Clustering Approaches Applied to All EDCs for Chronic Kidney Disease Study Participants

Method		Mild	Moderate	Severe
K-means	Clustering Group	2	2	3
	Purity	0.38	0.43	0.48
Spectral	Clustering Group	3	3	3
Clustering	Purity	0.39	0.39	0.37
Hierarchical	Clustering Group	1	1	3
Clustering	Purity	0.49	0.46	0.48

3.4.2. Relevant EDCs Data Subset

Diabetic Retinopathy:

For diabetic retinopathy study participants, the three clustering methods didn't stratify study participants well based on severity conditions using the relevant EDCs, since they can only divide them into less than two groups.

Table 17: Comparison of Clustering Approaches Applied to Relevant EDCs for Diabetic Retinopathy Study Participants

Method		Mild	Moderate	Severe
K-means	Clustering Group	1	3	3
	Purity	0.37	0.42	0.40
Spectral	Clustering Group	2	3	3
Clustering	Purity	0.37	0.45	0.47
Hierarchical	Clustering Group	3	3	3
Clustering	Purity	0.41	0.42	0.53

Glaucoma:

For glaucoma study participants, none of these methods were effective to stratify study participants into their gold standard severity level groups when using the relevant EDCs. The K-means method could divide the study participants into two severity groups. However, it could not separate study participants with mild glaucoma from those with severe glaucoma.

Table 18: Comparison of Clustering Approaches Applied to Relevant EDCs for Glaucoma Study Participants

Method		Mild	Moderate	Severe
K-means	Clustering Group	2	1	2
	Purity	0.45	0.36	0.44

Spectral	Clustering Group	3	3	3
Clustering	Purity	0.48	0.38	0.39
Hierarchical	Clustering Group	3	3	3
Clustering	Purity	0.44	0.44	0.47

Chronic Kidney Disease:

For chronic kidney disease study participants, K-means and hierarchal clustering methods performed better than the spectral clustering method. They were able to divide the study participants into two severity level groups.

Table 19: Comparison of Clustering Approaches Applied to Relevant EDCs for Chronic Kidney Disease Study Participants

Method		Mild	Moderate	Severe
K-means	Clustering Group	2	2	1
	Purity	0.53	0.42	0.37
Spectral	Clustering Group	2	2	2
Clustering	Purity	0.53	0.44	0.44
Hierarchical	Clustering Group	2	1	2
Clustering	Purity	0.49	0.48	0.39

4. DISCUSSION

4.1 Summary

This research describes methods and results for understanding patterns of comorbidities in a range of clinical conditions. Also, it achieved the objectives to determine which clustering methods applied to comorbidity profiles are the most effective at stratifying study participants within a disease category into severity groups.

4.2 Implications

The existing approaches to identify comorbidity patterns rely on descriptive measures of comorbidity such as the prevalence of coexisting conditions or the prevalence of comorbidities based on a particular disease or a specific population.[5] We aimed to identify and describe the patterns of comorbidities among study participant cohorts from different health care settings, which can help to decrease population selection bias. Also, we tried to find clustering methods that could be used to stratify participants into severity level groups for a condition (mild, moderate, severe) based on their comorbidities. It describes a first step to apply clustering methods to aid with identifying severity of a disease.

4.3 Limitation

The quality of the data is the main issue in this research. EHR data normally has noise and bias. Because of this, there are limitations to our use of administrative data from EHRs to group study participants in to mild, moderate and severe conditions for those three diseases. The primary purpose of clinical documentation is to help with patient treatment and this presents a fundamental issue to the secondary use those data for research. Physicians document their diagnoses and procedures under their clinical training and standards. They may document quickly and efficiently by giving brief but informative descriptions, which can be a challenge for coders who require more detailed information. For example, a physician may list a glomerular filtration rate to indicate the stage or severity of chronic kidney disease, however, coders may have hard

time to interpret the range without clinical experience. In such case, it may be coded as "chronic kidney disease, unspecified". [32] Our result is highly reliant on the quality of how physicians code the severity of conditions explored in this research.

Second, the sample size for diabetic retinopathy study participants within each severity level group was small, especially for severe diabetic retinopathy. Therefore, the comorbidity patterns for study participants with severe diabetic retinopathy is hard to precisely describe.

Third, the research was limited by the data source used to define severity level. We were confident that the selected study participants had the selected conditions but were less confident able the severity level classifications. Study participants with one of the three selected conditions were previously identified according to eMERGE EHR phenotype definitions. EMERGE researchers used variety of data to define phenotypes including structured and unstructured formatted electronic health records, billing codes, laboratory results, medication data and so on. [33] In order to group study participants by severity level, however, we had access only to ICD-9-CM diagnosis codes. ICD codes are commonly used for reimbursement in hospital.[34] When estimating disease severity levels for study participants, other data may be useful, such as procedure codes. [35] Additional validation through the review of clinical notes by experts could also improve our confidence in the severity level groupings.

4.4 Future Expectation

Genome-wide association studies (GWAS) have been conducted in many research efforts. The GWAS is normally conducted based on one phenotype. However, when we care about more complex diseases, we cannot ignore the existence of other relevant conditions.[36]We want to further this research into genome sides: For those different comorbidity clustering groups, what are the underlying genome differences? And how can we guide study participants into better care? What are the mediation steps you can make based on your

conclusion? It would be valuable to evaluate this K-means approach to other diseases and see what results can be uncovered.

5. CONCLUSION

Findings from this research described comorbidities present for individuals from a range of healthcare institutions with diabetic retinopathy, glaucoma, or chronic kidney disease, and assessed the performance of three common clustering methods applied to comorbidity profiles to stratify individuals by disease severity. In summary, we found no significant differences in the number of comorbidities among different severity levels for individuals with diabetic retinopathy ($p = 0.8261$) or with glaucoma ($p = 0.5748$). However, there was a statistical difference in the number of comorbidities among severity levels for chronic kidney disease ($p = 0.0008497$). Also, we find that for individuals with diabetic retinopathy, K-means and spectral clustering methods that used all EDCs could stratify the study participants into three clustering groups that corresponded to disease severity. Finally, clustering approaches applied to datasets that included all EDCs performed better than those that only included relevant EDCs based on clinical input.

References

1 Term: comorbidity/comorbidities. 2006. Retrieved from http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?printer=Y&definitionID=102446.

2 Schäfer I, von Leitner EC, Schön G, Koller D, Hansen H, Kolonko T, et al. Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions. PloS one. 2010; 5(12): e15941.

3 Marengoni A, Rizzuto D, Wang HX, Winblad B, Fratiglioni L. Patterns of chronic multimorbidity in the elderly population. Journal of the American Geriatrics Society. 2009; 57(2): 225-230.

4 Miaskowski C, Aouizerat BE, Dodd M, Cooper B. Conceptual issues in symptom clusters research and their implications for quality-of-life assessment in study participants with cancer. Journal of the National Cancer Institute Monographs. 2007; 37: 39-46.

5 Ng SK, Holden L, Sun J. Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. Statistics in medicine. 2012; 31(27): 3393-3405.

6 John R, Kerby DS, Hennessy CH. Patterns and impact of comorbidity and multimorbidity among community-resident American Indian elders. Gerontologist. 2003; 43:649–660. DOI: 10.1093/geront/43.5.649.

7 Cornell JE, Pugh JA, Williams Jr JW, Kazis L, Lee AFS, Parchman ML, et al. Multimorbidity clusters: Clustering binary data from multimorbidity clusters: Clustering binary data from a large administrative medical database. Applied Multivariate Research. 2007; 12:163–182.

8 eMERGE Network. Retrieved on October 29, 2018. https://emerge.mc.vanderbilt.edu/.

9 Stanaway, I. B., Hall, T. O., Rosenthal, E. A., Palmer, M., Naranbhai, V., Knevel, R., ... & Linder, J. (2019). The eMERGE genotype set of 83,717 subjects imputed to~ 40 million variants genome wide and association with the herpes zoster medical record phenotype. Genetic epidemiology, 43(1), 63-81.

10 Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., ... & Brilliant, M. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. Genetics in Medicine, 15(10), 761.

11 Fossey R, Kochan D, Winkler E, Pacyna J, Olson J, Thibodeau S, et al. Ethical considerations related to return of results from genomic medicine projects: the eMERGE Network (Phase III) experience. Journal of personalized medicine. 2018; 8(1): 2.

12 Aronson, S., Babb, L., Ames, D., Gibbs, R. A., Venner, E., Connelly, J. J., ... & Liang, W. H. (2018). Empowering genomic medicine by establishing critical sequencing result data flows: the eMERGE example. Journal of the American Medical Informatics Association, 25(10), 1375-1381.

13 Facts About Diabetic Eye Disease. 2015. Retrieved from https://nei.nih.gov/health/diabetic /retinopathy.

14 Diabetic retinopathy. 2018. Retrieved from https://www.mayoclinic.org/diseases-conditions/diabetic-retinopathy/symptoms-causes/syc-20371611.

15 Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. Br J Ophthalmol. 2006;90(3):262–267.

16 Stephens C. Symptoms, causes, and treatment of chronic kidney disease. 2017 Dec [cited 2017 Dec 13]. Retrieved from https://www.medicalnewstoday.com /articles/172179.php.

17 Diabetic kidney disease. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-kidney-disease.

18 Johns Hopkins Bloomberg School of Public Health. The Johns Hopkins ACG® System, Version 11.0. 2019 Jun [cited 2019 Jun 24]. http://acg.jhsph.org.

19 Kan HJ, Kharrazi H, Chang HY, Bodycombe D, Lemke K, Weiner JP. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression data subsets in predicting health care costs in older adults. PloS One. 2019; 14(3): e0213258.

20 Bleich SN, Chang HY, Lau B, Steele K, Clark JM, Richards T, et al. Impact of Bariatric Surgery on Healthcare Utilization and Costs among Study participants with Diabetes. Med Care. 2012; 50(1): 58-65.

21 Concept: Adjusted Clinical Groups – Overview. 2019 June [2019 Jun 20]. http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?printer=Y&conceptID=1304.

22 Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559-572.

23 Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.

24 MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

25 Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. Departmental Papers (CIS), 107.

26 Doshi N. 2019. Retrieved from https://towardsdatascience.com/spectral-clustering-82d3cff3d3b7.

27 Rokach, L., & Maimon, O. (2005). Clustering methods. In Data mining and knowledge discovery handbook (pp. 321-352). Springer, Boston, MA.

28 Clustering. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/clustering.html.

29 Catherine McCarty and Peggy Pessig. Marshfield Clinic Research Foundation. Diabetic Retinopathy. PheKB; 2012 Available from: https://phekb.org/phenotype/11

30 Glaucoma validation of diagnosis. Retrieved from https://phekb.org/implementation/ glaucoma-validation-diagnosis

31 KPWA Chronic Kidney Disease. https://phekb.org/implementation/kpwa-chronic-kidney-disease

32 Lucyk, K., Tang, K., & Quan, H. (2017). Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. BMC health services research, 17(1), 766.

33Phenotyping: cohort discovery using EHR data. Retrieved from https://emerge.mc.vanderbilt.edu/phenotyping-cohort-discovery-using-ehr-data/

34 National Committee on Vital Health Statistics Web site. Retrieved from http:// www.ncvhs.hhs.gov/031105a1.htm.

35 Utter, G. H., Miller, P. R., Mowery, N. T., Tominaga, G. T., Gunter, O., Osler, T. M., ... & Brown, C. V. (2015). ICD-9-CM and ICD-10-CM mapping of the AAST Emergency General Surgery disease severity grading systems: conceptual approach, limitations, and recommendations for the future. Journal of Trauma and Acute Care Surgery, 78(5), 1059-1065.

36 Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009;10(4):241-51.

TING HE

Baltimore, Maryland

443 467 4493 the14@jhmi.edu

EDUCATION

Johns Hopkins University (JHU) School of Medicine July 2017 - Present
Master of Science, Health Science Informatics GPA: 3.8/4.0
Course: Statistical Machine Learning, Biostatistics, Survival Analysis, Longitudinal Analysis, Statisti cal Computing, HIT Standards and Interoperability

JHU Bloomberg School of Public Health Jan 2018 - Present
Graduate Certificate, Health Economics GPA: 3.0/4.0
Course: Clinical Decision Support, Economics Evaluation, Health Economics

Georgia Institute of Technology College of Computing Jan 2019 - Present
Master of Science, Online Computer Science program

University of Calgary School of Medicine Jun 2016 - Dec 2016
Visiting Student, Bioinformatics

Dalian University of Technology School of Life and Sciences Jun 2013 - Aug 2017
Bachelor of Science, Bioinformatics GPA: 3.7/4.0
Course: Bioinformatics, Data Mining, Calculus, Linear Algebra, Probability and Statistics, statistics, Perl, Bio-Java, Database, Molecular Biology, Genetics

Computer Languages	R, Python, SAS (certificate), SQL, Shell, HTML
Data Science	Clustering and Classification, Modeling, Machine Learning
Health IT	EHR, Claim, Oncology (ICD10,UMLS,etc)

Johns Hopkins Division of Health Science Informatics July 2017 - Present
Graduate Researcher

Goal: To stratify the risk of various comorbidities in eMERGE cohort by understanding the comorbidity cluster

Identifying and describing the patterns of comorbidities in a range of clinical conditions

Defining the gold standards to determine the efficiency of different clustering methods which can best stratify the severity conditions

Identifying symptom clusters in women experiencing pre-term birth

Goal: Identify symptom pairs to help guide symptom assessment and management for pregnant women

Performed quantitative and qualitative methods to clean data, analyze data and visualize data set to explore the psychological symptom clusters in 2836 pregnant women by R

Presented a poster at 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics

Comorbidity Characterization Among eMERGE Institutions: A Pilot Evaluation of the Johns Hopkins ACG System

Goal: Using the Johns Hopkins ACG system to characterize comorbidities to discover new genetic association

Conducted literature review to identify the current clustering methods using in co morbidity field
Published a paper in AMIA and helped with background part

Customized Recommendations for Healthcare Management about Incidental Findings

Goal: Leveraging the Disease Ontology to identify condition-specific phenotypes for incidental findings
Worked on 6 Disease Ontology Knowledge database to identify the incidental ndings

Created a simple Incidental Finding Notification Dashboard by Python Flasks to present our research results
Advisor: Casey Overby Taylor, Ph.D. Assistant professor of Health Sciences Informatics

PROJECT EXPERIENCE

Breast Cancer Data Search July 2016 - Aug 2016
Undergraduate Executor

Built Breast Cancer Related Mutation Query web application based on Java, HTML, and MySQL
Used Decision Tree model to predict unknown classes of patients through WEKA machine learning tool

Health Doctor - Android App July 2015 - Sep 2015
Undergraduate Team Leader

Led a 6-student team to develop app providing general health improvement recommendations based on user medical data

Presented results and development plans in a poster presentation for provincial competition and won 1st Prize

PUBLICATION & PRESENTATION

Ongoing Master Thesis Comorbidity Clusters in Clinical Conditions: An Analysis of Electronic Health Record Data

Publication Taylor CO, Lemke KW, Richards TM, Roe KD, He T. Comorbidity characterization among eMERGE institutions: a pilot evaluation of the Johns Hopkins ACG system. 2019 AMIA Informatics

Poster Presentation He T, Taylor CO. (Sept, 2018) Identify symptom cluster in pregnant women experiencing preterm birth. 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics

WORK EXPERIENCE

Johns Hopkins Division of Health Sciences Informatics Sep 2018 - Dec 2018 Graduate Teaching Assistant

Master Program Core Course: Introduction to Biomedical and Public Health Informatics; Informatics and Clinical Research Lifecycle: Tools, Techniques ,and Processes

Synyi AI Dec 2018 - Jan 2019 Winter Intern Medical data governance, mining Startup in Shanghai, China

Served in Bioinformatics team conducting 5 Joint SNV analysis a GWAS study to compare the detection powers in different methods and performed Elastics Net to find relationships between individual SNV and disease

Conducted research, generated a report about healthcare data governance market and presented it to C-level