2022

Developing Artificial Intelligence tools to investigate the phenotypes and correlates of Chronic Kidney Disease patients in West Virginia

Marzieh Amiri Shahbazi
WVU, ma00030@mix.wvu.edu

Developing Artificial Intelligence tools to investigate the phenotypes and correlates of Chronic Kidney Disease patients in West Virginia

Marzieh Amiri Shahbazi

A thesis submitted to the
Statler College of Engineering and Mineral Resources at
West Virginia University
In partial fulfillment of the requirements
for the degree of

Master of Science
in
Industrial Engineering

Imtiaz Ahmed, Ph.D., Chair
Mohammad Abdullah Al-Mamun, Ph.D.
Ashish Nimbarte, Ph.D.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia
November, 2022

Keywords: Chronic Kidney Disease(CKD); Acute Kidney Injury(AKI); Clustering; Machine Learning

ABSTRACT

Developing Artificial Intelligence tools to investigate the phenotypes and correlates of Chronic Kidney Disease patients in West Virginia

Marzieh Amiri Shahbazi

Chronic kidney disease (CKD) is responsible for disrupting the lives of 37 million people just in the USA, which is about 1 in 7 adults. CKD results in a gradual loss of kidney function over time. Sometimes CKD doesn't produce any significant symptoms until it reaches an advanced stage. On the other hand, acute kidney injury (AKI) accounts for a sudden decline in the kidney's function. As a result, the kidneys fail to filter waste materials from the blood and cause an increase in blood pressure. High blood pressure can cause heart disease and, in the long-term, induce CKD. Literature to date says AKI leads to long-term adverse kidney outcomes and linked to CKD. AKI diagnosis, its severity, treatment, and recovery process have a major impact on the likelihood of a future diagnosis of CKD. This research attempts to understand the patient's trajectory toward developing CKD after AKI diagnosis, key triggers contributing to this trajectory and ultimately develop an Artificial intelligence-based prognosis tool. To comprehend the role of AKI and previous hospitalization in the progress of CKD, various cohorts of CKD patients are created: i) AKI after hospitalization before CKD ii) Random AKI before CKD, and iii) No AKI before CKD. Prior comorbidities, medications, lab results, and pertinent procedures are considered, and for each cohort of patients, the most prevalent phenotypes are identified. The patient cohorts required for this analysis are generated from CKD patients residing in West Virginia. The data is provided by TriNetx, a global network platform. K-means clustering and the latent class analysis (LCA) approach is used to identify and group the phenotypes of CKD for each cohort. The high-risk patient groups generated by the clustering algorithms are compared with each other. These results will help clinicians to understand the risk factors of CKD and the overall trajectory of the development of CKD. This research suggests that a single method of care does not work for all patients since phenotypes vary for distinct groups of patients and categorizing patients into distinct groups allows for the allocation of different resources and strategies for the care of different groups of patients. From this research, it is evident that patients' risk profiles change over the years before developing CKD. There are also similarities as well as differences across the cohorts for each year, which suggests that CKD risk factors may be linked to prior AKI, hospitalization, or inpatient care.

Contents

Copyright i

ABSTRACT ii

1 Introduction 1

1.1 Overview . . . 1
1.2 Motivation . . . 2
1.3 Objectives of the research . . . 4
1.4 Research questions . . . 4

2 Literature Review 5

2.1 Research domain . . . 5
2.2 AKI, phenotypes, diagnosis and prediction . . . 5
2.3 CKD, phenotypes, diagnosis and prediction . . . 5
2.4 AKI to CKD trajectory . . . 6
2.5 Methodology review . . . 6
2.6 Knowledge gap . . . 8

3 Research Design and Methodology 10

3.1 Research dataset . . . 10
3.2 Study design and setting . . . 10
3.3 Data preparation . . . 10
3.3.1 Patients with CKD . . . 11
3.3.2 Patients with and without AKI . . . 11
3.3.3 Patients with AKI prior to three years of CKD . . . 12
3.3.4 In-patients hospitalized . . . 12
3.3.5 Medical procedures . . . 13
3.4 Study cohorts . . . 13
3.5 Phenotypes profiling . . . 14
3.6 Creating binary files . . . 16
3.7 Adjusting for missing values . . . 17
3.8 Dividing cohorts by year . . . 18

4 Data Mining and Machine Learning Methods 19
4.1 Clustering . . . 19
4.1.1 Dimensional reduction and variable selection . . . 20
4.1.2 Logistic PCA . . . 20
4.1.3 K-means clustering . . . 21
4.1.4 Latent class analysis (LCA) . . . 22
4.1.5 Random forest . . . 23
4.2 Outcome measures . . . 24
4.3 Similarity measures . . . 25

5 Results 26
5.1 Patient population . . . 26
5.2 Logistic PCA and K-means clustering results . . . 29
5.3 Latent class analysis (LCA) and variable selection results . . . 34
5.4 Comparison of LCA clusters with K-Means clusters . . . 41
5.5 Cohort comparison and insights . . . 44
5.5.1 Diagnosis . . . 46
5.5.2 Procedures . . . 50
5.5.3 Medications . . . 53
5.5.4 Lab results . . . 56
5.5.5 Vital Signs . . . 59

6 Discussion and Conclusions 61
6.1 Discussion . . . 61
6.2 Strengths of the study . . . 62
6.3 Limitation of the study . . . 62
6.4 Contributions . . . 62
6.5 Future works . . . 62
6.6 Conclusions . . . 63

A 68

B 70

C **72**

D **74**

E **80**

F **82**

List of Figures

1.1 Long-term consequences after AKI[7] 3

3.1 Flowchart of inclusion for the final study population 15

4.1 Logistic PCA steps 21

5.1 The number of principal components for cohort 1. (a) year 1; (b) year 2; (c) year 3 30

5.2 The Elbow Method showing the optimal number of clusters (k). (a) year 1; (b) year 2; (c) year 3 31

5.3 Clusters of K-means for cohort 1. (a) year 1; (b) year 2; (c) year 3. 33

5.4 Variable prevalence in different clusters based on the LCA method for cohort 1. (a) year 1; (b) year 2; (c) year 3 36

5.5 Variable prevalence in different clusters based on the LCA method for cohort 2. (a) year 1; (b) year 2; (c) year 3 37

5.6 Variable prevalence in different clusters based on the LCA method for cohort 3. (a) year 1; (b) year 2; (c) year 3 38

5.7 Comparison of K-means clustering with latent class analysis (LCA) in year 1 cohort 1 41

5.8 Comparison of K-means clustering with latent class analysis (LCA) in year 2 cohort 1 42

5.9 Comparison of K-means clustering with latent class analysis (LCA) in year 3 cohort 1 43

5.10 Jaccard similarity index for cohort 1. 45

A.1 The number of principal components for cohort 2. (a) year 1; (b) year 2; (c) year 3 68

A.2 The number of principal components for cohort 3. (a) year 1; (b) year 2; (c) year 3 69

B.1 The elbow method to find the number of clusters for cohort 2. (a) year 1; (b) year 2; (c) year 3 70

B.2 The elbow method to find the number of clusters for cohort 3. (a) year 1; (b) year 2; (c) year 3 71

C.1 Clusters of K-means for cohort 2. (a) year 1; (b) year 2; (c) year 3 72
C.2 Clusters of K-means for cohort 3. (a) year 1; (b) year 2; (c) year 3. 73

D.1 Comparison of K-means clustering with latent class analysis(LCA) in year 1 cohort 2 74
D.2 Comparison of K-means clustering with latent class analysis(LCA) in year 2 cohort 2 75
D.3 Comparison of K-means clustering with latent class analysis(LCA) in year 3 cohort 2 76
D.4 Comparison of K-means clustering with latent class analysis(LCA) in year 1 cohort 3 77
D.5 Comparison of K-means clustering with latent class analysis(LCA) in year 2 cohort 3 78
D.6 Comparison of K-means clustering with latent class analysis(LCA) in year 3 cohort 3 79

E.1 Jaccard similarity index for cohort 2. (a) year 1; (b) year 2; (c) year 3 80
E.2 Jaccard similarity index for cohort 3. (a) year 1; (b) year 2; (c) year 3 81

List of Tables

3.1 Data details 11

5.1 Clinical characteristic in all cohorts 28
5.2 Principal Components 29
5.3 Number of clusters selected for K-means clustering by the elbow method 29
5.4 Demographic information based on the clusters for cohort 1 . 40
5.5 Demographic information based on the clusters for cohort 2 . 40
5.6 Demographic information based on the clusters for cohort 3 . 40
5.7 Year 1 diagnosis comparison 48
5.8 Cohort comparison for diagnosis 49
5.9 Year 1 procedure comparison 52
5.10 Cohort comparison for procedure 53
5.11 Year 1 medication comparison 55
5.12 Cohort comparison for medications 56
5.13 Year 1 lab comparison 58
5.14 Cohort comparison for lab 59
5.15 Year 1 vital sign comparison 60
5.16 Cohort comparison for vital signs 60

F.1 Year 2 diagnosis comparison 82
F.2 Year 3 diagnosis comparison 83
F.3 Year 2 procedure comparison 84
F.4 Year 3 procedure comparison 85
F.5 Year 2 medication comparison 86
F.6 Year 3 medication comparison 87
F.7 Year 2 lab comparison 88
F.8 Year 3 lab comparison 89
F.9 Year 2 vital sign comparison 90
F.10 Year 3 vital sign comparison 90

1 Introduction

1.1 Overview

The primary function of the kidneys is to remove excess water and waste from blood to produce urine. The kidneys balance the salts and minerals that circulate in the blood, such as calcium, phosphorus, sodium, and potassium, to keep the body functioning properly. Additionally, the kidneys produce hormones that maintain bone density, create red blood cells, and aid in blood pressure regulation [1].

Acute Kidney Injury (AKI) also known as acute kidney failure or renal failure, occurs when the kidney suddenly fails to perform its function and doesn't filter waste from the blood[2]. Consequently, waste products in the blood increase dramatically and cause high blood pressure and injure the heart. It can rapidly develop over a few days or weeks. To recover the kidney from the injury, it is extremely important to identify the abnormal function of the kidney in time . However, there exists a lag in diagnosing the injury which makes the recovery process more difficult[3].

AKI is common in adults over 60 years of age who have been hospitalized, especially for critical illnesses or intensive care[4, 5]. Nearly 25% of hospitalized patients develop AKI after being admitted. Low- and middle-income countries are where this disease is more prevalent.[6].

There are several reasons that could result in AKI. One of the major reasons is decreased blood flow. Many diseases and conditions can reduce the blood flow in the kidney such as low blood pressure (also known as hypo-tension) or shock, blood or fluid loss such as in bleeding and severe diarrhea, heart attack, heart failure, and other conditions that affect the heart's ability to pump blood, overuse of some medicines (used to treat headaches, the flu, colds, and other conditions by reducing swelling or reducing discomfort), significant allergic responses, burns, an accident, major surgery, etc[1]. The second main factor is kidney damage brought on by direct trauma, such as a sort of severe, hazardous infection, a form of cancer, blood vessel inflammation, or an allergic reaction to specific kinds of medications. An obstruction in the urinary system is the third cause.[4].

The symptoms of AKI include decreased urine output, swelling in the legs, ankles, and around the eyes, difficulty breathing, weakness, nausea, confusion, and exhaustion, abnormal heartbeat, chest pressure, or seizures and coma in severe cases. However, in some cases, there are no symptoms at all and it's diagnosed through lab tests. AKI diagnosis is based on changes in serum creatinine and urine output(an increase in serum creatinine (SCr) or a decrease in urine output (UO))[7–9].Tests such as blood urea nitrogen (BUN), fractional excretion of sodium, and urine microscopy could be carried out to help clinicians determine the severity and diagnose possible underlying cause of the AKI[10].

AKI has both short-term and long-term outcomes depending on the recovery success of initial AKI . Initial AKI can be followed by a number of different recovery patterns, such as

sustained AKI reversal, recurrence after full or partial recovery from the first episode, and AKI without reversal. Depending on the patterns of healing, long-term outcomes varied greatly amongst patients. Those with AKI, who relapsed or never recovered had longer hospital stays and increased mortality in the nine months after the original AKI occurrence than patients who fully recovered[6]. The patient may have comorbidities like diabetes, hypertension, chronic kidney disease (CKD), cardiovascular disease (CVD), liver disease, lung disease, sepsis and surgery and exposure to nephrotoxic drugs[7]. Heart failure and sepsis are the most important potential adverse outcomes of AKI because the sudden dysfunction of the kidney dramatically increases the blood pressure, and high blood pressure causes heart failure[11].

One of the long-term outcomes of AKI is chronic kidney disease (CKD)[12]. CKD is a gradual loss of kidney functions. Kidneys with chronic renal disease are harmed and unable to filter blood as they should. Because the kidney damage occurs gradually over a long period of time, the condition is referred to as ”chronic”. Wastes may accumulate in the body as a result of this harm. Other health issues can also result from CKD[**4**].

Over time, renal disease frequently worsens and might result in kidney failure. Dialysis or a kidney transplant are required for CKD patients to maintain their health after the kidney failure.

Patients with CKD may not exhibit many symptoms or indicators in the early stages. It's possible that they won't be diagnosed until it's progressed to an advanced stage. They can take action to save their kidneys as soon as they become aware that they have renal disease. Other health issues, like heart disease, can also develop as a result of kidney disease. A heart attack is more likely to occur in persons with kidney dysfunction. High blood pressure can work as both the cause and the effect of renal disease. Kidney injury makes it hard to maintain a healthy blood pressure. Furthermore, people with CKD are more prone to develop a sudden change in kidney function as a result of a disease, accident, or medicine. So, AKI can lead to CKD and CKD patients are more vulnerable to future AKI[7].

The diagram below depicts how AKI can progress to CKD and then back to AKI in the long run.
Research has revealed that patients who fully recover from early AKI following surgical treatment have a comparable risk of CKD and long-term mortality to those who do not, hence this study will concentrate on CKD as a long-term outcome of AKI[13].

1.2 Motivation

Rising number of CKD patients in recent years poses significant burden to the healthcare systems in USA (CKD patients increased by 52.6% between 2002 and 2016[**2**]). It results in damaged kidney over time and contributes to many other significant diseases such as high blood pressure, heart disease, stroke, etc [14]. CKD is a silent disease and almost 9 out of 10 people do not know that they have CKD [**3**]. Given the significance of the kidneys to the body and the fact that CKD gradually impairs kidney function while occasionally exhibiting

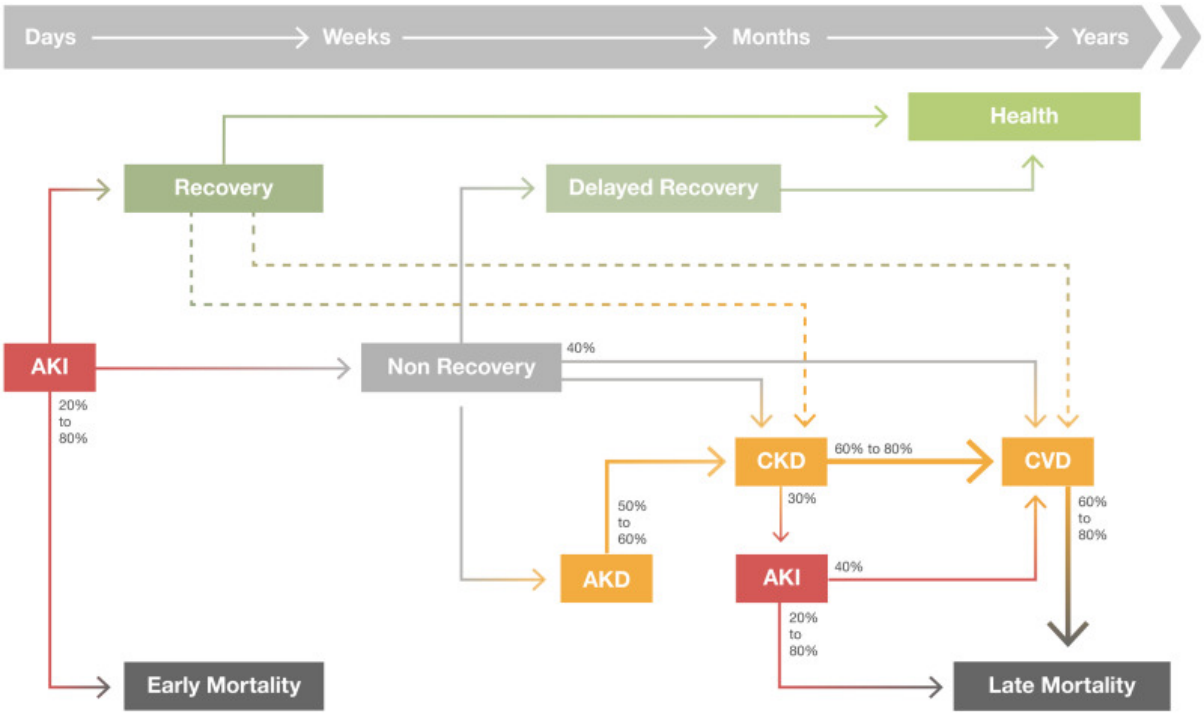


Figure 1.1: Long-term consequences after AKI[7]

.

no symptoms, it is crucial to identify high-risk patients at an early stage. AKI increases the likelihood of developing CKD and historically shows an association with CKD [15]. So, an intervention just after the AKI occurrence can possibly mitigate the risk of CKD in long term. To take action in this phase, a deeper comprehension of the AKI to CKD transition is required. Common comorbidity increases the risk of CKD and makes it desirable to identify. Due to the potential for an insidious progression of the disease, CKD is typically disregarded in the clinic until it is severe, and this is another reason for the importance of CKD prediction.

The rising number of CKD patients, its related direct and indirect expenditures, as well as the rising costs resulting from its severity, are what motivate our research to better understand how patient diagnosed with CKD after having a prior AKI. The cost of CKD has an impact on society, caregivers, and patients. Identification, management, and treatment of CKD are all included in the cost of care for CKD patients. Managing comorbidities including diabetes, hypertension, and heart disease is also a part of treating CKD, as they all increase the cost of treatment. Patients' and carers' productivity is impacted. Patients' productivity falls due to absences and overall inefficiency at work, whereas carers' productivity rises as a result of investing more time and effort in their charge. The time and money spent by patients, carers, employers, and healthcare professionals also have an impact on society.[16].The CKD cost for non-dialysis patients is around $15,000 per patient for a one-year care period.[17]. Therefore, the increasing frequency and progression of chronic kidney disease (CKD) raises questions about how to efficiently manage the financial burden it places on patients, caregivers, and society. The societal costs of CKD and end-stage renal disease are substantial and increases as the disease progresses. [16]. All of these impacts inspired

us to investigate the AKI to CKD trajectory and understand the factors responsible for CKD. Once the patients groups who are more vulnerable to CKD and their distinct characteristics are identified, specific intervention strategies can be designed to reduce the risk of long term CKD and associated complications.

1.3 Objectives of the research

The objectives of the research are as follows:

i. To develop different CKD cohorts for understanding the patient's trajectory from the onset of AKI toward CKD diagnosis.

ii. To develop clustering algorithms for creating several distinct patient groups with similar characteristics across cohorts.

iii. To compare among cohorts across years to find out the similarities and dissimilarities in terms of important clinical variables.

1.4 Research questions

The overarching theme of this research is to understand patient trajectory between AKI and CKD as its long-term outcome. To understand the patient trajectory the thesis attempts to answer the following research questions:

a) Is CKD more prevalent in patients with prior hospital induced AKI and compared to the patients with no AKI?

b) Are there any differences between hospital induced AKI injury compared to the random AKI injuries (who didn't have an induced condition for the AKI prior to CKD and hospitalization) in long term CKD development?

c) What are the roles of prior comorbidities, medications, lab results and relevant procedures in developing CKD?

d) What are the most prevalent phenotypes in different CKD patient groups?

e) Can we find similarities and dissimilarities in the phenotypes over a time period prior CKD diagnosis?

2 Literature Review

The remainder of this thesis is organized as follows. Section 2 highlights the relevant works in the literature. Section 3 describes the dataset, study design, and preparation of cohorts. Section 4 illustrates the data mining methods and machine learning approaches that have been utilized in this research. Section 5 discusses the results and conclusions.

2.1 Research domain

This review has been carried out using the papers available on PubMed Central (PMC), which is a free full-text archive of biomedical and life sciences journal articles at the the United States National Institute of Health's National Library of Medicine(NIH/NLM). Google Scholar is also used as an additional web search tool. The main emphasis is provided on the recent works that discussed the possible association between AKI and CKD and their respective phenotypes, and diagnosis. We also included papers where machine learning (ML) methods and data mining algorithms are proposed to predict AKI risk variables and outcomes.

2.2 AKI, phenotypes, diagnosis and prediction

AKI occurs when the kidney function suddenly decreases. It will be followed by an increase in serum creatinine or a decrease in urine output[18]. It is one of the major reasons for patients' mortality[19]. Identifying the phenotypes for AKI helps to identify the disease early and reduce the threat of AKI. The phenotypes for AKI can be divided into two groups. One is chronic phenotypes, which includes chronic kidney disease, chronic liver disease, congestive heart failure, hypertension, and atherosclerotic coronary vascular disease which all have the same risk. The second one is acute risks, including low pH, nephrotoxin exposure, severe infection/sepsis, mechanical ventilation, and anemia, where the first two poses the highest level of risk, followed by the second two acute factors, and finally anemia has the lowest risk among the acute phenotypes[19]. Different ML methods were used to predict AKI including logistic regression[19, 20] and k-nearest neighbors[18] that had the high accuracy in prediction.

2.3 CKD, phenotypes, diagnosis and prediction

CKD results in a steady decline of the normal kidney activity and, as a result, destroys kidney functionality in the long run. When the kidney fails completely, dialysis is needed for the survival. This is a health burden. Around 1.4 million CKD patients in the world needs to replace their renal. Apart from being painful, it is extremely expensive [21, 22]. It is diagnosed at the advanced stages either by decline in estimated glomerular filtration rate (eGFR) and/or the presence of significant albuminuria [23]. Individual's genetics, age, race, gender, and

family history are important phenotypes for CKD, also, smoking, obesity, hypertension, and diabetes mellitus can also lead to CKD[22].

The most crucial variables for predicting the start and progression of early CKD in people with diabetes were albuminuria and eGFR. The performance of prediction is only slightly enhanced by the addition of demographic, clinical, and other laboratory factors[24]. Different methods have been used to predict CKD, like linear regression, generalized additive models, Gaussian process regression, regression trees, and k-nearest neighbor for continuous outcomes. For binary outcomes, logistic regression, classification tree, support vector machine, and k-nearest neighbor have been used[25].

2.4 AKI to CKD trajectory

Studies about CKD followed by AKI show that there is a substantial link between AKI and CKD[7, 10, 26]. Patients with AKI have a 41-fold increased chance of developing CKD than patients without the condition, and patients with prior AKI have a 28-fold higher chance of developing advanced CKD than those without prior AKI[26]. The severity of the AKI before the CKD also influences how severe it is thereafter. More severe CKD will occur after serious AKI. Moreover, experimental trail for AKI models confirms the statistical correlation and provides insights on the molecular mechanisms by which AKI contributes to future CKD[27].

On sometimes, AKI cannot make a suitable or acceptable adjustment to the situation[10]. As a result, in this situation, CKD will worsen over time. Age, and comorbidities, among other clinical variables, can also affect how CKD develops. If people who are more likely to undergo late CKD progression can be identified early, further research is still needed to answer this question. An important research goal is still to develop new treatments to reduce the risk of CKD after AKI.
Effectiveness of model risk predictions following AKI, shows that 1 in 7 patients with AKI have its side effects, and 1 in 3 AKI patients don't fully recover before being released from the hospital[28]. However, more study is needed to determine which AKI outcomes are more likely to occur, when they are more likely to happen, and what may be done to minimize serious consequences in high-risk individuals. There is a gap for post-AKI patient follow-up that, if filled, will reduce mortality, readmission, CKD, cardiovascular, and other AKI outcomes.

2.5 Methodology review

One of the key methods in data mining is clustering. In order to uncover the structure of the data and hidden groups among the patients also, it can be helpful to categorize individuals or patients based on their similarity in phenotypes. In general, the first stage in clustering entails dividing up related data into groups that are similar to one another. The second phase involves giving each group a name based on the similarities among classes and differences

with other groups. As a result, care can be delivered in the various facets of the healthcare industry more effectively and efficiently.

Data mining from clinical and diagnostic data is utilized in healthcare to come up with early detection in this field.[29]. The vector quantization method, as one of the clustering methods, has been used in healthcare to predict re-admissions in intensive medicine. The algorithms used in the vector quantization method are k-means, k-medoids, and the K-means obtained the best results. [30].

Clustering has been used in data mining even in an area where it didn't work much due to the presence of substantially skewed distribution like budget data. One of the area that clustering in data mining had been used, is to identify costs changing in the end stage of kidney disease who initiated hemodialysis. Clustering methods like, k-means CA and hierarchical CA, can be used in this area and seems they work good. It is important to note that the k-means CA method works better for data with highly skewed[31].

Latent class analysis(LCA) and k-means clustering analysis were employed, as two clustering techniques to find complex patient profiles among people who take several medications. K-means and latent class analysis produced identical groupings based on the quantitative findings. It was discovered that complex patient profiles can be divided into various categories, which aids the healthcare system in managing treatment and allocating resources.[32].

Hierarchical cluster analysis (HCA) is another method for clustering that had been used for complex patterns of concurrent medication use In order to categorize pharmaceutical exposure [33]. HCA had been used to determine which strategy is most appropriate for a certain sort of data, some studies compared the HCA method with methods from past studies that used k-means and latent class analysis. The k-means method is suitable for quantitative variables; however, it is only applicable to quantitative variables with interval scales, and it is restricted in its applicability when dose volume information is missing because it is not suitable for categorical variables. This approach employs a conventional vector of distance toward the center. Latent class analysis, which is more flexible than the k-means method is also more suited for variables of mixed scale types[34]. However, it needs data reduction when we have a high number of variables. Some research points out that HCA is the best clustering method for identifying high-risk exposure patterns. That is because HCA allows flexible data reduction; it is good for categorical variables and longitudinal data; it can be used for any type of variable; and it allows for the use of a custom distance metric. After HCA, the optimal number of clusters needs to be identified, which was based on the dendrogram and algorithms that automatically find the optimal number of clusters. Research shows that the definition of risk can be improved by uncovering actual patterns of prescription use and can be used to detect risks related to various exposure profiles for other drugs, and past analysis has shown[35].

Numerous hierarchical clustering techniques exist. Every one has a distinct definition. They discover that the points fit into each cluster in a unique fashion, although everything

works well when clustering various kinds of data[36].

Determining the group of individuals with various chronic diseases who receive the major portion of Medicare beneficiaries was another area where cluster-based techniques had been used. Researches showed that one-third of adults who have three or more chronic conditions are responsible for two-thirds of healthcare expenditures in the USA [37]. Agglomerative hierarchical clustering as one of CA methods had been used and discovered that there is within-group pattern between patients with multiple chronic disease. Patients with heart failure and kidney disease makes up the majority of the group that spent almost one-fourth of the total budget. Agglomerate hierarchical clustering approach is a preferred method for two reasons. First, it doesn't need a pre-assigned number of clusters. and the second one is that clusters are mutually exclusive. This method starts with one single cluster, then they divide it into smaller clusters with the most similarity. This nesting cluster helps the number of clusters chosen by the statistical criteria at each step [38].

In general, clustering methods in healthcare can be divided into three methods: a) Partition based clustering which employs K-means and K-medoid to group n data points into K classes. b) Hierarchical Clustering. This method can be classified as Agglomerative and Divisive and finally c) Density based clustering method also play a very important role in biomedical research because they are capable of handle any cluster of arbitrary shape [39].

2.6 Knowledge gap

Mostly previous literature selected their variables by medical selection only. However this research has been carried out to utilize artificial intelligence tools for variable selection and analyzing CKD risks in patient groups after AKI besides the medical selection.
We chose to use machine learning (ML) methods over more traditional techniques because they are faster and more accurate when applied to large data sets.Additionally, it can be updated by passing time. On the other hand, ML can identify hidden patterns in the dataset which traditional statistical methods cannot.
Also, we defined different cohorts of patients to consider the effects of multiple criteria on AKI prior to CKD at the same time. For instance, one cohort has been designed to see whether any patients who develop AKI after hospitalization can get CKD or not. In general, AKI-related hospitalization has been considered. Another cohort considered AKI not related to hospitalization to find out how other factors can affect CKD. The last cohort is a control cohort. This cohort controls the effects of AKI on CKD. Each cohort has been explained in more detail at Part 3.4.
In general, it is important to understand how patients in cohorts with previous AKI are different from patients in another cohort without prior AKI in terms of different health characteristics.Comprehensive clustering-based patient profiling which is a novel method Implications had been used in this research to help clinicians to understand the transition from AKI to CKD better and how AKI groups are different from the control group. Also, it will help to

understand early the patient trajectory or profiles from AKI to CKD health status.
Some other long term outcomes for AKI like cardiovascular disease has been considered and investigated for its association with AKI but not CKD. As CKD is considered to be one of the most distressful diseases for a patient and a major financial burden for the government, identifying high-risk patients for CKD and then monitoring them to detect CKD at an early stage is critically important and that is what this study is targeting to achieve.

3 Research Design and Methodology

Chapter Contents

1.1 Overview 1
1.2 Motivation 2
1.3 Objectives of the research 4
1.4 Research questions 4

3.1 Research dataset

Data for this study was provided by TriNetX, a global network platform that has been used by healthcare institutions practically everywhere in the world to gather historical patient data and make use of this information collected from over 29 countries. TriNetX retrieves data from the electronic health record (EHR), verifies its accuracy, and saves it in accordance with a common format. This platform categorizes the data as belonging to the patient's demographics, diagnosis, procedures, measurement of the lab results, prescribed drugs, and vital signs. Patients with chronic kidney disease (CKD) in West Virginia, USA, were the subject of this study. A total of 75033 CKD patients of various disease stages participated in the study. 7696 of them experienced AKI in the three years prior to CKD, and all of them had hospitalization/admission visits in the 90 days prior to AKI.

3.2 Study design and setting

This study considered the possibility of CKD after a maximum of five years following AKI. This is due to the presumption that the patients' features would significantly alter after three to five years. We also focus on the patients who were hospitalized in the 90 days before AKI, if they had AKI prior to CKD, to find a relationship between AKI and hospitalization too.

3.3 Data preparation

Data preparation is the most important part and the first step of data analysis. It accounts for nearly 80 percent of the problems solved in this area. The goal of this part is to gather relevant data to make sure analytic tools provide useful information and practical insights for identifying phenotypes and finding hidden patterns in clinical data related to AKI and CKD. In this section, we used several filtering, merging, and categorizing tools to find and analyze the important variables and their effects on the study.

Dataset		CKD patients in West Virginia
Sex		
	Male	37073(%49.42)
	Female	37937(%0.50.58)
Age		
	>60(years)	57581(%84)
	<60(years)	11244(%16)
BMI (range/mean)		0.18-50/30.8(kg/m2)
Race		
	Asian	104(%0.14)
	Black or African American	2596(%3.46)
	Native Hawaiian or Other Pacific Islander	38(%0.05)
	White	63512(%84.65)
	Unknown	8782(%11.7)

Table 3.1: Data details

3.3.1 Patients with CKD

To find out the start date when CKD was first diagnosed for a patient, we filtered data on the diagnosis file based on the codes which start with "N18". This code is used for CKD patients in the ICD-10-CM code system. The Diagnosis table contains the "patient id," a unique id for each patient; the "encounter id," a unique id for each encounter; and The "code system," in which the diagnosis is coded, encompasses ICD-9-CM and ICD-10-CM. The column "Code" is used to determine the type or name of the diagnosis, and "date" is the date the diagnosis was recorded, which we used as the CKD start date. There are several duplicates for each unique patient. This file keeps all the diagnoses that are identified for an individual patient, which includes different levels of CKD, AKI at different times, and the comorbidity available for each patient. Due to these reasons, we face several duplicates in this file. The results from this part are gathered in two different files. One is for CKD patients with different encounter ids at different dates. That means a patient was admitted multiple times with the ICD-10 code of CKD. The second file includes a unique CKD patient ID with the date of the first occurrence of the CKD.

3.3.2 Patients with and without AKI

Patients with CKD are at first divided into two groups: those who had AKI prior to CKD and those who did not have AKI prior to CKD. To find the first cohort, we find all AKI patients from the diagnosis file based on their codes that were saved by N17 for patients. This code is based on the ICD-10 system code. In this cohort, where all AKI patients are included, the first date or the smallest date is kept as the index date for the AKI (this is the date a patient

identified for AKI), and all other duplicates are dropped from this file.
Then among these patients, those who got AKI "prior" to CKD are considered in our study. So, based on the index date for CKD (first occurrence of CKD) and the index date for AKI (first occurrence of AKI), those whose index date for AKI is before the CKD diagnosis date are filtered and kept as the AKI prior CKD cohort. To consider the second cohort, who are patients that didn't have AKI before CKD, we segregate the patients whom were at CKD cohort but were not diagnosed for AKI prior CKD. This cohort is named as NO AKI prior CKD. So now we have three cohorts: 1) AKI after hospitalization prior CKD, 2)AKI at rondom prior CKD, and 3)No AKI prior CKD

3.3.3 Patients with AKI prior to three years of CKD

Cohort one is limited to patients who had AKI at most "three years" prior to CKD. We chose this timeline based on our literature review of the average time of CKD diagnosis after AKI. AKI in CKD patients prior to three years was filtered based on the time between the index dates and divided into two cohorts. If the time between the date diagnosed as CKD patients and AKI patients is zero, we removed them, as they considered prevalent cases of AKI. We created the cohort with patients who had time difference of less or equal to three years between CKD and AKI diagnosis dates.

3.3.4 In-patients hospitalized

We also need to know the patients who were hospitalized before being diagnosed with AKI, as it is an important factor for developing hospital induced AKI and later leading to CKD. Thus we must identify the hospitalized patients first.

Any interaction between a patient and a healthcare provider(s) for the purpose of delivering healthcare services or determining a patient's health condition is referred to as an encounter. Therefore, an encounter file must be created for each patient admitted as an inpatient to a hospital. There are many encounter categories, such as ambulatory (AMB), emergency (EMER), home health (HH), inpatient encounter (IMP), inpatient non-acute (NONAC), observation (OBSENC), pre-admission (PRENC), short stay (SS), and virtual (VR), according to the TriNetX data dictionary. These values are based on the ActEncounterCode value set from HL7 version 3.

In-patients are classified in the encounter file as having the type 'IMP' based on the TriNerX data dictionary. Because our focus is on patients who were hospitalized 90 days before developing AKI, we kept patients who had both AKI and CKD within three years of each other and who had also been hospitalized within 90 days of developing AKI. To find patients who had been hospitalized, we filtered patients with the type code 'IMP'.

We have five types of CKD patients. The first group includes patients whose diagnosis time between AKI-CKD is zero; the second group, whose diagnosis time between AKI-CKD is less than three years; and the third group, whose time between AKI-CKD is more than three

years but less than five. The fourth group, which is beyond our study, is patients with a time between AKI-CKD of more than five years. The last group are patients who never had AKI before CKD.

In this research, we separately identified hospitalized patients for each of these groups to consider phenotypes for patients in each group.

3.3.5 Medical procedures

In the medical field, a ”procedure” refers to the hierarchical actions doctors take to improve a patient’s condition. It includes figuring out, evaluating, or diagnosing the parameters or condition of the patient; treating, resolving, or restoring function, for instance, through surgery and physical therapy. Additionally, each activity and service that patients can receive from a healthcare practitioner is given a CPT code, which is a number and saved for each patient’s encounter in a procedure file.

In this study, admission visits that demonstrate the effectiveness of physicians should also be taken into account for each patient, in addition to hospitalization, which demonstrates the efficacy of healthcare providers and systems. As a result, we identified every patient who had a CPT code in the procedure file. These codes cover all ”inpatient visits” with: hospital inpatient initial care; hospital inpatient subsequent care; hospital observation initial care; hospital observation subsequent care; hospital observation per hour; hospital inpatient initial consult care; hospital inpatient initial care (between 8hrs to 24hrs) and discharge on the same day; nursing facility inpatient initial care. It also covers ”outpatient visits” with: outpatient established office care; outpatient new to office care; outpatient consult care. ”Critical care”: Critical care and ”Primary care centers”: Standard Office Visits, New Patient (comprehensive), New Patient (extensive), Established Patient, brief, Established patient (moderate), Established patient (in-depth), Established patient (extensive).

3.4 Study cohorts

We are considering three cohorts of the patients in this study.

a) **Cohort 1**: AKI after hospitalization/inpatient visits (Those patients who had AKI within 90 days of any inpatient services), first occurrence of AKI had been chosen, those patients who had AKI and CKD on the same date discarded in this group (time difference is 0), those patients who developed AKI after 90 days of inpatient visit discarded and saved in as a separate file. Three years is chosen for this cohort, as this group is with people who have AKI after hospitalization, and we hypothesize that this impact can be measured up to 3 years.
   In this cohort, we want to see whether any patients develop AKI after hospital service, can get CKD or not, and in general, AKI related to hospitalization.

b) **Cohort 2**: AKI at random ,including AKI prior to 3 years of CKD and hospitalized beyond 90 days prior AKI and AKI prior 5 years of CKD who are not in cohort 1.
In this cohort, we want to consider AKI when it isn't related to hospitalization. and to see how other things can affect CKD.

c) **Cohort 3**: Non-AKI-patients in other word, AKI never occurred for these CKD patients. This cohort is a control cohort, means that patients in this group did not have the disease of interests (without AKI).

The goal is to compare these cohorts for different criteria such as phenotypes for patients in each group in order to determine the impact of AKI, hospitalization, or admission visit on the progression of AKI and CKD.
Figure 3.1 shows the details of each cohort and the inclusions and exclusions of the study.

3.5 Phenotypes profiling

Following are the five categories of factors that are most widely viewed as significant factors for CKD and AKI in the literature:

i Diagnosis

ii Procedures

iii Lab results

iv Vitals

v Medications.

The phenotypes from each group, as listed above, considered, and the arbitrarily number of variables selected from each. According to what we've seen in the literature, we looked at the top 50 comorbidities, roughly 100 lab results and vitals, 100 procedures, and roughly 200 medications, which is how this type of data is normally provided. Therefore, the number of codes changes based on the type of data we are analyzing. These figures were chosen because they can more easily be translated into percentages and accurately reflect the bulk of the population.

In this research, we choose the top 50 comorbidities based on both ICD-9 and ICD-10 codes. In between these 50 comorbidities, we have 16 groups which include similar codes. The 16 groups are: hypertensive diseases, disorders of lipoprotein metabolism and other lipidemias, Type 2 diabetes mellitus, diseases of the blood and blood-forming organs, atrial fibrillation and flutter, chronic obstructive pulmonary disease, Heart failure, overweight and obesity, abnormalities of breathing, Osteoarthritis , Hypothyroidism, Chronic ischemic heart disease, Anxiety disorders, Pleural effusion, Gastro-esophageal reflux disease.

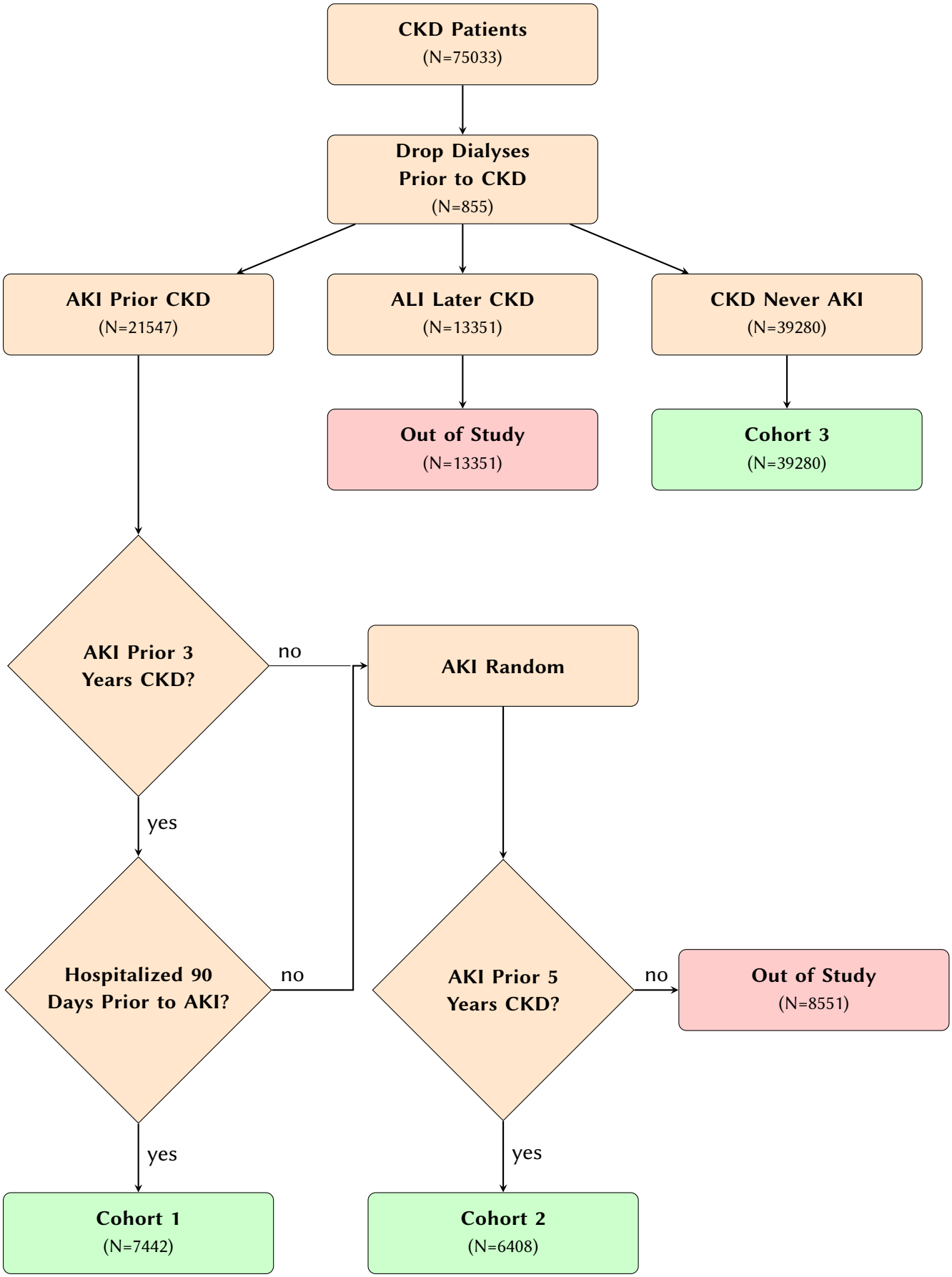


Figure 3.1: Flowchart of inclusion for the final study population

From procedure we found top 100 codes which are more frequently present in the file.

For the lab results we select a threshold to separate the positive and negative patients. We find all codes in the groups of: Urea nitrogen [mass/volume] in serum, plasma, or blood in the range of 6.0-24.0, Alkaline phosphatase [enzymatic activity/volume] in Serum, Plasma, or Blood in the range of 44-147, Protein concentration (mass/volume) in serum or plasma in range of 6-8, Folate [mass/volume] in Serum, Plasma, or Blood in range of 6-10, Cobalamin (Vitamin B12) [mass/volume] in Serum, Plasma, or Blood in range of 200-900, Hepatitis B virus surface Ab [Units/volume] in Serum for patients with value more than 12 for positive hep,Hemoglobin A1c/Hemoglobin.total in Blood with value more than 6.5 for positive diabetes, White blood cell /Leukocytes [/volume] in Blood with value between 4.5 - 11, Bilirubin.total [Mass/volume] in Serum, Plasma or Blood value between 0.1 - 1.2, and for vital includes Heart rate by value between 60-80, body temperature(36.4-37.2 Celsius), body height, body weight, BMI(18.5-24.9), blood pressure, systolic(120) and blood pressure, diastolic(80).

Finally for the medication, we found top 500 medication codes for CKD patients.

Other variables that were considered during this study were: heart rate, body temperature, body height, body weight, BMI, blood pressure(systolic) and blood pressure(diastolic) which are from vital file.

3.6 Creating binary files

Final step in data preparation is generating binary variables for individual patients in each cohort.

For each patient, we have around 1322 variables for cohort 1 and 1910 variables for cohorts 2 and 3, including: for each year, we considered 50 binary variables for the diagnosis; 100 binary variables for the procedures; 33 binary variables for the lab results and vital signs; and 500 binary variables for the medications. For diagnosis and procedure variables, we took into account whether each variable was positive (meaning a patient was diagnosed for any commodity or had any procedure) in the first year following AKI, the second year following AKI, or the third year following AKI. Consequently, for cohort 1, we end up with 150 variables for the diagnosis (50 for year 1, 50 for year 2, and 50 for year 3), as well as 300 for the procedure (100 for each year).

We have 33 variable for the lab results and vitals. which considered every three months for three years in cohort 1 and five years For cohorts 2 and 3, As a result, for cohort 1, there are 132 variables each year and 396 variables overall throughout the period of three years, and 660 overall variables for cohorts 2 and 3.

For medication, we simply looked to see if a patient had taken any of the top 500 medications prior to CKD, three years for cohort 1, and five years for cohorts 2 and 3.
A binary row was produced for each patient. For example, if a patient in each cohort had a diagnosis in the first year, second year, or third year, encountered any procedures in the first year, second year, or third year, had abnormal results for lab tests in the first three months,

second three months, and so on, or had taken any medication, we would insert 1 for the intersection of the variable and patient, and 0 otherwise.

3.7 Adjusting for missing values

We were facing too many missing values in each cohort. They can be divided into three groups:

1. **Missing value in some periods**
   Patients in each cohort had been considered over several periods of time for each class of variables. For instance, diagnosis and procedure variables are considered in three periods for cohort 1 and five periods for cohorts 2 and 3, or lab results and vitals are considered in 12 periods for cohort 1 (each three months during three years) and 20 periods for cohorts 2 and 3, and medications are considered in one period of 3 years for cohort 1 and 5 years for cohorts 2 and 3.

   Before altering missing values for patients in this group, we considered to see if they were related to labs and vitals. We examined whether the majority of the patient's records that were available were normal or abnormal. We filled in the missing values with 1 as abnormal when patients were found to have a majority of abnormal lab results or vital signs, and with 0 as normal when the majority of a patient's records were discovered to be normal. In all other cases, we set zero in place of any missing values for other classes of variables.

   For lab and vitals variables, there are three possible scenarios: first, patients who did not test(NaN) ; second, patients who tested and received normal results(0) ; and third, patients who tested and received abnormal results(1). However, for other variables, say medication, patients took a medicine(1) or not(1), or had a comorbidity(1) or not(1). Because of this, we only evaluated NaN in lab results and vital signs; otherwise, we interpreted missing values as zero.

2. **Missing values in some class of variables**
   In general, we are considering on five groups of variables (diagnosis, procedures, lab results, vitals, and medications). Some patients' records for certain classes are missing. For instance, some of them are not noted in procedures, while others are not noted in laboratories, and so on.

   Missing values are replaced by 0 if they pertain to diagnosis, procedures, and medications (because if a patient does not have a record indicating comorbidity, it means he was not diagnosed with it (0)). Otherwise, if they belong to labs and vitals, they drop from the data(when a patient doesn't have a record, we can not replace it by 0, which is used for tests with normal results).

3. **Patients with no records**
   In each cohort, there were some patients who had no record at all. As a result, they were not taken into account by our analysts and dropped from the datasets.

3.8 Dividing cohorts by year

As previously noted, we identified three cohorts and followed patients within cohort 1 for three years and cohorts 2 and 3 for five years before developing CKD. Then to make comparison between cohorts, we changed all cohorts to three years, and separated them yearly, such as one, two, and three years before to CKD, to investigate if the clinical features altered annually prior to CKD or not. The goal is to group patients within each cohort by passing time in order to more clearly identify the phenotypes for each year.

4 Data Mining and Machine Learning Methods

Chapter Contents

2.1 Research domain 5
2.2 AKI, phenotypes, diagnosis and prediction 5
2.3 CKD, phenotypes, diagnosis and prediction 5
2.4 AKI to CKD trajectory 6
2.5 Methodology review 6
2.6 Knowledge gap 8

So far, we have generated patient groups in accordance with the study's planned objective. The next step is to look at these cohorts to see if there are any patterns or similarities between them or even within them. The objectives of this portion of the statistical analysis are to identify common tendencies and to convert them into meaningful, comprehensive information. Data mining algorithms are statistical models that are typically used for analyze big data to increase the speed and accuracy of data processing. Before using machine learning (ML) techniques, data mining methods are frequently used to investigate data and build models to find important patterns, groupings, and trends between datasets. This makes ML algorithm quicker and more accurate. We made the decision to use several data mining approaches and select appropriate ML algorithms for each of them.

4.1 Clustering

Clustering is a data mining technique which divides the data into several distinct groups so that data points in a group are more similar to each other compared to the data points in a different group. On another words, within group variation is far less compared to the between group variation. The groups are created based on different combinations of the underlying variables. Clustering helps us to unwrap the hidden structure in the data, capture different characteristics in separate clusters and summarize the data in a meaningful way. Clustering is an unsupervised learning technique. Unlike classification, the clustering method has no predefined classes. It needs less or no information for analyzing the data. There are several clustering algorithms which can be used to analyze our data. In this research, we choose to utilize two clustering algorithms from two different families, namely K-means clustering which represents the traditional heuristic based approaches and latent class analysis (LCA) which represents the statistical model based approaches. They both are discussed in this section. However, to produce meaningful clusters and to interpret these clusters for decision making we need additional methods to make the data ready for clustering and finding important variables from resulting clusters. These methods are known as Dimensionality reduction and variable selection approach. They are discussed in more detail below.

4.1.1 Dimensional reduction and variable selection

For cohort 1, there are 1322 variables, and for cohorts 2 and 3, there are 1910 variables. Cohort 1 has been divided into three years, while cohorts 2 and 3 have been divided into five years. We have 782 variables for every year, including 50 diagnoses, 100 procedures, 132 lab results and vital signs (33 for every three months over a year), and 500 medications. We looked at patients on an annual basis to determine if their profiles changed over time.

The new cohorts still have an excessive number of variables. In order to assist ML algorithms learn and uncover significant patterns when dealing with big data, we first take into account as many variables as we can. However, this slows down the learning process for ML algorithms, increases overfitting, and makes the data difficult to interpret. Thus, dimensional reduction of the data is required. It helps the algorithms to reduce the number of features without losing much information and helps to improve the model.

Algorithms for dimensionality reduction are different. Based on the goal of the project and the available dataset, the best algorithm can be chosen. However, there are two groups into which different dimension reduction techniques can be divided:

1. **Features selection**: Techniques are employed to choose the most important variables among existing known characteristics and features. In this instance, a technique called random forest is applied.

2. **Features extraction**: The process of feature extraction comprises reducing the number of features and linearly merging them. The dimensionality reduction has resulted in several significant principle components, similar to those found using the PCA technique.

Both techniques for dimension reduction were applied in this study. Before using K-means, logistic PCA was utilized for cohort analysis, and the random forest method was employed to identify key factors for each cluster in LCA.

4.1.2 Logistic PCA

Principal component analysis (PCA) is the feature extraction method in data science that helps keep the maximum amount of information by decreasing the number of features . PCA removes correlated features and reduces over-fitting. Reducing the dimension means losing the information. In PCA, the cumulative variance of the remaining data helps us to maintain the desired accuracy.

PCA decreases the number of variables to P components, where each component is the linear combination of the original variables in the way that keep maximum variation of the original variables. Principal components are orthogonal at the points of their intersections. In another word, the original data project to the p vectors and the best fit is one which minimize the average squared distance of points to the p vectors. To project the data into a new subspace with fewer dimensions, PCA technically seeks the eigenvectors of a covariance

matrix with the greatest eigenvalues and uses those. Practically, PCA creates a new dataset of less than n features from a matrix of n features. In other words, it decreases the number of features by creating new, fewer variables that effectively capture a significant amount of the data included in the original features.

The previous data files were replaced by the new data files, which is obtained through logistic PCA. The number of patients is unchanged in the new data files; the number of variables changed to principal components, and the data value has changed to the logistic PCA score[40, 41]. Figure 4.1 shows the steps of logistic PCA.

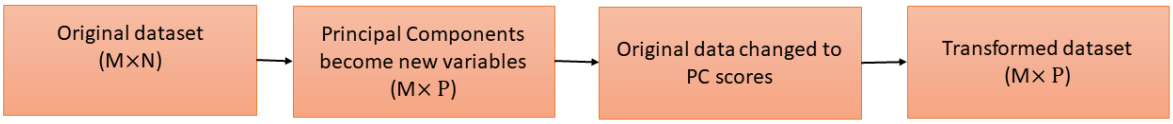


Figure 4.1: Logistic PCA steps

.

We have binary datasets, so we used logistic-PCA. It is a dimensionality reduction algorithm for binary data. For binary datasets like our dataset, the parameters project from the Bernoulli concentration model and minimize the Bernoulli variance.[42].

Before performing the K-means clustering, we extracted features using the R package LogisticPCA. This package needs two hyperparameters to be defined. m is a tuning parameter or principal component score, and k is the total number of principal components. Since the saturated parameters that come from natural parameters have both positive and negative values, tuning parameters are employed to regularize and approximate them. The m is utilized by cross validation to minimize over-fitting by choosing a number of random subsets of the data and measuring their average error, so the best hyperparameter is the value with the lowest cross validation value. On the other hand, k, which specifies the number of principal components, would be chosen in accordance with the original dataset variation that was desired to be maintained. The variance of the original dataset increases with a higher number of k. We decide to retain 80% of the variation from the original datasets in this section, and we then determine the k (number of components) that best serves our objectives.

4.1.3 K-means clustering

K-means clustering is one of the widely used data clustering methods which attempts to form k-centroids for k-clusters and assign data points to their nearest clusters based on their distances to these centroids. Initially, it starts with some random locations as centroids and then try to optimize their locations in an iterative fashion.This is not a model-based method and doesn't follow a statistical method. It is appropriate for large and complex data sets. It is relatively quick and can join clusters when necessary. It produces comparatively accurate

clustering if clusters are spherical and evenly sized. K-means algorithms can be applied using following steps:

a) First, we choose K as an arbitrary number of the clusters to start.

b) Random K points are chosen as the center for each cluster.

c) Each data point are assigned to its nearest centroid, which will create the K clusters that have been defined.

d) The k centroid elements are recomputed and modify their positions after all the objects in the space have been assigned. Random K points are chosen as the center for each cluster.

e) Steps (c) and (d) are repeated until the centroids reach a position where they no longer change with respect to the distances between all the elements of their group.

K-means clustering has one hyperparameter, so it must be manually defined. This hyperparameter is the number of clusters(K). Choosing a k is the initial stage in K-means clustering, as indicated above. The best choice is better to made because, More clusters would explain more variance, but fewer points would be contained in each cluster and cause decreasing persistence. The initial clusters are designed to contribute a lot of information and are thus required since there are actually that many groups in the data. However, as it is just partitioning the real groups, the additional information will rapidly decrease if the number of clusters gets too large in the data.

Elbow Method

The best k value can be determined using a variety of techniques. The elbow technique is one of these techniques that has been used in this research as it is easy to understand and is based on the variance of the points in each cluster. The elbow method measures the value of the within cluster sum of squares (wcss) for different numbers of clusters and plots them. The best number of clusters is the point where the wcss doesn't have sharp changes.

The k-means clustering method fitted k clusters based on the results of the elbow method, and then patients were assigned to the clusters..Elbow method and K-means have been implemented by the Sklearn package in Python.

4.1.4 Latent class analysis (LCA)

Latent class analysis (LCA) is a finite mixture model based clustering approach. It uses a probability model to describe the underlying distribution of the data and generates two or more subgroups of latent classes. Different from K-means algorithm, LCA develops latent classes and decide clustering membership based on these classes instead of the original variables. For each data point, LCA returns a probability of class assignment (soft assignment) instead of a hard assignment. It is considered to be more advantages than traditional clustering, as it

uses statistical methods for assigning points to clusters and can find the optimal number of clusters by following specific criteria. In LCA, we can statistically test each of the alternative models and choose a statistically significant model. Being backed up by statistical theories, LCA is more preferred in analyzing health care data. However, LCA is computationally much slower than the K-means clustering.

LCA classifies individuals based on their maximum likelihood within each class. Imagine each class is a disease and the patients belong to each class based on their disease. As a result, when a patient has multiple diseases, his symptoms are related. The purpose of LCA is to develop latent classes within which there is no longer any association between one symptom and another, so the symptoms will be conditionally independent because the class is the disease that causes their association. One important feature of latent classes is that their variables are independent of one another, and the relationship between variables is defined by the latent variable classes.

LCA provides us with more interpretable clustering because it is a probability distribution-based clustering and make more sense medically. It also shows us the prevalence of variables in each cluster. The prevalence is the proportion of patients who have specific variables in clusters. It helps us find the presence of a variable in each cluster and, generally, the change in its attendance in different clusters. These are also the reasons we used LCA clustering.

PoLCA, a R package for LCA, was used to implement LCA. Based on the Bayesian information criterion (BIC) and degree of freedom, we decided how many clusters to create. By minimizing the Bayesian information loss function, BIC estimates the accuracy of the fit. In this case, a smaller BIC is preferable. Higher degrees of freedom are required for more clusters, and the quantity of variables has an impact on this. PoLCA is applicable to positive integers, so we apply it to the original binary dataset by changing the value of zero to another integer. And we didn't use logistic PCA before applying LCA because the principal component scores were not positive integers.
LCA had applied for three cohorts, year by year over three years. LCA was applied to a couple of clusters (k = 4 through 8).Based on the degree of freedom, the number of clusters was constrained. We required extra variables to support more clusters. So, out of k = 4, 5, 6, 7, and 8, we decided on k = 5 for each of the three cohorts' three years based on the lowest value for The Bayesian information criterion (BIC).The BIC, is the standard by which a model is judged against a limited set of other models. For the majority of the cohort years, we established the BIC minimum value for k = 5, however according to a little variance in value, the best BIC was for more clusters in some of the cohorts. As a result, we ultimately opted to apply LCA for k = 5 for all years in all cohorts due to the little variance.

4.1.5 Random forest

To find the differences and similarities between LCA's clusters and name them after LCA has been implemented, we need to identify the most crucial variables in each cluster. In

order to accomplish this, we used feature selection. There are several advantages to feature selection: first, it makes the model easier to interpret; second, ML models work faster with fewer features; third, it will improve the accuracy of the model and decrease over-fitting.

One of the most popular ML methods for variable selection is random forest. The weight of each feature in a random forest is determined using gini importance or the mean decrease average (MDA). This is the reduction in likelihood ratio or accuracy caused by the removal of a variable. The significance of the variable increases with the size of the decline. Here, the mean decline plays a key role in determining which variables to use.This method is one of the subsets of embedded methods, and its benefit is that it decreases over-fitting. It is also easy to interpret. So, we apply this method to find our most important variables.

The random forest algorithm can be applied by the following steps:

i. First, from the given data set, random samples(e.g., 1000) are selected.

ii. Then this algorithm will construct a decision tree for every sample. and it will get the prediction result from every decision tree.

iii. Taking into account how many times each prediction is reported

iv. The most reported prediction is chosen as the final prediction.

In this research, we used the random forest package in R to find the most important variables in each cluster based on the MDA for each variable. Then we used Jaccard similarity index to find the similarity and difference of clusters in each cohort during different years. For example similarities between clusters in year 1 cohort 1 vs. year 2 cohort 1 vs. year 3 cohort 1 and so on.

4.2 Outcome measures

Three outcomes are measured by this study:

1. **Comparison of Clustering Methods**
   The matching of patients for class assignment using two clustering methods was investigated in order to determine how patients were assigned using K-means clustering and LCA. It was used to validate LCA as a statistical method.

2. **Important Variables**
   Important variables were discovered in order to find phenotypes in each cluster, year, and cohort in general.

3. **Prevalence Calculation**
   The prevalence of a certain attribute in a population is measured in each cluster, year, and cohorts in general.

4.3 Similarity measures

The Jaccard similarity index was used to measure similarity of patients clustering across years before CKD. We want to see how patients change their clusters by time.

The Jaccard similarity index determines the similarity between two sets of data. The range of the Jaccard similarity index is 0 to 1. If it is 1, it means that two datasets are fully comparable, and if it is 0, it means that the datasets are fully dissimilar. Jaccard similarity index divides the number of comparable data in two datasets by the total number of data in the two datasets.

5 Results

Chapter Contents

3.1 Research dataset 10
3.2 Study design and setting 10
3.3 Data preparation 10
3.3.1 Patients with CKD 11
3.3.2 Patients with and without AKI 11
3.3.3 Patients with AKI prior to three years of CKD 12
3.3.4 In-patients hospitalized 12
3.3.5 Medical procedures 13
3.4 Study cohorts 13
3.5 Phenotypes profiling 14
3.6 Creating binary files 16
3.7 Adjusting for missing values 17
3.8 Dividing cohorts by year 18

This chapter considers the results of clustering and data mining methods for the three study cohorts. In the case of comparisons, three years prior to CKD were considered and analyzed cohort by cohort and year by year. Patient profiles in general had been compared between three study cohorts, and for more details, cohorts were considered year by year.

5.1 Patient population

A total of 75033 CKD patients were studied, with 21547 (28.72%) having AKI prior to CKD, 13351 (17.79%) having AKI after CKD, and 39280 (52.35%) having never had AKI. Patients with AKI prior to CKD and those who had never experienced AKI created three study cohorts. Cohort 1 contains 7442 patients, or about 10% of all CKD patients; cohort 2 contains 6408 patients, or 8.5% of all CKD patients; and cohort 3 contains 39280 patients, or 52% of all CKD patients. Other CKD patients were left out of this study. Demographics, ethnicity, and comorbidities were compared for three study cohorts in table 5.1.
Patients in cohort 1 who were hospitalized 90 days before AKI had a significantly lower age (69.35 (62-79)), a significantly higher BMI(30.93(25.79 - 35.49)), a higher proportion of male (3820 (48.64%)) and white patients (6462 (86.83%)), a higher prevalence of comorbidities such as hypertensive diseases (92%), disorders of lipoprotein metabolism and other lipidemias (78.34%), type 2 diabetes mellitus (59.42%), Other long term (current) drug therapy (86.09%), disease of the blood and blood-forming organs (68.07%), long-term (current) aspirin

use (61.25%), atrial fibrillation and flutter (42.71%), atherosclerotic heart disease of the native coronary artery without angina pectoris (59.27%), unspecified chest pain (50.75%), and chronic obstructive pulmonary disease (43.36%) in contrast to patients in cohort 3.

Additionaly, patients in cohort 2 compared to patients in cohort 3, for same variables as cohorts 1 and 3. Patients in cohort 2 who had not hospitalized before AKI, had a significantly lower age (69.55(62 - 79)), a significantly higher BMI(30.97(25.86 - 35.49)), a higher proportion of male (3165 (49.39)) and white patients (5701(88.97)), a higher prevalence of comorbidities such as hypertensive diseases (94.4%), disorders of lipoprotein metabolism and other lipidemias (83.44%), type 2 diabetes mellitus (62.97%), Other long term (current) drug therapy (91.46%), disease of the blood and blood-forming organs (72.88%), long-term (current) aspirin use (67.21%), atrial fibrillation and flutter (45.88%), atherosclerotic heart disease of the native coronary artery without angina pectoris (64.42%), unspecified chest pain (58.61%), and chronic obstructive pulmonary disease (58.31%) in contrast to patients in cohort 3 (Table 5.1).

	Cohort 1	Cohort 3	P-value	Cohort 2	Cohort 3	P-value
Demographics						
Age(years)(IQR)	69.35(62 - 79)	71.07(65 - 80)	<0.0001	69.55(62 - 79)	71.07(65 - 80)	<0.0001
BMI(kg/m2)(IQR)	30.93(25.79 - 35.49)	30.62(25.9 - 34.67)	0.0003	30.97(25.86 - 35.49)	30.62(25.9 - 34.67)	0.0001
Gender(M/F)%	3820(48.64)/3620(51.33)	18317(46.63) /20951(53.34)	<0.0001	3165 (49.39)/3241(50.58)	18317(46.63) /20951(53.34)	0.0144
Ethnicity						
Asian(%)	10(0.13)	67(0.17)	0.4809	7(0.11)	67(0.17)	0.2582
Black or African American(%)	283(3.80)	1260(3.21)	0.0109	240(3.75)	1260(3.21)	0.0305
Native Hawaiian or Other Pacific Islander(%)	3(0.04)	23(0.06)	0.5408	3(0.05)	23(0.06)	0.7150
White(%)	6462(86.83)	32410(82.51)	0.006	5701(88.97)	32410(82.51)	<0.0001
Unknown(%)	684(0.19)	5520(14.05)	<0.0001	457(7.13)	5520(14.05)	<0.0001
Comorbidities						
Hypertensive diseases(%)	6847(92.00)	31182(79.38)	<0.0001	6050(94.4)	31182(79.38)	<0.0001
Disorders of lipoprotein metabolism and other lipidemias(%)	5830(78.34)	27929(71.1)	<0.0001	5347(83.44)	27929(71.1)	<0.0001
Type 2 diabetes mellitus(%)	4422(59.42)	17085(43.5)	<0.0001	4035(62.97)	17085(43.5)	<0.0001
Other long term (current) drug therapy(%)	6407(86.09)	24715(62.92)	<0.0001	5861(91.46)	24715(62.92)	<0.0001
Disease of the blood and blood forming organs(%)	5066(68.07)	14645(37.29)	<0.0001	4670(72.88)	14645(37.29)	<0.0001
Long term (current) use of aspirin(%)	4558(61.25)	14974(38.12)	<0.0001	4307(67.21)	14974(38.12)	<0.0001
Atrial fibrillation and flutter(%)	3179(42.71)	8822(22.46)	<0.0001	2940(45.88)	8822(22.46)	<0.0001
Atherosclerotic heart disease of native coronary artery without angina pectoris(%)	4411(59.27)	14310(36.43)	<0.0001	4128(64.42)	14310(36.43)	<0.0001
Chest pain, unspecified(%)	3777(50.75)	9466(24.090	<0.0001	3756(58.61)	9466(24.090	0
Chronic obstructive pulmonary disease(%)	3227(43.36)	9212(23.45)	<0.0001	3096(48.31)	9212(23.45)	<0.0001

Table 5.1: Clinical characteristic in all cohorts

5.2 Logistic PCA and K-means clustering results

The logistic PCA was applied to the original data set before applying the K-means clustering. K-means clustering is a centroid-based method, so it works better for smaller dimensions of the dataset. Logistic PCA reduces the dimension of the data by reducing the number of variables. The associated cumulative explained variance derived from logistic PCA for varying numbers of principle components between 1 and the total number of variables in each year of cohorts (774), had been kept to retain 80% of the variance of the original data. As a result the number of variables is reduced to the principal components as shown in Table 5.2.
Figure 5.1 shows the change in cumulative explained variance for different numbers of principal components for cohort 1. If all of the original variables were retained, the cumulative explained variance would be equal to one, as expected. Figures for cohorts 2 and 3 are available in the appendix A.

Cohort 1	P	Cohort 2	P	Cohort 3	P
Year 1	33	Year 1	55	Year 1	5
Year 2	23	Year 2	39	Year 2	2
Year 3	20	Year 3	30	Year 3	2

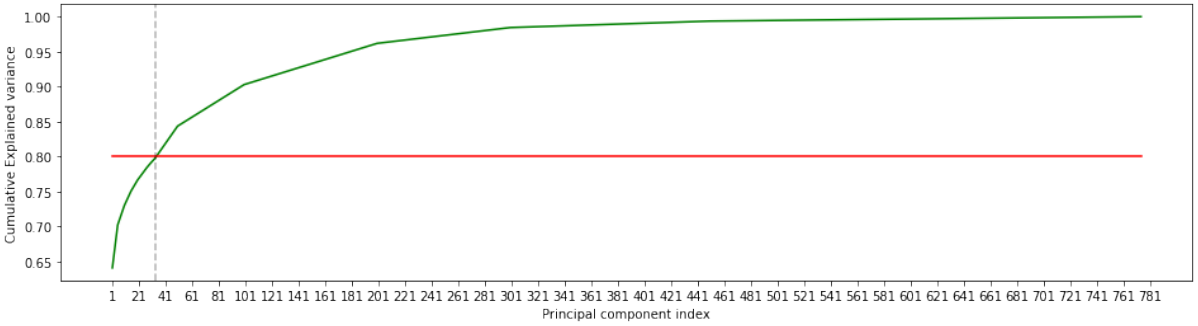
Table 5.2: Principal Components

After that we applied K-means clustering to analyze the results of the logistic PCA. The best number of clusters for K-means has been determined using the elbow method. The elbow method was designed to see how the variance was distorted or changed by varying the number of clusters between 1 and 9. The optimum number of clusters is the number of clusters after which the distortion of variance is minimized by increasing the number of clusters after it. It also depends on how it is interpreted.

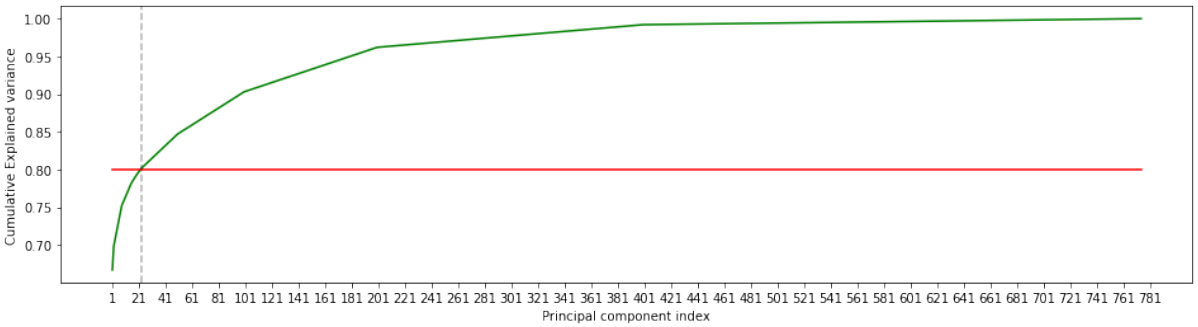
Figure 5.2 show the results of elbow method for the year 1 cohort 1, year 2 cohort 1, and year 3 cohort 1. Similar work was done for year 1 cohort 2, year 2 cohort 2, year 3 cohort 2, year 1 cohort 3, year 2 cohort 3, and year 3 cohort 3. Figures for cohorts 2 and 3 are available in the appendix B. Table 5.3 shows the number of clusters selected by the elbow method for each year in each cohort.

Cohort 1	K	Cohort 2	K	Cohort 3	K
Year 1	4	Year 1	4	Year 1	6
Year 2	4	Year 2	5	Year 2	4
Year 3	4	Year 3	4	Year 3	4

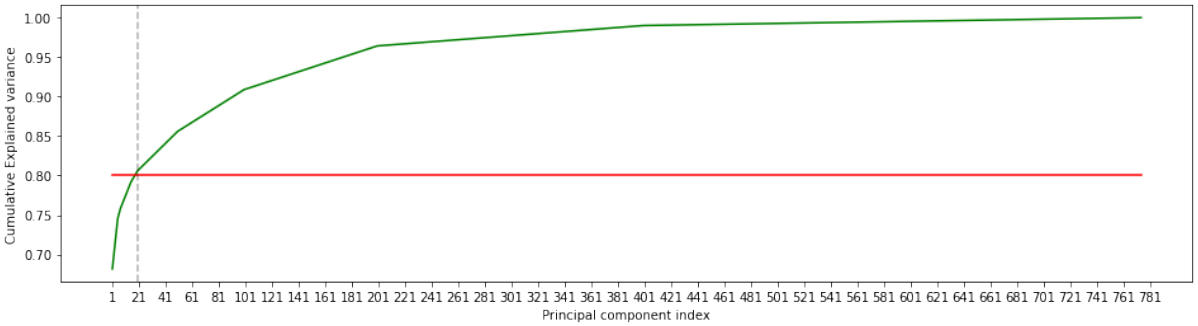
Table 5.3: Number of clusters selected for K-means clustering by the elbow method



(a)

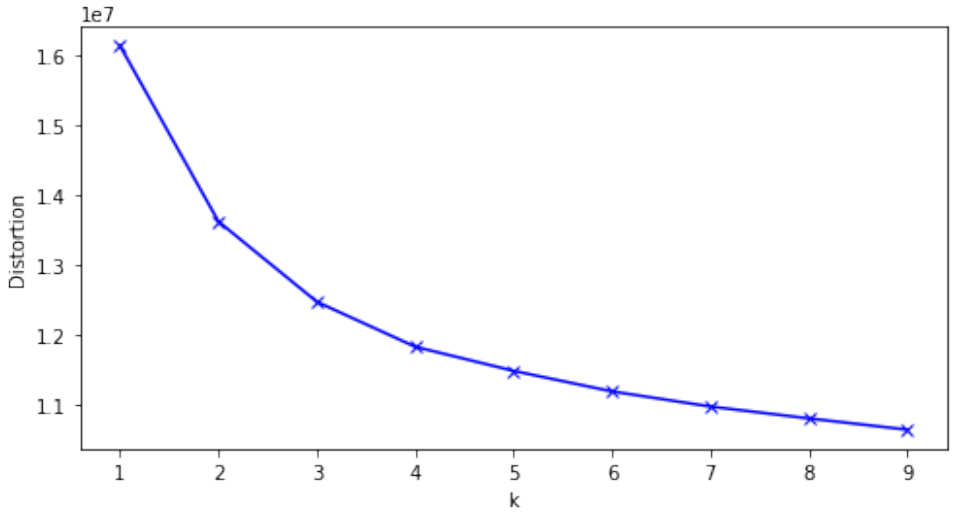


(b)

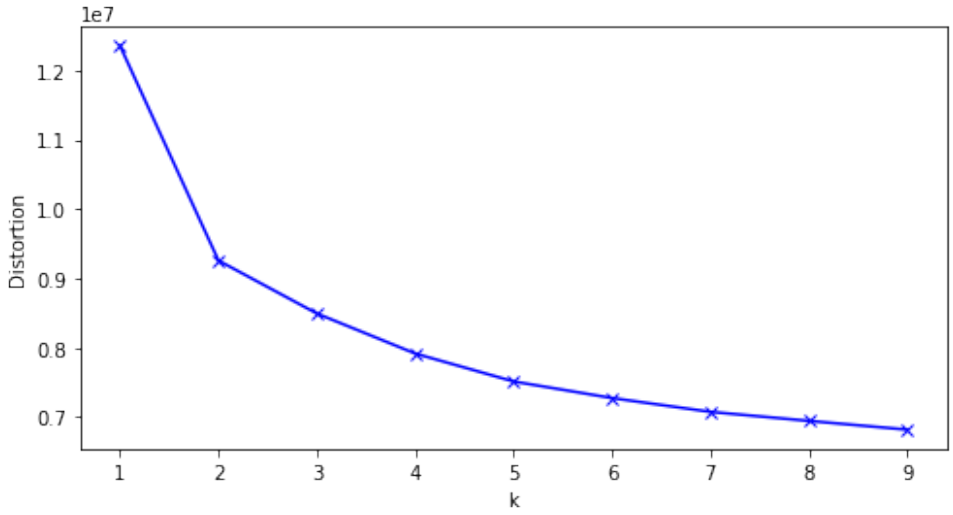


(c)

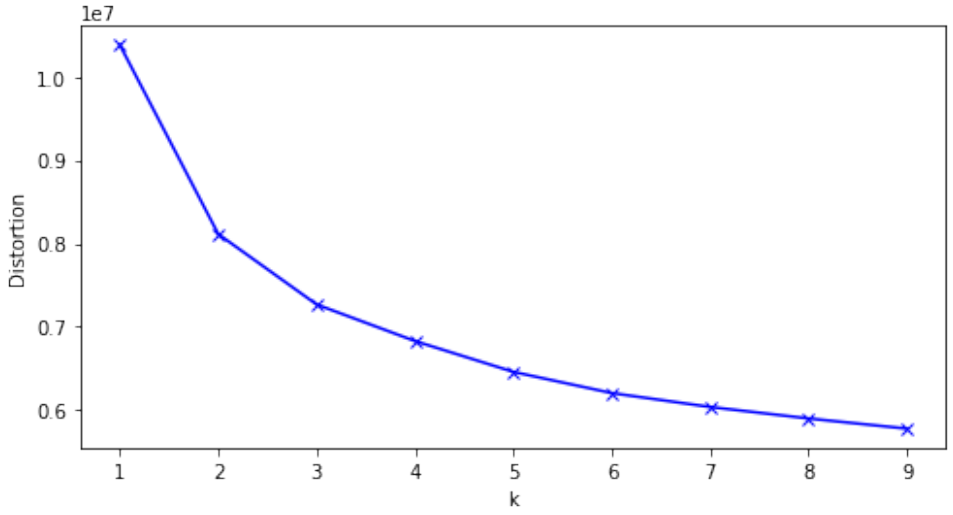
Figure 5.1: The number of principal components for cohort 1. (a) year 1; (b) year 2; (c) year 3 .



(a)



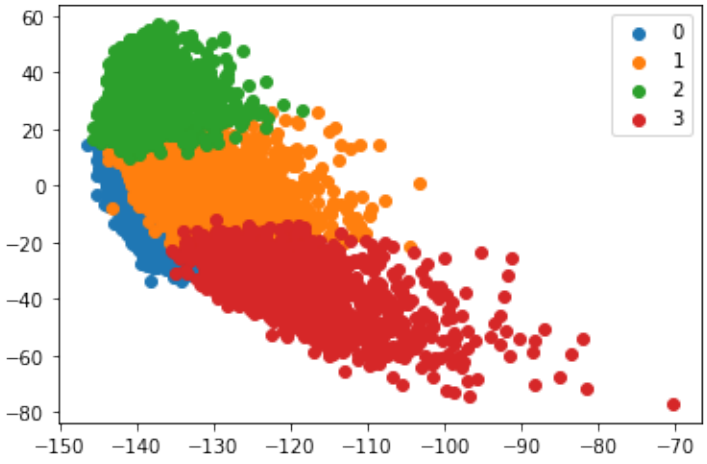
(b)



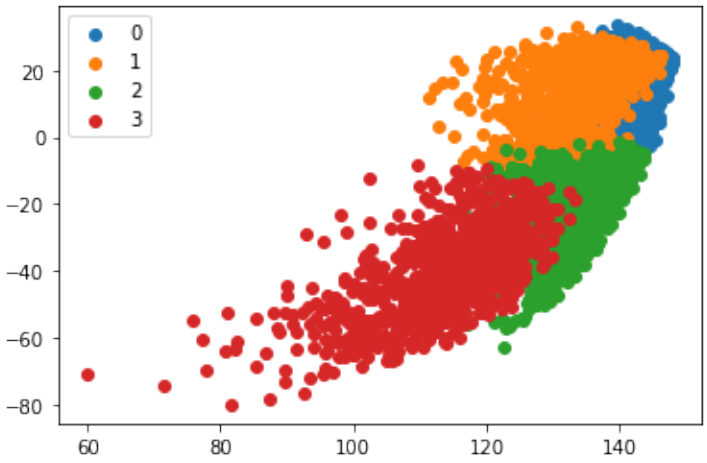
(c)

Figure 5.2: The Elbow Method showing the optimal number of clusters (k). (a) year 1; (b) year 2; (c) year 3 .

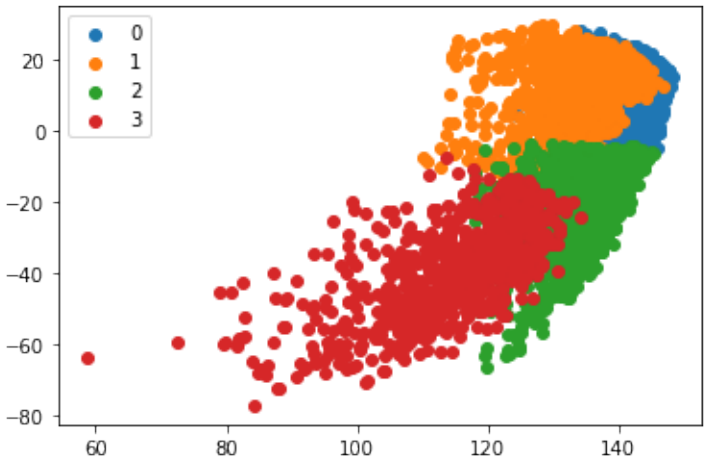
The next step after finding the optimum number of clusters is fitting K-means clustering to the data set for a number of clusters calculated by the elbow method. For example, years 1, 2, and 3 for cohort 1 are fitted to K-means for k = 4, year 1 cohort 2 is fitted to k = 4, year 2 cohort 2 is fitted to k = 5, and so on. Clusters of K-means are shown in figure 5.3 to help visualize and better understand the K-means clustering. These plots belong to the year 1 cohort 1, the year 2 cohort 1, and the year 3 cohort 1. Similar work was done for year 1 cohort 2, year 2 cohort 2, year 3 cohort 2, year 1 cohort 3, year 2 cohort 3, and year 3 cohort 3. Figures for cohorts 2 and 3 are available in the appendix C.



(a)



(b)



(c)

Figure 5.3: Clusters of K-means for cohort 1. (a) year 1; (b) year 2; (c) year 3.
To visualize K-means clustering, data points are projected on two axes; the values on the x and y axes are connected to these two PCs and do not exhibit a distinctive variable.

5.3 Latent class analysis (LCA) and variable selection results

LCA is only applicable to positive integer data, so we couldn't use it on the logistic PCA files because the PCA scores aren't positive integers. Instead, we used LCA on the original dataset with the original variables. After LCA, each patient was assigned a cluster between 1 and 5, and then a random forest was applied to the data. For each cluster in each year, we found the top 10 most important variables based on the mean MDA value that random forest gave us, from each group of variables (diagnosis, procedures, medications, lab results, and vitals). Additionally, we discovered each year's prevalence probability for every cluster. The rate of patients who have particular variables in each cluster is shown by the prevalence probability. For instance, if the variable is hypertension and it equals 0.9 for cluster 1, 90% of the patients in cluster 1 have hypertension, and so on.

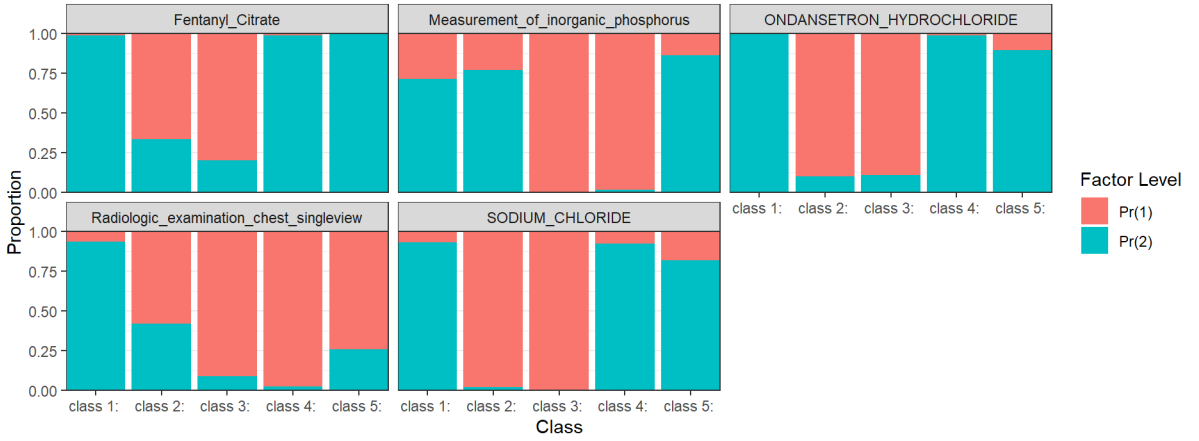
We could not label the clusters using the important clustering factors because we couldn't find any transparent instructions or information. There were too many common variables in clusters. Instead, we used the similarity and dissimilarity of important variables across clusters to find out the important variables across years and cohorts.

This study demonstrates that the risk varies depending on the patient group, and clustering helps highlight the significance. From the prevalence probability tables for each year, we choose five variables as an example to demonstrate this. All of these variables had a high prevalence probability in one or more clusters while having a low prevalence probability in others. Figure 5.4 illustrates it for three years in cohort 1, Figure 5.5 illustrates it for three years in cohort 2, and Figure 5.6 illustrates it for three years cohort 3. Two factor levels (Pr(1) denotes the proportion of variables' existence on a cluster, while pr(2) denotes its absence) are represented in these graphs by the colors orange and blue. and they illustrate the likelihood that a certain cluster will include a particular variable.

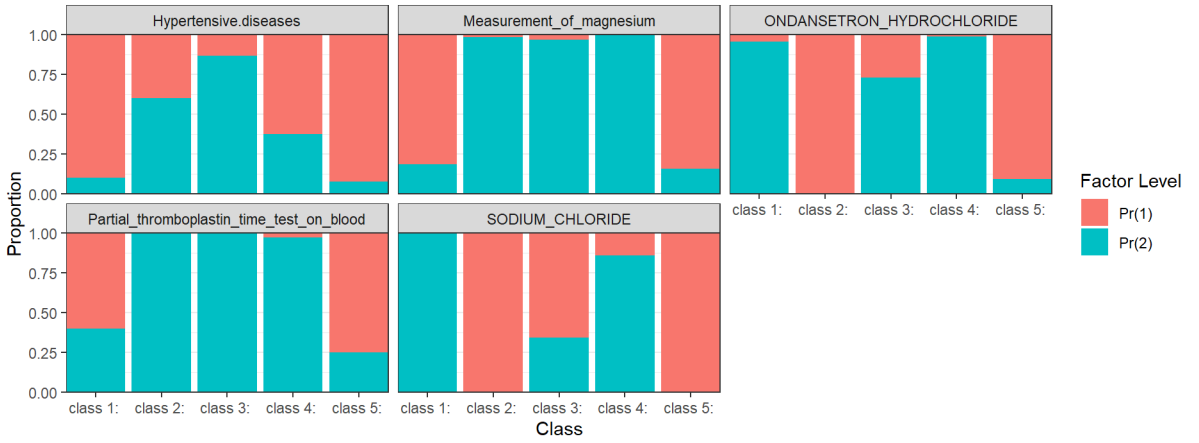
For instance, sodium chloride, which is one of the important factors in year 1 cohort 1, is an important phenotype for patients in classes 1, 4, and 5, with a probability of more than 80%; however, for patients in classes 2 and 3, it is a phenotype without any statistical test. On the other hand, hypertensive disease, three years prior to CKD for patients in cohort 1, classes two, three, and four, is an important phenotype, as almost all patients in these classes are diagnosed with it, but in classes one and five, it doesn't seem to be a phenotype. Ondansetron hydrochloride is a drug that has been identified as an important phenotype for most of the classes in cohort one during all three years. Two to three years before diagnosis, hypertensive diseases are significant phenotypes for CKD. For Cohort 2, the prothrombin time test and the partial thromboplastin time test are both phenotype for some patients during the first year prior to CKD. Two years prior to CKD, breathing abnormalities are one of the most important phenotype for some groups of patients. Three years prior to CKD-related hypertension disease, as in cohort 1, are important phenotype.

For cohort 3, cough was found as a prevent phenotype for year 1. Again, sodium chloride is a phenotype for year 1 in cohort 3 for specific groups of patients, same like year 1 in cohort

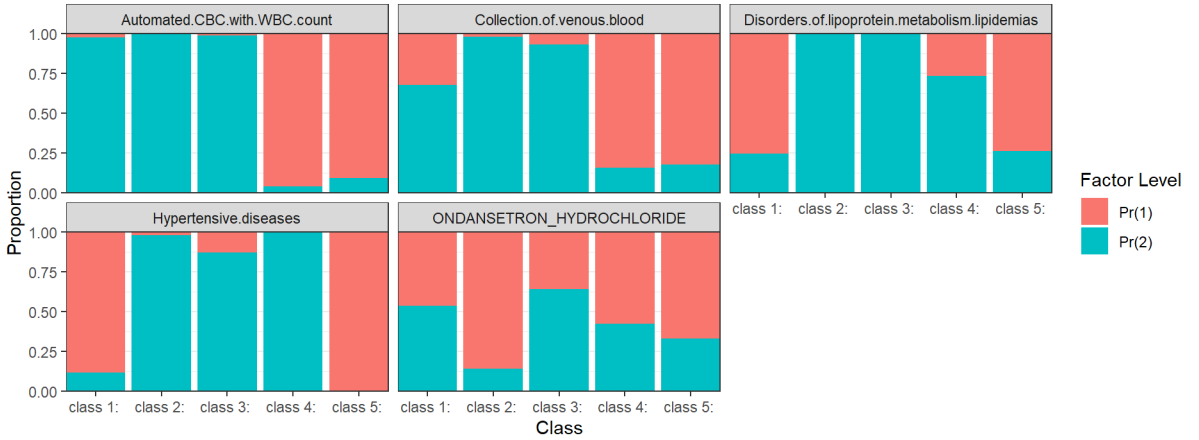
1. Collection of venous blood is a phenotype for some groups of patients two years prior to CKD in ”all three cohorts.” Fentalycitrate is a phenotype in year 2 for cohorts 2 and 3, but not for cohort 1. Hypertensive disease is a phenotype three years prior to CKD for all three cohorts.
Here are a few examples of how clustering can be used to identify phenotypes over time for various patient groups across various study cohorts.



(a)

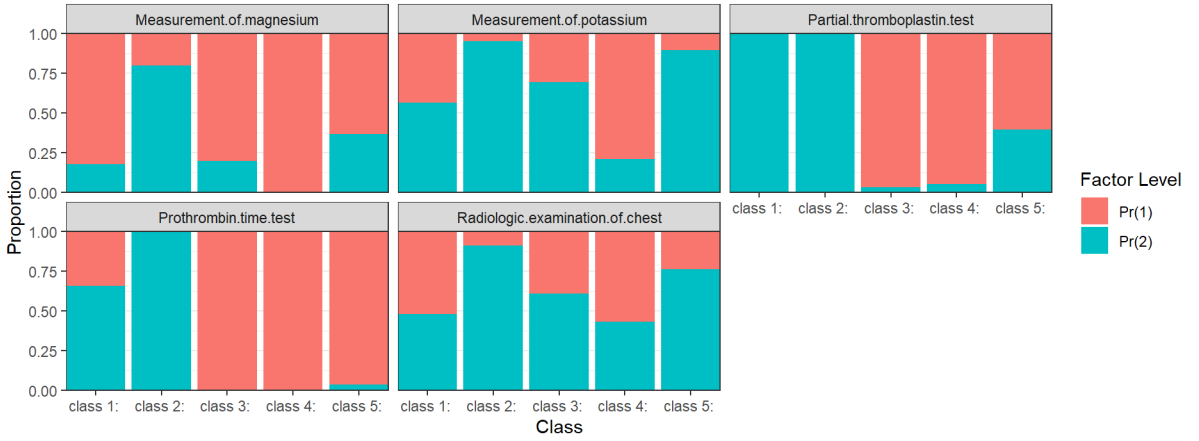


(b)

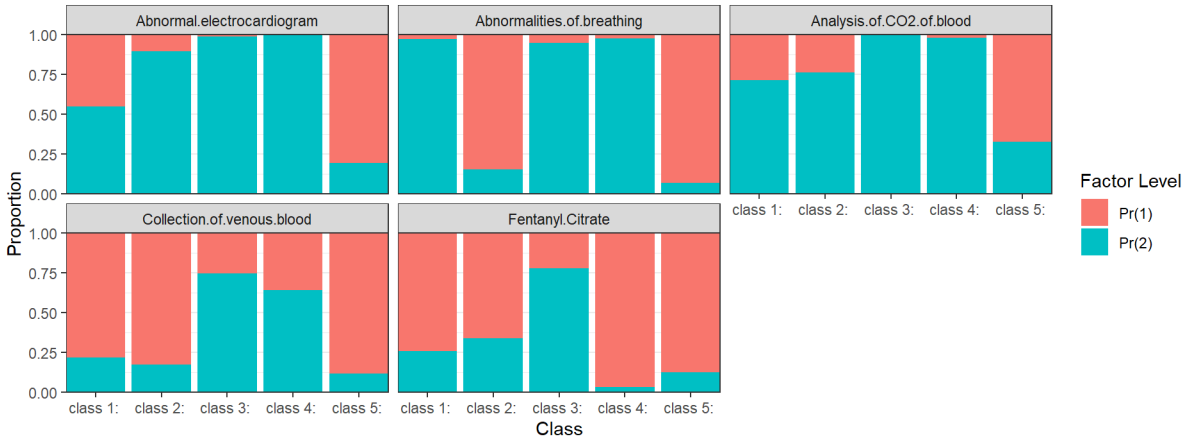


(c)

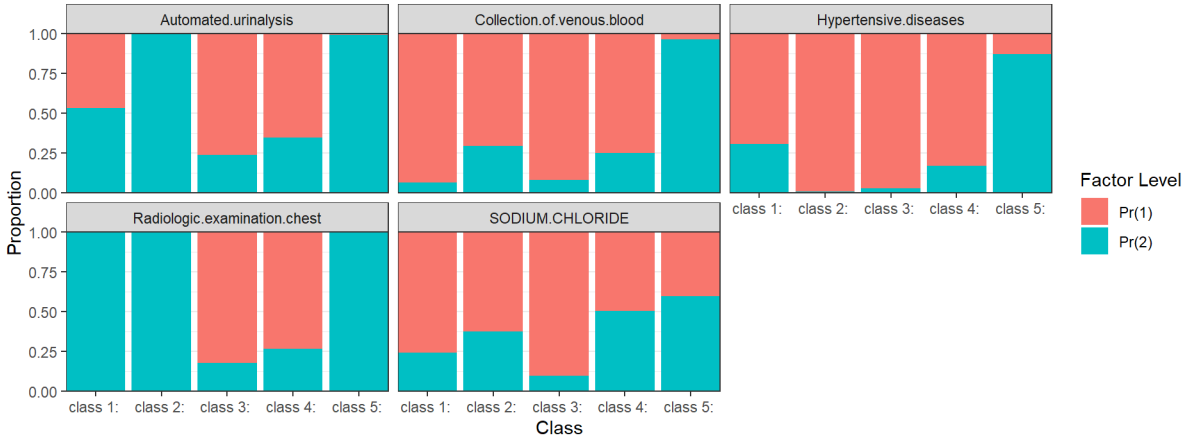
Figure 5.4: Variable prevalence in different clusters based on the LCA method for cohort 1. (a) year 1; (b) year 2; (c) year 3 .
Pr(1) denotes the proportion of variables' existence on a cluster, while pr(2) denotes its absence.



(a)

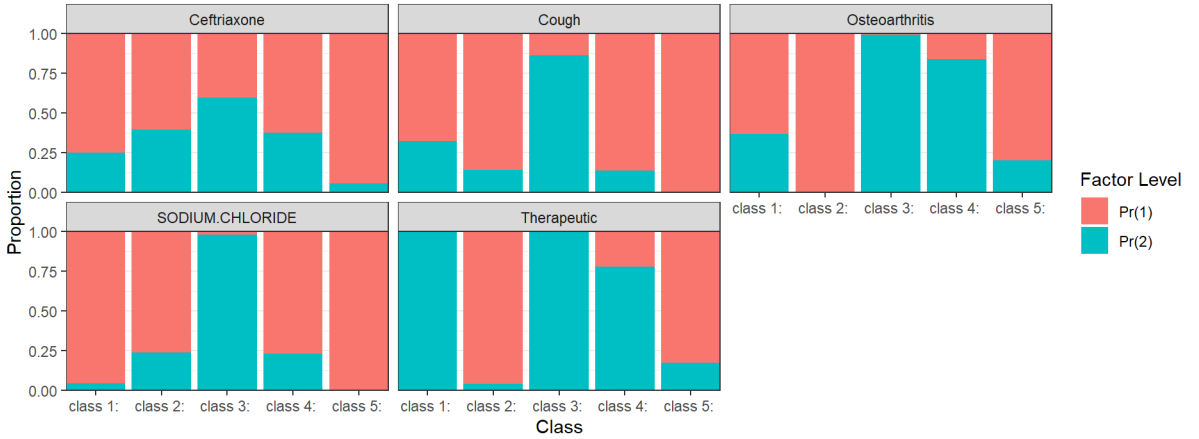


(b)

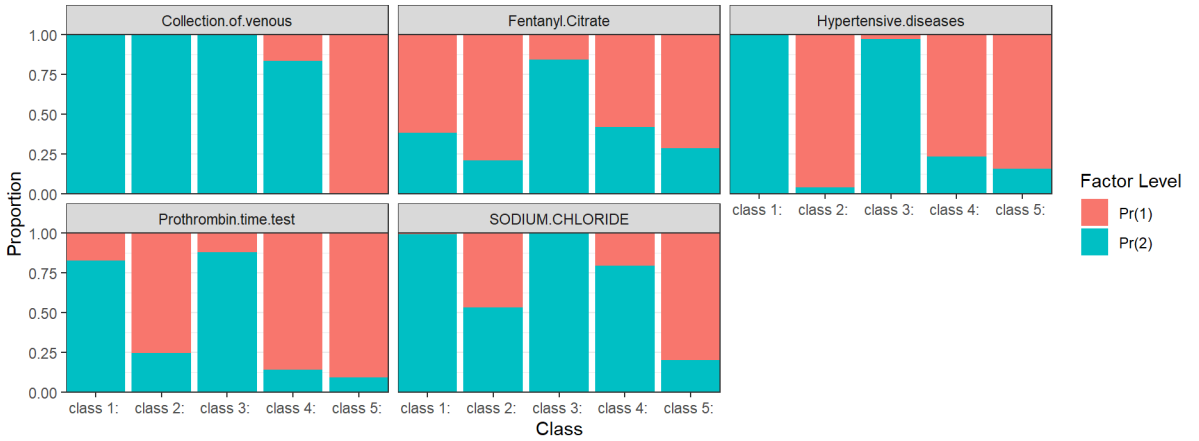


(c)

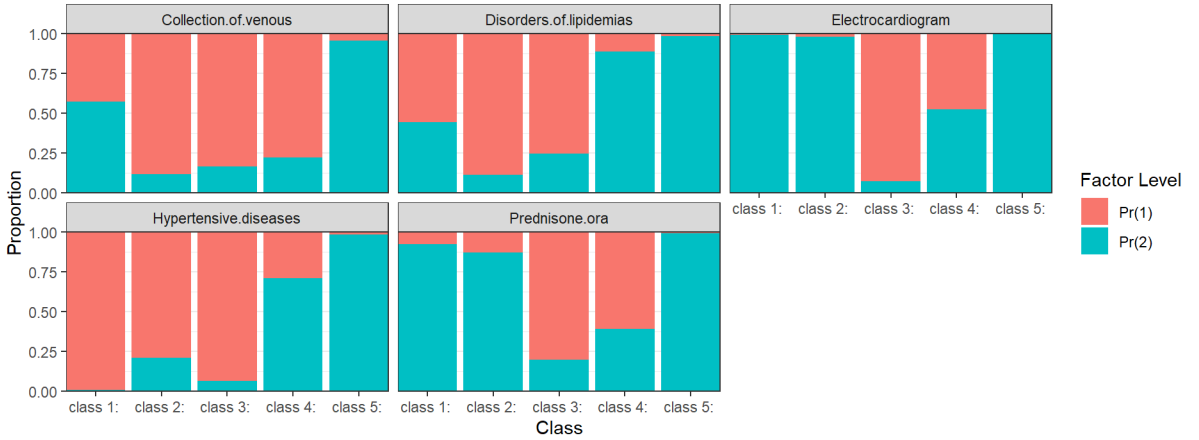
Figure 5.5: Variable prevalence in different clusters based on the LCA method for cohort 2. (a) year 1; (b) year 2; (c) year 3 .
Pr(1) denotes the proportion of variables' existence on a cluster, while pr(2) denotes its absence.



(a)



(b)



(c)

Figure 5.6: Variable prevalence in different clusters based on the LCA method for cohort 3. (a) year 1; (b) year 2; (c) year 3 .
Pr(1) denotes the proportion of variables' existence on a cluster, while pr(2) denotes its absence.

In order to determine whether or not demographic factors may be taken into account as phenotypes, we also took into account the age, sex, and race of patients in each class. As a result, as shown in Table 5.1, none of age, gender, or race can be considered phenotypes for the various cohort 1 groups because they are all close; the average age for patients in all clusters is close to 70 years old, and almost 90% of patients were white. The demographic data for cohorts 2 and 3 are shown in tables 5.2 and 5.3. Tables indicate that in Cohort 2 year 1, in the first cluster, the majority of people were female (66.67%). A partial thromboplastin time test is a phenotype in this group of patients. Additionally, the majority of patients in cohort 2 year 3 were female (66.67%) . Sodium chloride and venous blood collection and hypertensive disease are not phenotypes for this cluster, however, radiologic evaluation of the chest is.

		Year 1					Year 2					Year 3				
		C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Sex	Male	53.83	0	49.26	58.62	51.27	52.52	50.41	51.37	52.03	45.33	49.2	49.14	53.42	54.08	45.3
	Female	46.12	100	50.68	41.38	48.73	47.48	49.59	48.57	47.97	54.67	50.8	50.79	46.58	45.92	54.7
Age	mean(SD)	69.3(12.9)	20(0)	67.22(12.4)	67.5(15.6)	70.97(12.6)	70.16(13.13)	70.12(12.1)	69.28(12.6)	70.5(12.4)	67.2(12.7)	70.1(12.2)	68.81(12.8)	69.4(12.6)	70.8 (13.3)	68.3(12.3)
Race	Asian	0.23	0	0.25	0	0.05	0.27	0.14	0.13	0.11	0.11	0	0.15	0	0.34	0.24
	Black or African American	2.97	0	2.4	0	3.77	3.01	2.88	3	4.17	2	3.02	2.7	3.74	3.35	2.56
	Native Hawaiian or Other Pacific Islander	0	0	0	0	0.05	0	0	0	0.11	0	0	0.08	0	0	0
	White	86.31	100	91.88	82.76	88.53	87.07	92.33	88.53	91.33	93.67	92.9	87.23	92.71	88.33	92.92
	Unknown	10.49	0	5.47	17.24	7.6	9.65	4.66	8.35	4.28	4.22	4.09	9.84	3.56	7.98	4.27

Table 5.4: Demographic information based on the clusters for cohort 1

		Year 1					Year 2					Year 3				
		C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Sex	Male	53.83	0	49.26	58.62	51.27	52.52	50.41	51.37	52.03	45.33	49.2	49.14	53.42	54.08	45.3
	Female	46.12	100	50.68	41.38	48.73	47.48	49.59	48.57	47.97	54.67	50.8	50.79	46.58	45.92	54.7
Age	mean(SD)	69.3(12.9)	20(0)	67.22(12.4)	67.5(15.6)	70.97(12.6)	70.16(13.13)	70.12(12.1)	69.28(12.6)	70.5(12.4)	67.2(12.7)	70.1(12.2)	68.81(12.8)	69.4(12.6)	70.8 (13.3)	68.3(12.3)
Race	Asian	0.23	0	0.25	0	0.05	0.27	0.14	0.13	0.11	0.11	0	0.15	0	0.34	0.24
	Black or African American	2.97	0	2.4	0	3.77	3.01	2.88	3	4.17	2	3.02	2.7	3.74	3.35	2.56
	Native Hawaiian or Other Pacific Islander	0	0	0	0	0.05	0	0	0	0.11	0	0	0.08	0	0	0
	White	86.31	100	91.88	82.76	88.53	87.07	92.33	88.53	91.33	93.67	92.9	87.23	92.71	88.33	92.92
	Unknown	10.49	0	5.47	17.24	7.6	9.65	4.66	8.35	4.28	4.22	4.09	9.84	3.56	7.98	4.27

Table 5.5: Demographic information based on the clusters for cohort 2

		Year 1					Year 2					Year 3				
		C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Sex	Male	50.55	46.18	42.15	48.67	41.58	38.66	49.79	43.64	40.37	47.85	38.61	46.96	42.82	46.83	48.15
	Female	49.43	53.81	57.85	51.28	58.39	61.34	50.21	56.34	59.63	52.11	61.39	53	57.18	53.17	51.82
Age	mean(SD)	72.8(12.5)	70.7(12.7)	69.8(12.5)	71.4(12.8)	71.3(12.3)	71.28(12.1)	70.8(12.9)	72.05(11.8)	70.3(12.6)	71.3(12.9)	70.53(12.7)	71.41(12.6)	71.9(11.9)	70.28(12.9)	71.32(12.7)
Race	Asian	0.16	0.24	0.17	0.14	0.06	0	0.08	0.23	0.16	0.2	0.03	0.23	0.26	0.11	0.15
	Black or African American	3.62	2.69	3.1	2.84	1.95	2.02	3.26	2.64	2.64	2.9	2.3	3.19	3.23	2.22	2.92
	Native Hawaiian or Other Pacific Islander	0.1	0.04	0.05	0.06	0	0	0.06	0.02	0.03	0.09	0	0	0.03	0.05	0.1
	White	82.92	85.31	91.48	77.83	92.34	94.77	84.78	92.09	93.64	80.4	94.9	84.53	93.33	88.53	81.56
	Unknown	13.21	11.72	5.19	19.13	5.65	3.21	11.76	5.02	3.54	16.41	2.77	12.04	3.15	9.09	15.26

Table 5.6: Demographic information based on the clusters for cohort 3

5.4 Comparison of LCA clusters with K-Means clusters

For years 1, 2, and 3 in cohort 1, the elbow generating an ideal k of 4, while LCA produced an ideal number of clusters of 5. Based on the LCA and K-means, we attempted to determine how closely a patient's class fits, and we compared these two approaches to see if they produced comparable or dissimilar findings. So, we discovered patients who were assigned to cluster 1 using K-means and then see how they were assigned using LCA. Figure 5.7 compares K-means clustering and LCA for year 1 in cohort 1.

For year 1, as shown in figure 5.7, almost all patients in the first k-means cluster were in the third LCA cluster. The first LCA cluster contained about 75% of the patients in the second k-means cluster and 60% of the patients in the fourth k-means cluster. Eighty percent of the patients in the fifth LCA cluster are in the third K-means cluster. Patients allocated to classes 2 and 4 by LCA did not share classes when using the k-means clustering approach. It could be because there are more classes in LCA than in k-means. Moreover because LCA is a more trustworthy clustering technique than k-means, which is a centroid technique.

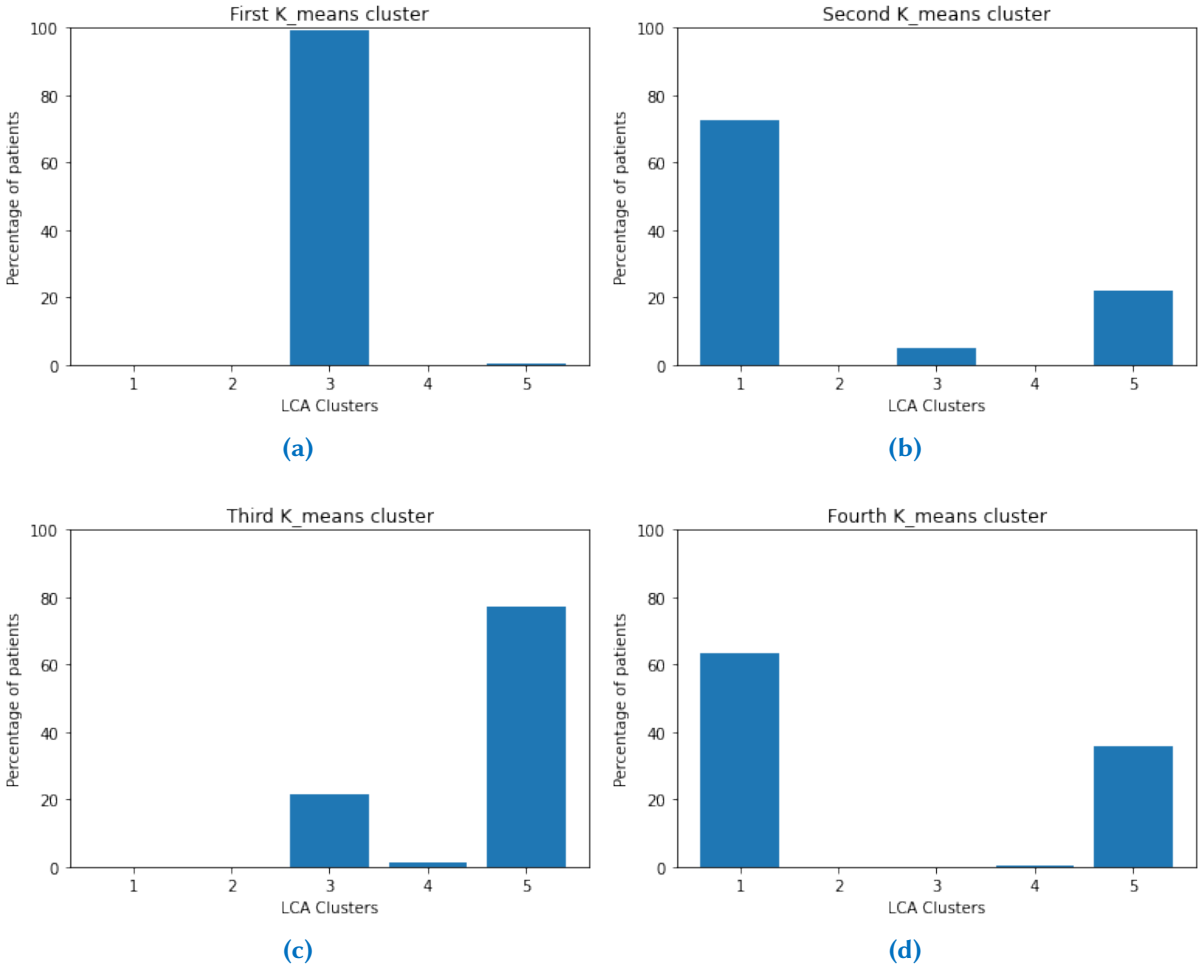


Figure 5.7: Comparison of K-means clustering with latent class analysis (LCA) in year 1 cohort 1

Patients in the first cluster were assigned by k-means 40% to cluster 2, and 50% to cluster 4 by LCA in year 2 of cohort 1. Likewise, 90% of patients in the second k-means cluster were assigned to the third LCA cluster. The first LCA cluster was assigned to 60% of patients in the third k-means cluster, and the fifth LCA cluster was assigned to nearly 90% of patients in the fourth k-means cluster.

Therefore, for year 2 cohort 1, the first k-means cluster corresponds to the second and fifth LCA cluster, the second k-means cluster corresponds to the fourth LCA cluster, the third k-means cluster corresponds to the first LCA cluster, and the fourth k-means cluster corresponds to the fourth LCA cluster. As a result, in year 2 of Cohort 1, k-means closely satisfy LCA clustering on patient profiles.

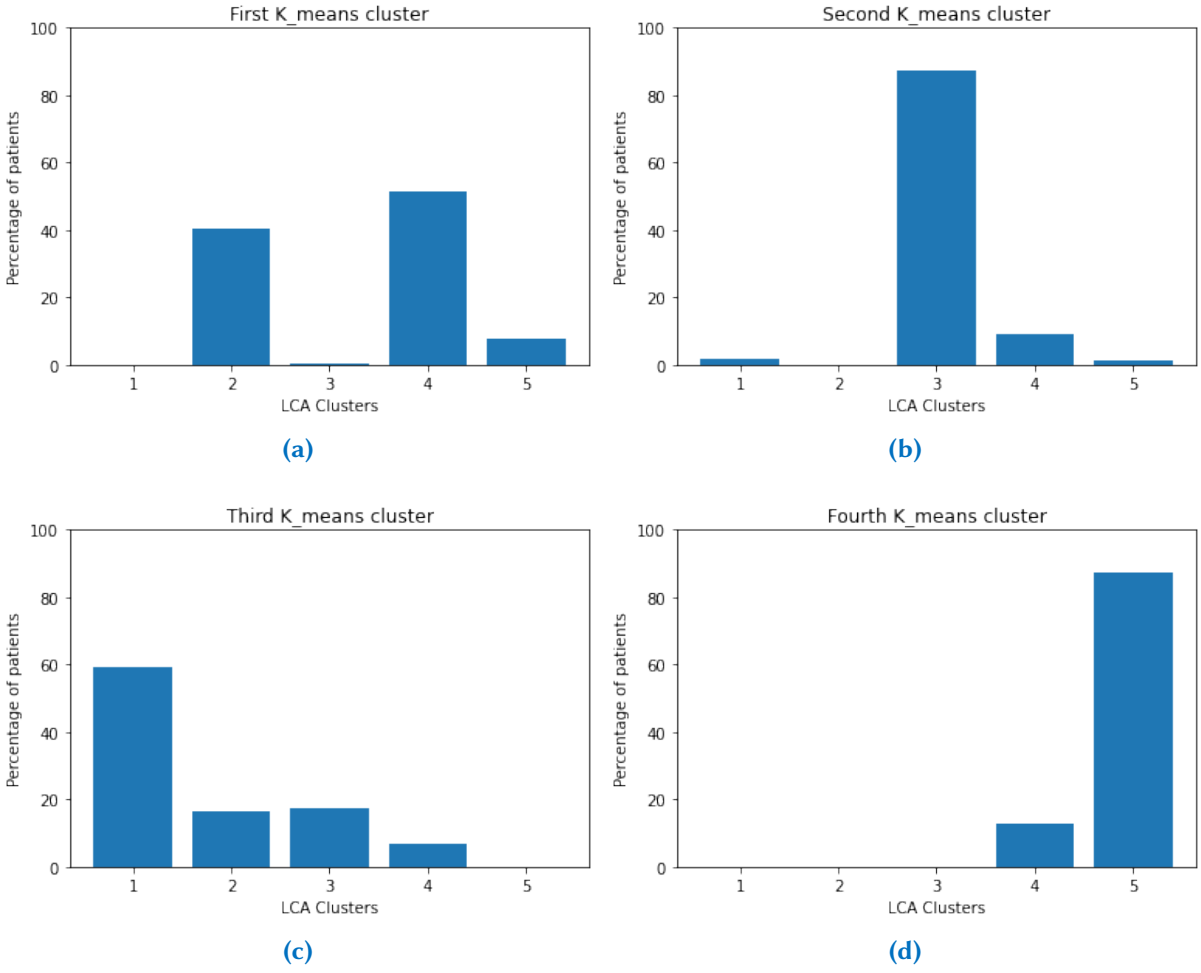


Figure 5.8: Comparison of K-means clustering with latent class analysis (LCA) in year 2 cohort 1

For cohort 1's year 3 patients, 65% of those in the first k-means cluster are matched with the second LCA cluster, 65% of those in the second k-means cluster are matched with the fourth LCA cluster, 55% of those in the third k-means cluster are matched with the first LCA cluster, and nearly all of those in the fourth k-means cluster are matched with the fifth LCA cluster. The clustering of patient profiles by LCA will be satisfied if, as in year 2, we are able to identify a corresponding cluster for each of the k-means classes in the LCA.
Because the LCA is a more credible technique than K-means, we chose its outcomes for our investigation. Additionally, we made comparable comparisons of years in Cohorts 2 and 3 using k-means and LCA. The appendix D contains the figures for similarity comparisons.

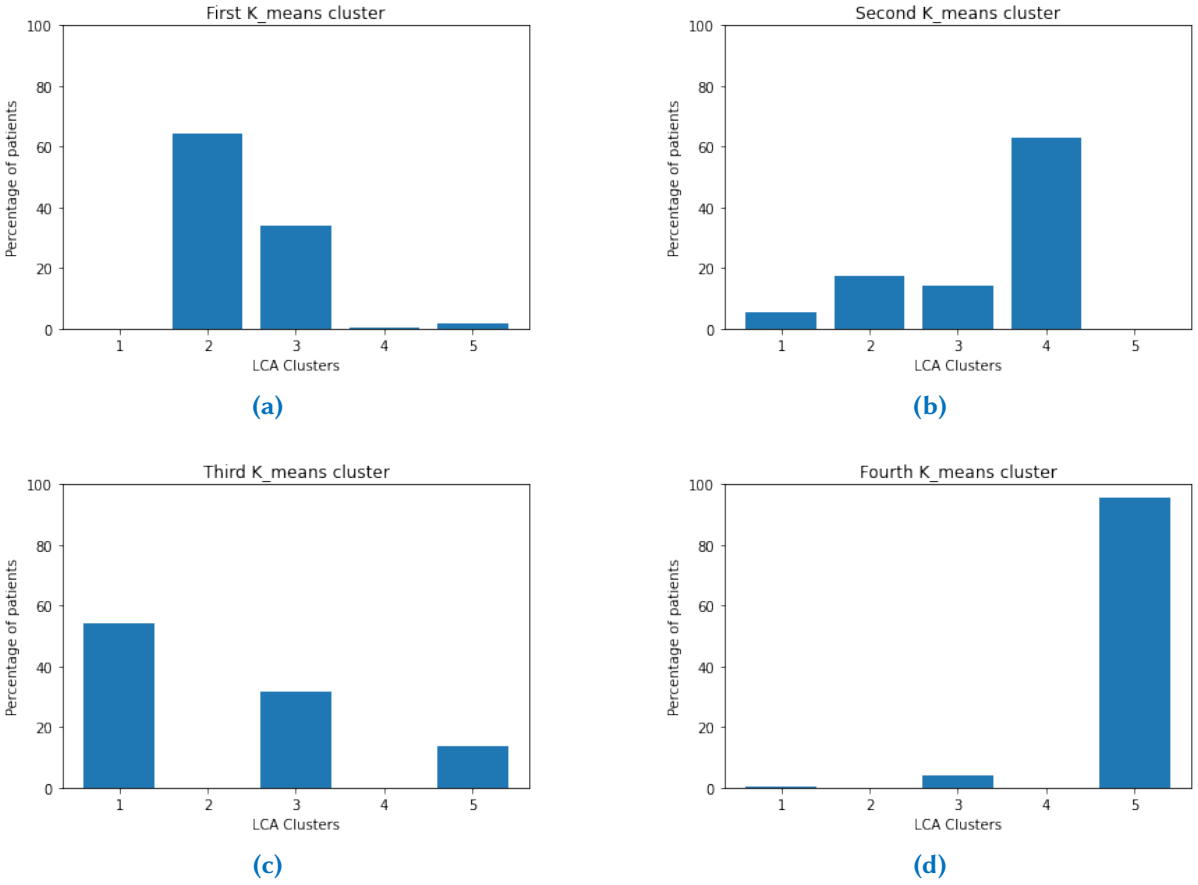


Figure 5.9: Comparison of K-means clustering with latent class analysis (LCA) in year 3 cohort 1

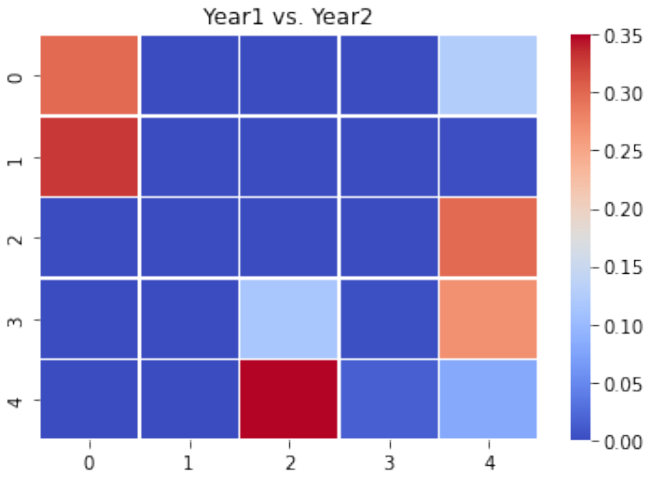
5.5 Cohort comparison and insights

In this part, we used the Jaccard similarity index to see if patients changed their cluster over the years. For instance, patients in class 1 year 1 move to class 1 year 2, or class 2 year 2, or class 3 year 2, and so on. We were attempting to determine whether or not there was a pattern for patients moving across years. During all three cohorts, we calculated the Jaccard similarity index between clusters for years 1 vs. 2, years 1 vs. 3, and years 2 vs. 3. Figure 5.10 shows the results for cohort 1. Jaccard similarity figures for cohorts 2 and 3 are available in the appendix E.

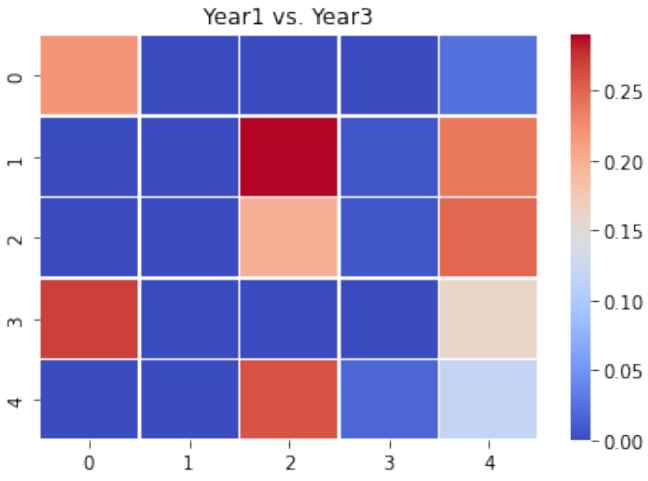
Jaccard's similarity index in general changes between 0 and 1. Index 0 indicates that there is no similarity, while index 1 indicates that they are completely similar. Low levels of patient similarity are evident in clusters, as seen in Figure 5.10. The greatest similarity index between the first and second years is 0.35; between the first and third years is 0.29; and between the second and third years is 0.46. The approach to indicate the Jaccard similarity index value and pair the colors with numbers can be seen in the line on the right of figure 5.10. This indicates that most patients' clusters change over time. Jaccard similarity, thus, failed to provide us with a clear direction across clusters in years; as a result, we decided to consider cohorts year by year.
In the rest of this research, we considered the similarities and dissimilarities among cohorts across years. For example, cohort 1 year 1 vs. cohort 2 year 1 vs. cohort 3 year 1. Then we repeated the process for the following years. for diagnosis, procedure, medication, vitals, and labs. Also, we considered similarities and dissimilarities between Cohort 1 vs. Cohort 2 vs. Cohort 3.

As previously stated, we used random forest to determine the top ten significant variables for each cluster based on the MDA. It is important to take into account all categories of factors, such as diagnosis, procedures, medication, lab results, and vital signs. There were too many variables that were shared between clusters. We then eliminated similar factors from the data and maintained just the most significant ones for each year in each cohort in order to compare and contrast patients in each year. Then, in order to compare them and discover similarities and differences between cohorts in the same year prior to CKD, we gathered all variables for several cohorts in the same year at one table. Similarly, we gathered the significant variables in each cohort to compare three cohorts. During the three years, important variables for each cohort were gathered and compared with those of the other cohorts.

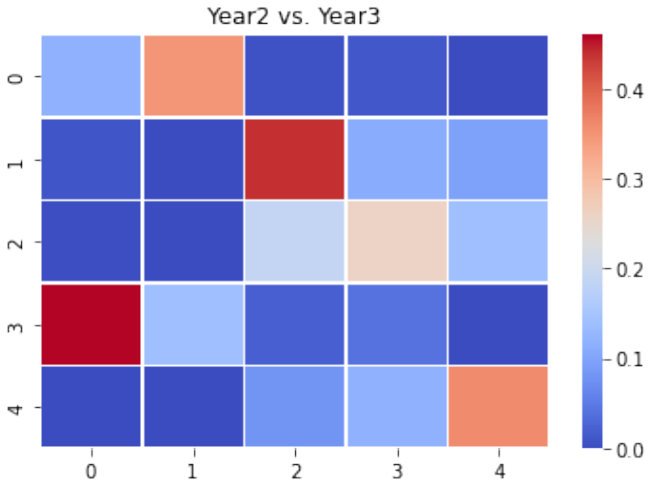
Along with the comparison of cohorts overall, each set of factors had its own considerations for the comparison of cohorts in each year and generally over three years. Details regarding the diagnosis, procedures, medications, lab results, and vital signs can be found below. All tables in this chapter are sorted based on the importance score (MDA) of variables.



(a)



(b)



(c)

Figure 5.10: Jaccard similarity index for cohort 1.

5.5.1 Diagnosis

Important phenotypes during the year prior to CKD for all three cohorts are gathered on table 5.4. We can find ”abnormal electrocardiogram [ECG] [EKG],” ”abnormalities of breathing,” and ”cardiomegaly.” ”Chronic ischemic heart disease,” ”heart failure,” ”hypertensive diseases,” ”major depressive disorder, single episode, unspecified,” ”other long-term (current) drug therapy,” ”other specified postprocedural states,” and ”pleural effusion” are common phenotypes in all three cohorts during year 1, so we can say that they can be phenotypes one year prior to CKD because this is a common criteria in all three cohorts.

In cohorts 1 and 2, common phenotypes include ”anxiety disorders,” ”hypokalemia,” ”localized edema,” and ”other chronic pain.” These factors are not in cohort 3, which was one year prior to CKD, so as the patients in cohorts 1 and 2 had prior AKI, we can say that these factors can be an outcome of AKI, which can increase the risk of CKD.

We have some factors that were present in cohort 1 during year 1 but were not in cohorts 2 or 3, such as ’Dorsalgia, unspecified’, ’Long term (current) use of aspirin’, ’Diarrhea, unspecified’, ’Long term (current) use of anticoagulants’, ’Constipation, unspecified’, ’Encounter for follow-up examination after completed treatment for conditions other than malignant neoplasm’, ’Unspecified abdominal pain’, ’Atrial.fibrillation.and.flutter’, ’Chest pain, unspecified’, ’Atherosclerotic heart disease of native coronary artery without angina pectoris’, ’Urinary tract infection, site not specified’, ’Encounter for other preprocedural examination’, ’Overweight.and.obesity’, ’Hyperkalemia’, ’Nonrheumatic mitral (valve) insufficiency’. These elements can be referred to as phenotypes for individuals who were hospitalized or had an inpatient visit 90 days prior to AKI.
On the other hand, there are some variables uniquely for cohort 3 during year 1 such as ’Encounter for general adult medical examination without abnormal findings’, ’Encounter for immunization’, ’Encounter for screening mammogram for malignant neoplasm of breast’, ’Encounter for screening for malignant neoplasm of colon’, ’Encounter for other preprocedural examination’, ’Long term (current) use of anticoagulants’, ’Chest pain, unspecified’, ’Atrial.fibrillation.and.flutter’, ’Nonrheumatic mitral (valve) insufficiency’, ’Long term (current) use of aspirin’, ’Unspecified abdominal pain’, ’Encounter for follow-up examination after completed treatment for conditions other than malignant neoplasm’. Because patients in cohort 3 never had AKI, these characteristics can be referred to as CKD phenotypes that are unrelated to AKI.
Similar conclusions may be drawn for the two and three years prior to the onset of CKD for all cohorts based on the data available in similar tables in the appendix F.
To compare the cohorts in general, we gathered the most important variables of each cohort over the course of three years in Table 5.5, regardless of yearly changes in patient profile. As this table shows, we can conjecture that some phenotypes that are common between cohorts 1 and 2, such as ”hypokalemia,” ”major depressive disorder, single episode, unspecified,” and ”other specified postprocedural states,” can be named as phenotypes of CKD for

patients who had experienced AKI. Additionally, the three cohorts shared substantial characteristics that might be referred to as phenotypes for CKD, including ”abnormal electrocardiogram [ECG] [EKG], abnormalities of breathing, hypertensive diseases, and other long-term (current) pharmaceutical therapy.”. Furthermore, ”long-term (current) use of aspirin,” ”unspecified diarrhea,” ”chronic ischemic heart disease,” ”cardiomegaly,” ”unspecified chest pain,” ”atherosclerotic heart disease of the native coronary artery without angina pectoris,” ”encounter for other pre-procedural examination,” and ”overweight and obesity” as unique factors in cohort 1 can be considered phenotypes for CKD for patients who were hospitalized or had an inpatient visit. And finally, ”encounter for general adult medical examination without abnormal findings,” ”encounter for immunization,” ”encounter for screening mammogram for malignant neoplasm of the breast,” ”encounter for screening for malignant neoplasm of the colon,” ”encounter for other preprocedural examination,” ”Vitamin D deficiency, unspecified,” and ”Long term (current) use of anticoagulants” as unique factors in cohort 3 can be identified as phenotypes for CKD for patients who have never experienced AKI.

Cohort1	Cohort2	Cohort3
Dorsalgia, unspecified	Other long term (current) drug therapy	Encounter for general adult medical examination without abnormal findings
Long term (current) use of aspirin	Other specified postprocedural states	Encounter for immunization
Major depressive disorder, single episode, unspecified	Other chronic pain	Encounter for screening mammogram for malignant neoplasm of breast
Other long term (current) drug therapy	Hypertensive diseases	Encounter for screening for malignant neoplasm of colon
Other specified postprocedural states	Abnormal electrocardiogram [ECG] [EKG]	Encounter for other preprocedural examination
Diarrhea, unspecified	Abnormalities of breathing	Abnormal electrocardiogram [ECG] [EKG]
Chronic ischemic heart disease	Anxiety.disorders	Hypertensive diseases
Long term (current) use of anticoagulants	Major depressive disorder, single episode, unspecified	Vitamin D deficiency, unspecified
Constipation, unspecified	Chronic ischemic heart disease	Other long term (current) drug therapy
Encounter for follow-up examination after completed treatment for conditions other than malignant neoplasm	Cardiomegaly	Low back pain
Abnormalities of breathing	Localized edema	Long term (current) use of anticoagulants
Abnormal electrocardiogram [ECG] [EKG]	Pleural.effusion	Chest pain, unspecified
Hypokalemia	Low back pain	Atrial fibrillation and flutter
Unspecified abdominal pain	Disorders of lipoprotein metabolism and other lipidemias	Heart.failure
Cardiomegaly	Other fatigue	Other specified postprocedural states
Anxiety disorders	Vitamin D deficiency, unspecified	Chronic ischemic heart disease
Hypertensive diseases	Edema, unspecified	Abnormalities of breathing
Other chronic pain	Disease of the blood and blood forming organs	Cardiomegaly
Pleural effusion	Heart failure	Pleural effusion
Atrial fibrillation and flutter	Hypokalemia	Disorders of lipoprotein metabolism and other lipidemias
Chest pain, unspecified		Nonrheumatic mitral (valve) insufficiency
Atherosclerotic heart disease of native coronary artery without angina pectoris		Long term (current) use of aspirin
Urinary tract infection, site not specified		Major depressive disorder, single episode, unspecified
Encounter for other preprocedural examination		Unspecified abdominal pain
Overweight and obesity		Encounter for follow-up examination after completed treatment for conditions other than malignant neoplasm
Hyperkalemia		
Localized edema		
Heart.failure		
Nonrheumatic mitral (valve) insufficiency		

Table 5.7: Year 1 diagnosis comparison

Cohort1	Cohort2	Cohort3
Long term (current) use of aspirin	Other long term (current) drug therapy	Encounter for general adult medical examination without abnormal findings
Major depressive disorder, single episode, unspecified	Other specified postprocedural states	Encounter for immunization
Other long term (current) drug therapy	Hypertensive diseases	Encounter for screening mammogram for malignant neoplasm of breast
Other specified postprocedural states	Abnormal electrocardiogram [ECG] [EKG]	Encounter for screening for malignant neoplasm of colon
Diarrhea, unspecified	Abnormalities of breathing	Encounter for other preprocedural examination
Chronic ischemic heart disease	Major depressive disorder, single episode, unspecified	Abnormal electrocardiogram [ECG] [EKG]
Abnormalities of breathing	Pleural effusion	Hypertensive diseases
Abnormal electrocardiogram [ECG] [EKG]	Disorders of lipoprotein metabolism and other lipidemias	Vitamin D deficiency, unspecified
Hypokalemia	Disease of the blood and blood forming organs	Other long term (current) drug therapy
Cardiomegaly	Hypokalemia	Long term (current) use of anticoagulants
Hypertensive diseases		Abnormalities of breathing
Chest pain, unspecified		Disorders of lipoprotein metabolism and other lipidemias
Atherosclerotic heart disease of native coronary artery without angina pectoris		
Encounter for other preprocedural examination		
Overweight and obesity		

Table 5.8: Cohort comparison for diagnosis

5.5.2 Procedures

This section discusses the procedure's phenotypes over three years and three study cohorts, as well as the consideration given for the diagnosis. The important variables for procedures during the first year, second year, and third year prior to CKD for the three study cohorts are considered. Table 5.6 shows these variables one year prior to CKD.

One year prior to CKD, there are several common variables in cohorts 1 and 2 of patients who experienced AKI before CKD. These variables are "blood CO2 analysis," "ABO blood typing," "calcium ionized measurement," "potassium in plasma measurement," "therapeutic procedure, one or more areas, each lasting 15 minutes," and "therapeutic exercises to develop strength and endurance, range of motion, and flexibility". Once more, all three cohorts share some factors, such as 'Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count', 'Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral', 'Measurement of inorganic phosphorus (phosphate)', 'Measurement of lactate (lactic acid)' 'Prothrombin time test'.

For Cohort 1, one year prior to CKD, there are some unique procedure variables such as "drug test(s), definitive," "measurement of glucose in blood," "nonpressurized inhalation treatment for acute airway obstruction," "all-inclusive clinic visit rendered in a FQHC or CHC," Initial hospital care, per day, for the evaluation and management of a patient; qualitative drug screening; multiple drug classes by high complexity test method (e.g., immunoassay, enzyme assay); antibody screening, RBC, each serum technique; blood urea nitrogen (BUN) measurement; "Aerobic and anaerobic blood bacterial culture," "measurement of sodium in plasma," "measurement of aspartate amino transferase (AST) (SGOT)," "measurement of carbon dioxide," "radiologic examination of the chest, single view," "measurement of lactate dehydrogenase (LD)," and "blood typing; Rh (D)." These variables may be phenotypes for CKD and may have been related to hospitalization prior to AKI.

Patients in cohort 3, one year before CKD were at risk for 'Level 4 outpatient visit for established patient with problem of moderate to high severity 25 minutes', 'Level 3 outpatient visit for established patient with problem of low to moderate severity 15 minutes', 'Direct measurement of high density cholesterol', 'Emergency department visit for the evaluation and management of a patient,', 'Aerobic and anaerobic bacterial culture of blood', 'Measurement of albumin in plasma', 'Drug test(s), definitive,', 'Measurement of total calcium', 'Measurement of vitamin D 25 hydroxy', 'Measurement of thyroid stimulating hormone (TSH)', 'Hospital discharge day management; 30 minutes or less.', 'Automated complete blood cell count'.

Using similar tables available in appendix F, two and three years prior to CKD can be similarly interpreted, and phenotypes between and within study cohorts can be considered and analyzed.

Then, we looked at procedure variables generally throughout the three years previous to CKD to evaluate procedure-related phenotypes and how they evolved for three research

groups. Table 5.7's findings reveal that between cohorts 1 and 2, "measurement of lactate (lactic acid)" is the sole procedural variable in common. For all individuals who had AKI before developing CKD, this characteristic may be a phenotype for CKD. Regardless of prior AKI,'Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count','Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral',and 'Prothrombin time test' were shared by the three groups and may all be phenotypes for CKD. Cohort 1 had two distinct variables over the course of three years: "Drug test(s), definite," and "Analysis of CO2 of blood." For AKI patients who are hospitalized, these may be phenotypes. Finally, There were some specific variables for cohort 3 incloudes, such as "Measurement of Troponin," "Partial thromboplastin time test on blood," "Level 3 outpatient visit for established patient with problem of low to moderate severity," "Direct measurement of high density cholesterol," "Measurement of albumin in plasma," "Drug test(s), definitive," "Measurement of vitamin D 25 hydroxy," and "Measurement of thyroid stimulating hormone (TSH).

Cohort1	Cohort2	Cohort3
Drug test(s), definitive,	Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral	Measurement of Troponin
Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral	Non-covered item or service	Radiologic examination, chest; single view, frontal.
Measurement of lactate (lactic acid)	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report
Measurement of glucose in blood	Prothrombin time test	Prothrombin time test
Nonpressurized inhalation treatment for acute airway obstruction.	Measurement of magnesium	Level 4 outpatient visit for established patient with problem of moderate to high severity 25 minutes
Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count	Measurement of inorganic phosphorus (phosphate)	Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral
Measurement of potassium in plasma	Partial thromboplastin time test on blood	Partial thromboplastin time test on blood
All-inclusive clinic visit rendered in a FQHC or CHC.	Measurement of lactate (lactic acid)	Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count
Initial hospital care, per day, for the evaluation and management of a patient	Radiologic examination, chest; single view, frontal.	Level 3 outpatient visit for established patient with problem of low to moderate severity 15 minutes
Drug screen, qualitative; multiple drug classes by high complexity test method (e.g., immunoassay, enzyme assay), per patient	Measurement of Troponin	Non-covered item or service
Analysis of CO2 of blood	Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count	Direct measurement of high density cholesterol
Therapeutic procedure, one or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	Measurement of Calcium; ionized	Collection of venous blood by venipuncture
Measurement of inorganic phosphorus (phosphate)	Comprehensive metabolic panel	Measurement of lactate (lactic acid)
Antibody screen, RBC, each serum technique	Collection of venous blood by venipuncture	Subsequent hospital care, per day, for the evaluation and management of a patient
Blood typing; ABO	Assay of natriuretic peptide	Emergency department visit for the evaluation and management of a patient,
Measurement of blood urea nitrogen (BUN)	Analysis of CO2 of blood	Assay of natriuretic peptide
Aerobic and anaerobic bacterial culture of blood	Therapeutic procedure, one or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	Aerobic and anaerobic bacterial culture of blood
Prothrombin time test	Measurement of potassium in plasma	Measurement of albumin in plasma
Measurement of sodium in plasma	Subsequent hospital care, per day, for the evaluation and management of a patient	Comprehensive metabolic panel
Measurement of aspartate amino transferase (AST) (SGOT)	Measurement of alanine amino transferase (ALT) (SGPT)	Drug test(s), definitive,
Measurement of carbon dioxide	Smear from primary source with Giemsa stain for bacteria	Measurement of total calcium
Measurement of Calcium; ionized	Blood typing; ABO	Measurement of vitamin D 25 hydroxy
Measurement of chloride in blood		Measurement of inorganic phosphorus (phosphate)
Radiologic examination of chest, single view		Measurement of thyroid stimulating hormone (TSH)
Measurement of lactate dehydrogenase (LD), (LDH)		Measurement of magnesium
Blood typing; Rh (D)		Hospital discharge day management; 30 minutes or less.
		Automated complete blood cell count

Table 5.9: Year 1 procedure comparison

Cohort1	Cohort2	Cohort3
Drug test(s), definitive,	Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral	Measurement of Troponin
Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral	Non-covered item or service	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report
Measurement of lactate (lactic acid)	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	Prothrombin time test
Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count	Prothrombin time test	Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral
Analysis of CO2 of blood	Measurement of magnesium	Partial thromboplastin time test on blood
Prothrombin time test	Measurement of inorganic phosphorus (phosphate)	Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count
	Measurement of lactate (lactic acid)	Level 3 outpatient visit for established patient with problem of low to moderate severity 15 minutes
	Radiologic examination, chest; single view, frontal.	Non-covered item or service
	Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count	Direct measurement of high density cholesterol
	Comprehensive metabolic panel	Collection of venous blood by venipuncture
	Collection of venous blood by venipuncture	Subsequent hospital care, per day, for the evaluation and management of a patient
	Subsequent hospital care, per day, for the evaluation and management of a patient	Measurement of albumin in plasma
		Comprehensive metabolic panel
		Drug test(s), definitive,
		Measurement of vitamin D 25 hydroxy
		Measurement of thyroid stimulating hormone (TSH)

Table 5.10: Cohort comparison for procedure

5.5.3 Medications

Medicines were investigated to determine how they connected to the patient profile in the years before CKD and generally for all research cohorts, just as the previous two sets of diagnoses and procedures. Tables 5.8 and 5.9 were also created in order to compare and contrast cohorts one year prior to CKD and overall. In this study, we considered 500 different medications, and now we can observe that both Cohorts 1 and 2, one year prior to CKD, contain three significant medications. These variables are '50 ML Albumin Human', 'Furosemide', and 'Magnesium Sulfate Heptahydrate'. Also, we have 10 similar variables across all three cohorts, one year prior to CKD, which include the following: 'Ceftriaxone (as ceftriaxone sodium)', 'Docusate Sodium', 'Enoxaparin Sodium', 'Fentanyl Citrate, 'Lidocaine Hydrochloride', 'Ondansetron Hydrochloride', 'Polyethylene Glycol', 'Rocuronium Bromide', 'Sodium Chloride ', 'Sodium Chloride'. Cohort 1, as the only group of patients with hospitalization or inpatient visits, has unique variables, one year prior to CKD, such as 'Heparin Sodium', 'Ipratropium Bromide', 'Docusate Sodium', 'Dextrose Monohydrate', 'Water 100 g in 100 ml Irrigation Irrigant [veterinary sterile water for irrigation]', 'Robinul', 'Metformin Hydrochloride', 'Epinephrine;Isopropyl Alcohol', 'Bisacodyl', 'Levofloxacin[Levaquin]', 'Calcium Chloride', 'Culturelle Probiotics', 'Ondansetron', 'Product Containing Precisely Prochlorperazine Maleate'. moreover, cohort 3, as the only cohort with patients who have never experienced AKI, has unique variables, one year prior to CKD, like 'Dexamethasone Sodium Phosphate', 'Robinul', 'Ephedrine Sulfate ', 'Amoxicillin', 'Lorazepam', 'Trazodone Hydrochloride', 'Donepezil Hydrochloride', 'Nicotine', 'Oxycodone Hydrochloride', 'Hydrocodone Bitartrate',

'Docusate Sodium', "Heparin Sodium', 'Prednisone[Deltasone]'.

Similarity and dissimilarity between various cohorts two and three years prior to CKD, as well as one year prior to CKD, are taken into account using similar tables in the appendix F.

Then we compared three cohorts over the three years preceding CKD, and the only similarity between cohorts 1 and 2 was 'Magnesium Sulfate Heptahydrate', which belonged to the year preceding CKD. Similarly, for three years, eight medications were shared among three cohorts. Those variables are 'Ceftriaxone (as ceftriaxone sodium) 1 g powder for solution for injection vial', 'Docusate Sodium', 'Enoxaparin Sodium', 'Fentanyl Citrate', 'Lidocaine Hydrochloride', 'Ondansetron Hydrochloride', 'Sodium Chloride', 'Sodium Chloride'.

The medications that were unique in cohort 1 during three years prior to CKD were 'Heparin Sodium', 'Ipratropium Bromide', 'Furosemide', 'Polyethylene Glycol'. These medications can be phenotypes related to hospitalization or an inpatient visit prior to AKI. 'Dexamethasone Sodium Phosphate', 'Hydrocodone Bitartrate', 'Docusate Sodium', 'Heparin Sodium' were specific medications taken by cohort 3 over the course of three years that could be considered phenotypes but were not linked to AKI.

Cohort1	Cohort2	Cohort3
Sodium Chloride	Fentanyl Citrate	Sodium Chloride
Ondansetron Hydrochloride	Sodium Chloride	Lidocaine Hydrochloride
Fentanyl Citrate	Acetaminophen	Dexamethasone Sodium Phosphate
Heparin Sodium	Ondansetron Hydrochloride[ondansetron]	Fentanyl Citrate
Ipratropium Bromide	Sodium Chloride	Robinul
Ceftriaxone (as Ceftriaxone Codium)	Ceftriaxone (as Ceftriaxone Sodium)	Ephedrine Sulfate [Premierpro Rx Ephedrine Sulfate]
Docusate Sodium	Polyethylene glycol	Amoxicillin
Dextrose Monohydrate	Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water]#4	Rocuronium Bromide
Docusate Sodium	Docusate Sodium[senna-s]	Acetaminophen
Furosemide	Furosemide	Sodium Chloride
Polyethylene Glycol	50 ml Albumin Human	Lorazepam
Sodium Chloride	Rocuronium Bromide	Ondansetron Hydrochloride [Ondansetron]
Magnesium Sulfate Heptahydrate	Sodium Chloride	Trazodone Hydrochloride
Enoxaparin Sodium	Famotidine	Ceftriaxone (as Ceftriaxone Sodium)
Lidocaine Hydrochloride	2 ml Sugammadex	Donepezil Hydrochloride [Donepezil Hydrochloride]
Rocuronium Bromide	Chlorhexidine Gluconate	Nicotine [Nicotine Transdermal System Step 1]
Water/ Irrigation Irrigant [veterinary sterile water for irrigation]	Flagyl	Oxycodone Hydrochloride
Robinul	Enoxaparin Sodium	Hydrocodone Bitartrate
50 ml Albumin Human	Lidocaine Hydrochloride	Docusate Sodium
Metformin Hydrochloride		Heparin Sodium
Epinephrine; Isopropyl Alcohol		Prednisone [Deltasone]
Bisacodyl		Docusate Sodium
Levofloxacin		Enoxaparin Sodium
Calcium Chloride		Polyethylene Glycol
Culturelle Probiotics 15 Billion CFU Capsule		
Ondansetron [Zofran]		
Product Containing Precisely Prochlorperazine Maleate		

Table 5.11: Year 1 medication comparison

Cohort1	Cohort2	Cohort3
Sodium Chloride	Fentanyl Citrate	Sodium Chloride[Veterinary Lactated]
Ondansetron Hydrochloride [ondansetron]	Sodium Chloride [veterinary lactated]	Lidocaine Hydrochloride
Fentanyl Citrate	Acetaminophen	Dexamethasone Sodium Phosphate
Heparin Sodium	Ondansetron Hydrochloride [ondansetron]	Fentanyl Citrate
Ipratropium Bromide	Sodium Chloride	Acetaminophen
Ceftriaxone (as Ceftriaxone Sodium)	Ceftriaxone (as Ceftriaxone Sodium)	Sodium Chloride
Docusate Sodium	Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water] #4	Ondansetron Hydrochloride [Ondansetron]
Furosemide	Docusate Sodium	Ceftriaxone (as Ceftriaxone Sodium)
Polyethylene Glycol	Rocuronium Bromide	Hydrocodone Bitartrate
Sodium Chloride [Veterinary Lactated]	Enoxaparin Sodium	Docusate Sodium
Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water] #4	Lidocaine Hydrochloride	Heparin Sodium
Enoxaparin Sodium		Docusate Sodium
Lidocaine Hydrochloride		Enoxaparin Sodium

Table 5.12: Cohort comparison for medications

5.5.4 Lab results

In this study, lab data were also taken into account to identify individuals who had high levels of anomalies across several lab identifiers. Finding risk variables in this section is crucial since lab tests are one of the most crucial diagnostic tools for CKD. As part of our work, we gathered important lab variables during each year for each cohort. Tables 5.10 shows this information for one year prior to CKD. Similar tables are available for the two and three years prior to CKD in the appendix F. Then we compared different cohorts during different years prior to CKD based on the similarity and dissimilarity of variables between cohorts.

For one year prior to CKD, based on table 5.10, we see that there is no similarity between cohorts 1 and 2 in the year 1, which we could relate to patients with prior AKI only. However, we see the similarity between three cohorts during the year prior to CKD. These similarities are due to common variables such as 'Alkaline phosphatase Enzymatic.activity volume in Serum Plasma or Blood','Bilirubin total Mass volume in Serum Plasma or Blood','Folate Mass volume in Serum Plasma or Blood' ,'Glucose', 'Hematocrit' ,'Hemoglobin A1c Hemoglobin total in Blood' ,'Hepatitis B virus surface Ab Units volume in Serum' ,'Platelets' ,'Urea nitrogen Mass volume in Serum Plasma or Blood' , and 'White blood cell Leukocytes volume in Blood'. Additionally, there were no specific lab tests for Cohort 1 to demonstrate if it might be connected to inpatient stays or hospitalizations. Cohort 3 was the only cohort without AKI, and it only had one distinctive variable—protein mass volume in serum or plasma, one year prior to CKD.

For cohort comparison during the three years, as table 5.11 shows, the only similarity between cohorts 1 and 2 was "glucose," which can be related to AKI prior to CKD. Among

the three cohorts, we see that ”hematocrit,” ”platelets,” ”urea nitrogen mass volume in serum plasma or blood,” and ”white blood cell leukocyte volume in blood” were common during the three years prior to CKD. These factors can be phenotypes for CKD without being related to AKI. ”Bilirubin total mass volume in serum plasma or blood,” ”Alkaline phosphatase enzymatic activity volume in serum plasma or blood,” ”Folate mass volume in serum plasma or blood,” and ”Hemoglobin A1c hemoglobin total in blood” are unique variables in cohort 1, which may be related to hospitalization and inpatient visits, and Alkaline phosphatase Enzymatic activity volume in serum, plasma, or blood, hemoglobin A1c, hemoglobin total in blood, and protein mass volume in serum or plasma were unique variables in cohort 3, the only cohort without prior AKI.

Cohort1	Cohort2	Cohort3
White blood cell Leukocytes volume in Blood	White blood cell Leukocytes volume in Blood	Platelets
Hepatitis B virus surface Ab Units volume in Serum	Hematocrit	Urea nitrogen Mass volume in Serum Plasma or Blood
Platelets	Urea nitrogen Mass volume in Serum Plasma or Blood	Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood
Bilirubin total Mass volume in Serum Plasma or Blood	Platelets	Hemoglobin A1c Hemoglobin total in Blood
Glucose	Hemoglobin A1c Hemoglobin total in Blood	Glucose
Hematocrit	Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood	Hepatitis B virus surface Ab Units volume in Serum
Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood	Folate Mass volume in Serum Plasma or Blood	Hematocrit
Urea nitrogen Mass volume in Serum Plasma or Blood	Glucose	Bilirubin total Mass volume in Serum Plasma or Blood
Folate Mass volume in Serum Plasma or Blood	Hepatitis B virus surface Ab Units volume in Serum	White blood cell Leukocytes volume in Blood
Hemoglobin A1c Hemoglobin total in Blood	Bilirubin total Mass volume in Serum Plasma or Blood	Protein Mass volume in Serum or Plasma
		Folate Mass volume in Serum Plasma or Blood

Table 5.13: Year 1 lab comparison

Cohort1	Cohort2	Cohort3
White blood cell Leukocytes volume in Blood	White blood cell Leukocytes volume in Blood	Platelets
Platelets	Hematocrit	Urea nitrogen Mass volume in Serum Plasma or Blood
Bilirubin total Mass volume in Serum Plasma or Blood	Urea nitrogen Mass volume in Serum Plasma or Blood	Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood
Glucose	Platelets	Hemoglobin A1c Hemoglobin total in Blood
Hematocrit	Glucose	Hematocrit
Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood	Hepatitis B virus surface Ab Units volume in Serum	White blood cell Leukocytes volume in Blood
Urea nitrogen Mass volume in Serum Plasma or Blood		Protein Mass volume in Serum or Plasma
Folate Mass volume in Serum Plasma or Blood		
Hemoglobin A1c Hemoglobin total in Blood		

Table 5.14: Cohort comparison for lab

5.5.5 Vital Signs

This section examines whether or not particular vital signs were crucial for the three research groups in the 1, 2, and 3 years prior to CKD. We identified five vital sign variables, and as tables 5.20, 5.21, and 5.22 demonstrate, they were significant criteria in the three years before CKD because patients had abnormal values for these criteria. We may argue that these factors were equally important criterion for all cohorts because they occurred for all cohorts throughout all years.

There was no specific vital sign between cohorts 1 and 2 related to AKI; however, there were three common vital signs between the three cohorts during the three years in the cloud: ”blood pressure systolic,” ”body temperature,” and ”heart rate.” related to CKD only. BMI for cohort 2 and blood pressure diastolic for cohort 1 were not phenotypes.

Cohort1	Cohort2	Cohort3
Body temperature	Heart rate	Body temperature
BMI	Blood Pressure Diastolic	Blood Pressure Diastolic
Blood Pressure Diastolic	Body temperature	Blood Pressure Systolic
Blood Pressure Systolic	Blood Pressure Systolic	BMI
Heart rate	BMI	Heart rate

Table 5.15: Year 1 vital sign comparison

Cohort1	Cohort2	Cohort3
Body temperature	Heart rate	Body temperature
BMI	Blood Pressure Diastolic	Blood Pressure Diastolic
Blood Pressure Systolic	Body temperature	Blood Pressure Systolic
Heart rate	Blood Pressure Systolic	BMI
		Heart rate

Table 5.16: Cohort comparison for vital signs

6 Discussion and Conclusions

Chapter Contents

4.1 Clustering 19
4.1.1 Dimensional reduction and variable selection 20
4.1.2 Logistic PCA 20
4.1.3 K-means clustering 21
4.1.4 Latent class analysis (LCA) 22
4.1.5 Random forest 23
4.2 Outcome measures 24
4.3 Similarity measures 25

6.1 Discussion

A total of 75033 CKD patients were considered. Patients with AKI after CKD are excluded from this study. Otherwise, they had been classified into three study cohorts with different clinical characteristics. Cohort 1 includes 7442 patients who had AKI within 90 days of hospitalization prior to 3 years of CKD. Cohort 2 includes 6408 patients those with AKI who were not hospitalized within 90 days prior to 3 years of CKD, as well as those with AKI prior to 5 years of CKD who were not in Cohort 1, and cohort 3 with 39280 CKD patients with never experienced AKI. The clinical characteristics of patients in cohorts 1 and 2 and cohort 3 had been considered in terms of identifying the effects of AKI and hospitalization on CKD. Age and BMI, white ethnicity, as well as comorbidities listed in Table 5.1, are considered significant variables. Gender is considered significant for Cohort 1.

Variables are chosen in two steps: the first is based on clinical expert opinion, and the second is based on data mining methodology. A clinical expert selected the diagnosis, procedures, medications, lab results, and vital signs. Top variables were selected from the datasets, and then, using data mining methods, the most important variables were selected based on MDA and the prevalence of the variables.

K-means and LCA methods were selected to group patients in each cohort, year by year, during the 3 years before CKD. K-means was selected as one of the most widely used clustering algorithms because it is fast for big data and easy to understand. LCA, as a model-based method, is more interpretable and preferred in healthcare analysis and was selected to apply to the datasets. Patients' assignment by each clustering method was compared. As a result, K-means clusters were identified pretty close to LCA clusters. The LCA clusters are chosen for data analysis because it is considered to be more reliable method compared to the K-means for generating meaningful clusters from high dimensional binary data.

The clustering of patients showed different phenotypes in each cluster for each year within a cohort with the same clinical characteristics. which shows the importance of clustering. The high similarity between phenotypes in each cluster prevents us from naming the clusters.

Similarity between years is calculated to determine whether patients in a cohort change their groups over time. Because of the low similarity between patient groups, we decided to compare patient profiles in three cohorts on a yearly basis. Variable similarity and dissimilarity between cohorts are evaluated and interpreted on a yearly basis based on their health characteristics. Similarly, comparisons of variables between cohorts had been considered and interpreted.

6.2 Strengths of the study

This research analyzed high-dimensional clinical data for three large cohorts of patients with different health characteristics. Phenotypes for a patient were studied for years, even up to three years before CKD, to see how they changed over time. Also, the roles of diagnosis, procedures, medication, lab results, and vital signs have been considered separately.

6.3 Limitation of the study

The absence of non-CKD patients prevents us from developing an algorithm for predicting health outcomes in the CKD and non-CKD groups. Also, as this study was done on CKD patients in West Virginia and thus limited to a particular geographic region, it may not be generalized to all CKD patients.

6.4 Contributions

A clustering-based patient profiling system has been developed where we can find time-based phenotypes for a patient and use them to understand the patient's journey from AKI to CKD. Then, two existing methods were compared to determine which was superior. The results show there are numerous phenotypes that may impact or contribute to CKD.

6.5 Future works

One of our research objectives was to develop AI tools, but we could not do that because of the scale of the data we had. Applying AI to data with a large number of variables will take much more time, which is beyond the scope of this thesis. AI tools will be developed in the future and validated by the clustering results obtained by this research.

Important phenotypes obtained from this study can be used in the future to predict CKD patients in various study cohorts generally and across years differently based on diagnosis, procedures, medications, lab results, and vital signs.

6.6 Conclusions

The findings from clustering suggests that phenotypes are different for different classes of patients, so a single care method doesn't work for all patients. Categorizing patients into distinct groups allows for the allocation of different resources and strategies for the care of different groups of patients.

According to the results of the Jaccard similarity test, patient profiles change in the years before developing CKD; hence, research cohorts need to take each year into account in order to identify the phenotypes.

Three research cohorts' key factors were compared annually and generally, and the results revealed some similarities as well as some differences across the cohorts for each year. These findings suggest that specific CKD phenotypes may be linked to prior AKI, hospitalization, inpatient care, and CKD regardless of prior AKI.

References

(1) KO, S.; CO, O.; O, E. O. *African journal of food science* **2010**, *4*, 200–222.

(2) `https://www.kidney.org/atoz/content/AcuteKidneyInjury`.

(3) Vaidya, V. S.; Ferguson, M. A.; Bonventre, J. V. *Annual review of pharmacology and toxicology* **2008**, *48*, 463.

(4) Anderson, S.; Eldadah, B.; Halter, J. B.; Hazzard, W. R.; Himmelfarb, J.; Horne, F. M.; Kimmel, P. L.; Molitoris, B. A.; Murthy, M.; O'Hare, A. M., et al. *Journal of the American Society of Nephrology* **2011**, *22*, 28–38.

(5) Haines, R. W.; Powell-Tuck, J.; Leonard, H.; Crichton, S.; Ostermann, M. *Scientific reports* **2021**, *11*, 1–9.

(6) James, M. T.; Bhatt, M.; Pannu, N.; Tonelli, M. *Nature Reviews Nephrology* **2020**, *16*, 193–205.

(7) Gameiro, J.; Marques, F.; Lopes, J. A. *Clinical Kidney Journal* **2021**, *14*, 789–804.

(8) Wahl, T. S.; Graham, L. A.; Morris, M. S.; Richman, J. S.; Hollis, R. H.; Jones, C. E.; Itani, K. M.; Wagner, T. H.; Mull, H. J.; Whittle, J. C., et al. *JAMA surgery* **2018**, *153*, e182009–e182009.

(9) Brothers, T. N.; Strock, J.; LeMasters, T. J.; Pawasauskas, J.; Reed, R. C.; Al-Mamun, M. A. *SAGE Open Medicine* **2022**, *10*, 20503121221099359.

(10) Fiorentino, M.; Grandaliano, G.; Gesualdo, L.; Castellano, G. In *Acute Kidney Injury-Basic Research and Clinical Practice*; Karger Publishers: 2018; Vol. 193, pp 45–54.

(11) Ronco, C.; Kaushik, M.; Valle, R.; Aspromonte, N.; Peacock IV, W. F. In *Seminars in nephrology*, 2012; Vol. 32, pp 129–141.

(12) Goldberg, R.; Dennen, P. *Advances in chronic kidney disease* **2008**, *15*, 297–307.

(13) Kellum, J. A.; Sileanu, F. E.; Bihorac, A.; Hoste, E. A.; Chawla, L. S. *American journal of respiratory and critical care medicine* **2017**, *195*, 784–791.

(14) Sarnak, M. J.; Levey, A. S. *American journal of kidney diseases* **2000**, *35*, S117–S131.

(15) Hsu, R. K.; Hsu, C.-y. In *Seminars in nephrology*, 2016; Vol. 36, pp 283–292.

(16) Wang, V.; Vilme, H.; Maciejewski, M. L.; Boulware, L. E. In *Seminars in nephrology*, 2016; Vol. 36, pp 319–330.

(17) Manns, B.; Hemmelgarn, B.; Tonelli, M.; Au, F.; So, H.; Weaver, R.; Quinn, A. E.; Klarenbach, S.; Solutions, C. S.; to Overcome Chronic Kidney Disease, I. *Canadian Journal of Kidney Health and Disease* **2019**, *6*, 2054358119835521.

(18) Gameiro, J.; Branco, T.; Lopes, J. A. *Journal of clinical medicine* **2020**, *9*, 678.

(19) Malhotra, R.; Kashani, K. B.; Macedo, E.; Kim, J.; Bouchard, J.; Wynn, S.; Li, G.; Ohno-Machado, L.; Mehta, R. *Nephrology Dialysis Transplantation* **2017**, *32*, 814–822.

(20) Harrois, A.; Soyer, B.; Gauss, T.; Hamada, S.; Raux, M.; Duranteau, J. *Critical Care* **2018**, *22*, 1–10.

(21) Islam, M. A.; Akter, S.; Hossen, M. S.; Keya, S. A.; Tisha, S. A.; Hossain, S. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp 952–957.

(22) Kazancioğlu, R. *Kidney international supplements* **2013**, *3*, 368–371.

(23) Schanstra, J. P.; Zürbig, P.; Alkhalaf, A.; Argiles, A.; Bakker, S. J.; Beige, J.; Bilo, H. J.; Chatzikyrkou, C.; Dakna, M.; Dawson, J., et al. *Journal of the American Society of Nephrology* **2015**, *26*, 1999–2010.

(24) Dunkler, D.; Gao, P.; Lee, S. F.; Heinze, G.; Clase, C. M.; Tobe, S.; Teo, K. K.; Gerstein, H.; Mann, J. F.; Oberbauer, R. *Clinical Journal of the American Society of Nephrology* **2015**, *10*, 1371–1379.

(25) Roy, J.; Shou, H.; Xie, D.; Hsu, J. Y.; Yang, W.; Anderson, A. H.; Landis, J. R.; Jepson, C.; He, J.; Liu, K. D., et al. *Clinical Journal of the American Society of Nephrology* **2017**, *12*, 1010–1017.

(26) Belayev, L. Y.; Palevsky, P. M. *Current opinion in nephrology and hypertension* **2014**, *23*, 149.

(27) Hukriede, N. A.; Soranno, D. E.; Sander, V.; Perreau, T.; Starr, M. C.; Yuen, P. S.; Siskind, L. J.; Hutchens, M. P.; Davidson, A. J.; Burmeister, D. M., et al. *Nature Reviews Nephrology* **2022**, *18*, 277–293.

(28) Sawhney, S.; Tan, Z.; Black, C.; Marks, A.; Mclernon, D. J.; Ronksley, P.; James, M. T. *American Journal of Kidney Diseases* **2021**, *78*, 28–37.

(29) Jothi, N.; Husain, W., et al. *Procedia computer science* **2015**, *72*, 306–313.

(30) Veloso, R.; Portela, F.; Santos, M. F.; Silva, A.; Rua, F.; Abelha, A.; Machado, J. *Procedia Technology* **2014**, *16*, 1307–1316.

(31) Liao, M.; Li, Y.; Kianifard, F.; Obi, E.; Arcona, S. *BMC nephrology* **2016**, *17*, 1–14.

(32) Grant, R. W.; McCloskey, J.; Hatfield, M.; Uratsu, C.; Ralston, J. D.; Bayliss, E.; Kennedy, C. J. *JAMA network open* **2020**, *3*, e2029068–e2029068.

(33) Kaufman, L.; Rousseeuw, P. J., *Finding groups in data: an introduction to cluster analysis*; John Wiley & Sons: 2009.

(34) Rokach, L.; Maimon, O. In *Data mining and knowledge discovery handbook*; Springer: 2005, pp 321–352.

(35) Egan, B. M.; Sutherland, S. E.; Tilkemeier, P. L.; Davis, R. A.; Rutledge, V.; Sinopoli, A. *PloS one* **2019**, *14*, e0217696.

(36) Yoo, I.; Hu, X. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 2006, pp 220–229.

(37) Salvatore, S.; Domanska, D.; Wood, M.; Nordeng, H.; Sandve, G. K. *PLoS One* **2017**, *12*, e0190101.

(38) Everitt, B. S.; Landau, S.; Leese, M.; Stahl, D. *Cluster analysis* **2011**, *5*, 71–110.

(39) Ahmad, P.; Qamar, S.; Rizvi, S. Q. A. *International Journal of Computer Applications* **2015**, *120*.

(40) `https://cran.r-project.org/web/packages/logisticPCA/vignettes/logisticPCA.html`.

(41) Collins, M.; Dasgupta, S.; Schapire, R. E. In *NIPS*, 2001.

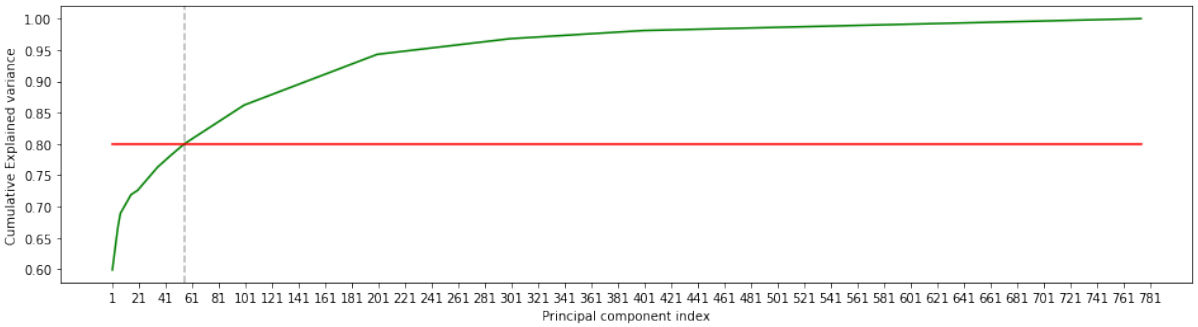
(42) Landgraf, A. J.; Lee, Y. *Journal of Multivariate Analysis* **2020**, *180*, 104668.

Appendix

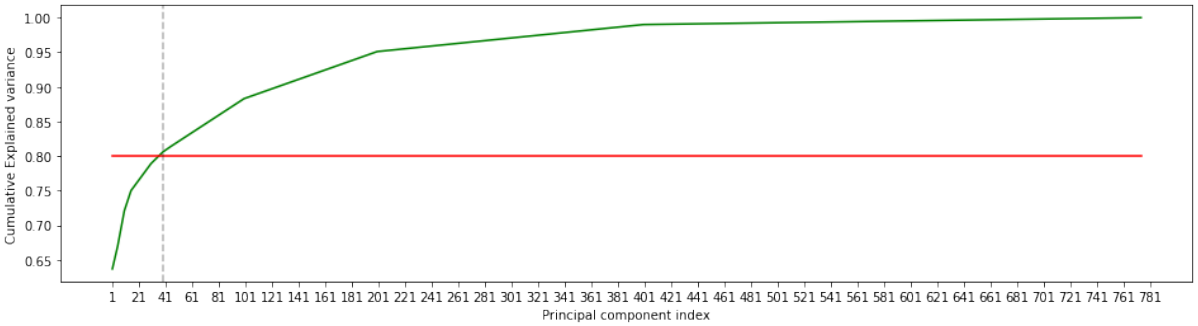
A

The number of principal components needed for logistic PCA to keep 80% of the variance of the original dataset.

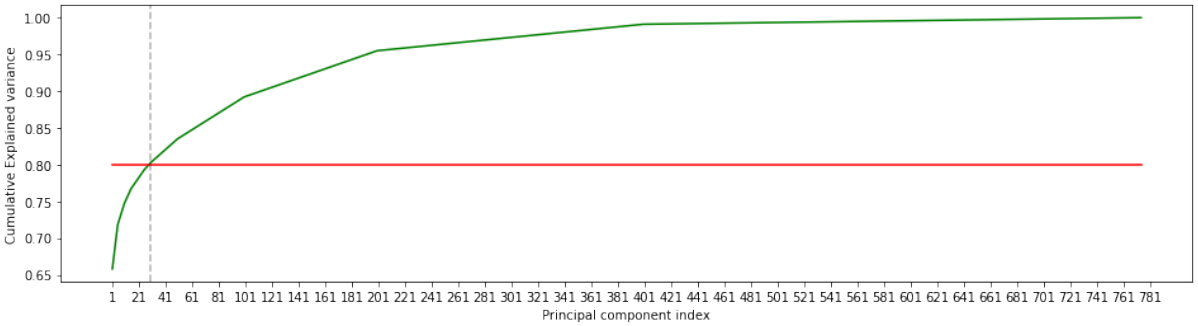
a. Cohort 2



(a)



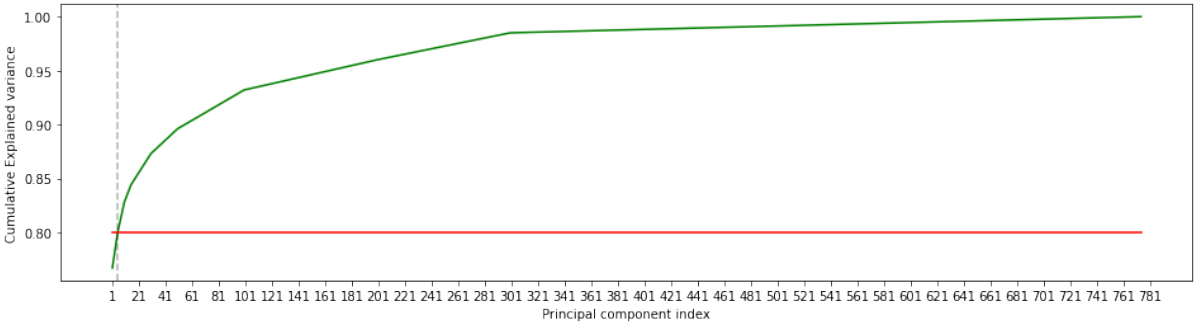
(b)



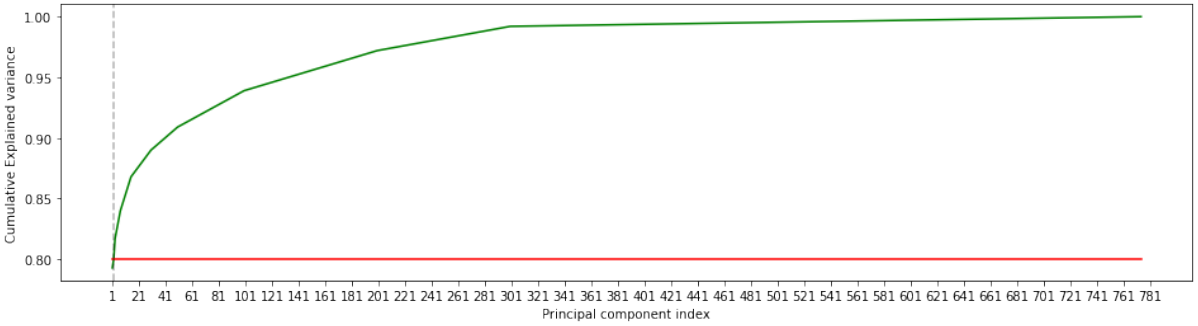
(c)

Figure A.1: The number of principal components for cohort 2. (a) year 1; (b) year 2; (c) year 3 .

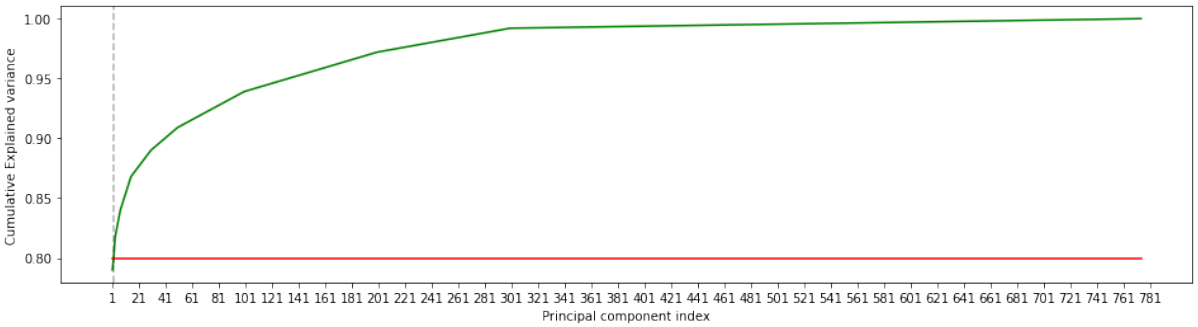
b. Cohort 3



(a)



(b)

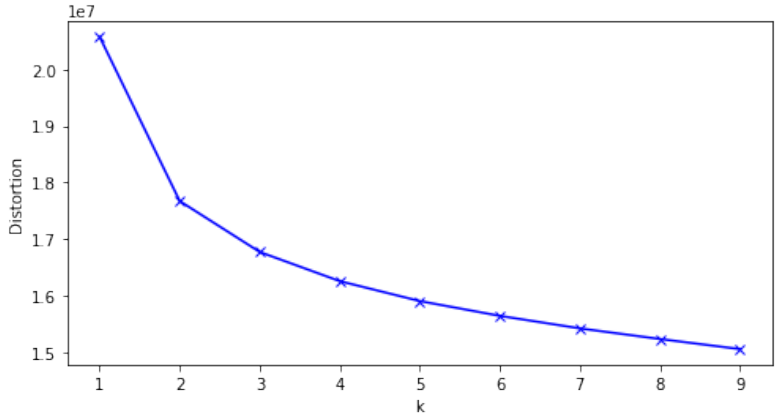


(c)

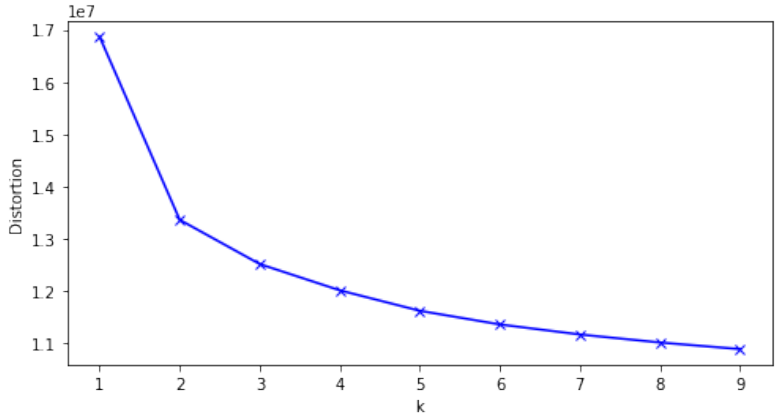
Figure A.2: The number of principal components for cohort 3. (a) year 1; (b) year 2; (c) year 3 .

The elbow method demonstrates the best number of clusters (k) for K-means.

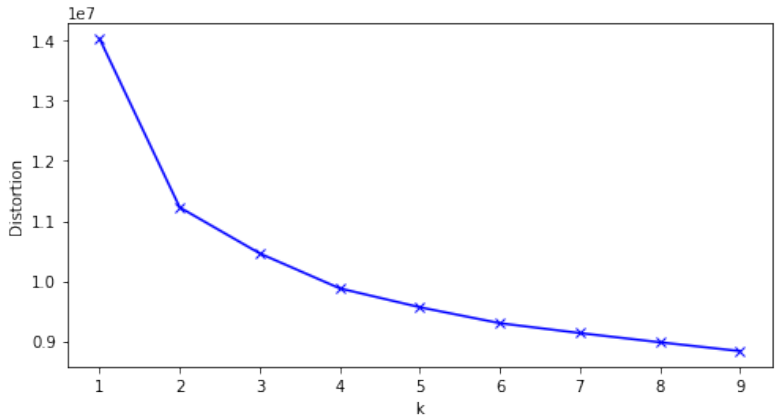
a. Cohort 2



(a)



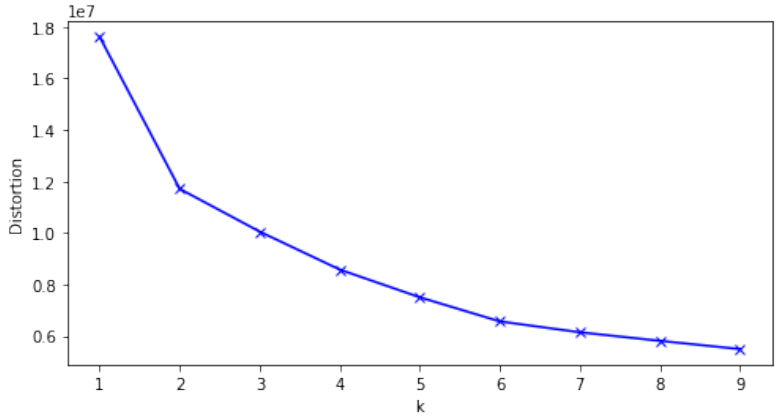
(b)



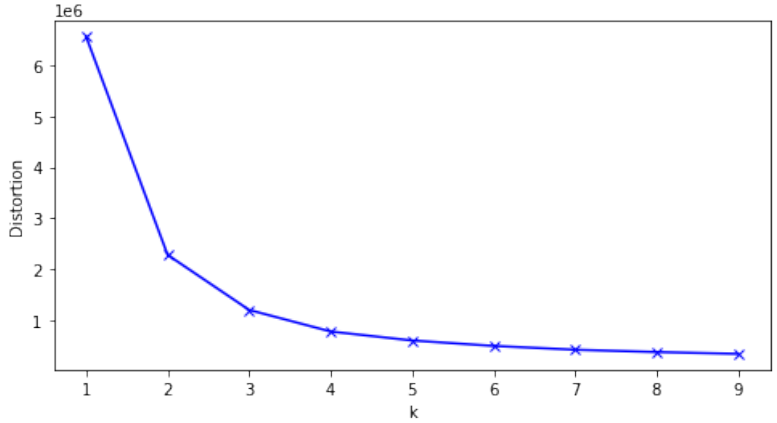
(c)

Figure B.1: The elbow method to find the number of clusters for cohort 2. (a) year 1; (b) year 2; (c) year 3 .

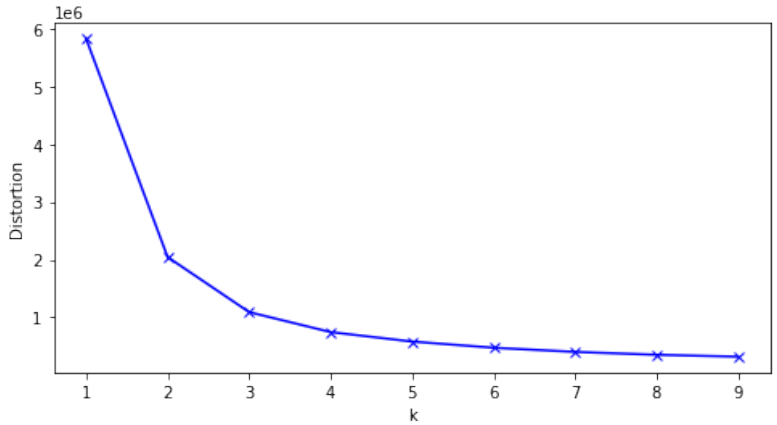
b. Cohort 3



(a)



(b)

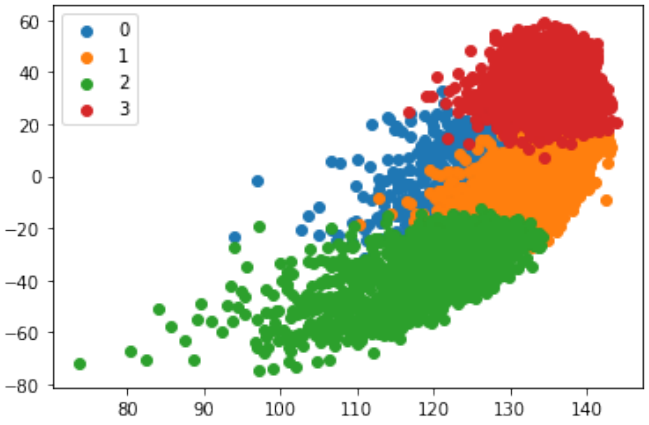


(c)

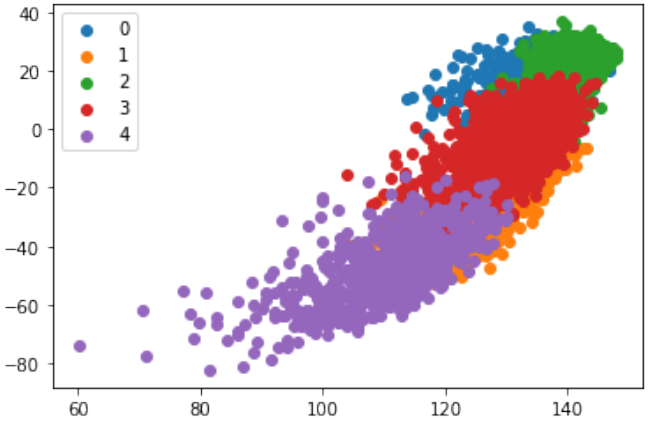
Figure B.2: The elbow method to find the number of clusters for cohort 3. (a) year 1; (b) year 2; (c) year 3 .

Visualizing clusters of K-means

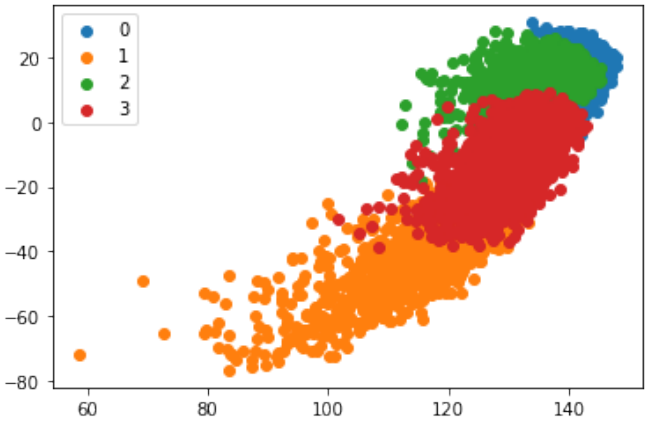
a. Cohort 2



(a)



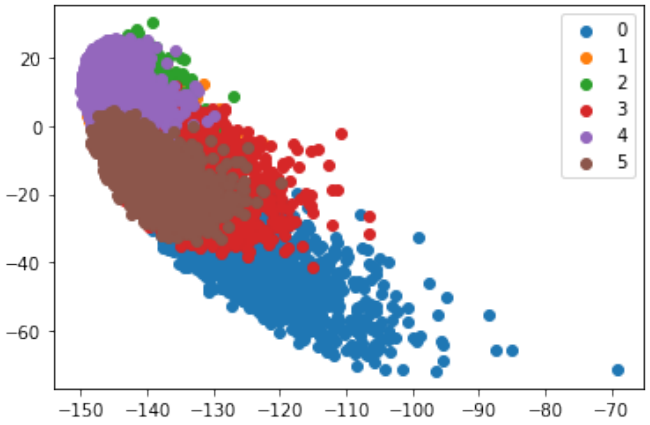
(b)



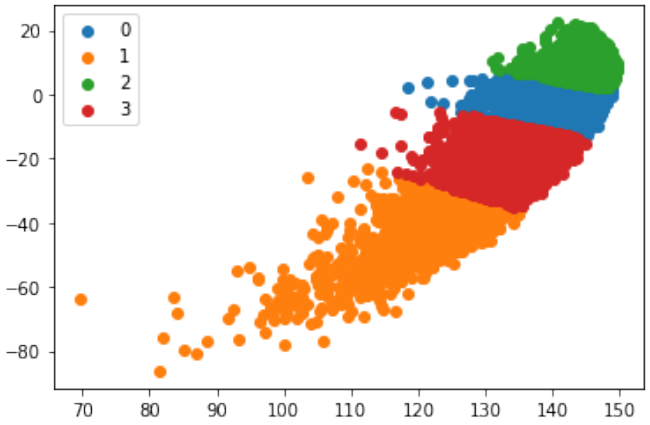
(c)

Figure C.1: Clusters of K-means for cohort 2. (a) year 1; (b) year 2; (c) year 3 .
To visualize K-means clustering, data points are projected on two axes; the values on the x and y axes are connected to these two PCs and do not exhibit a distinctive variable.

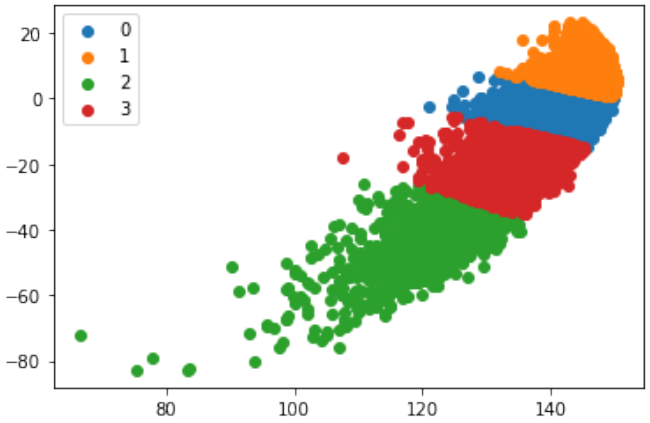
b. Cohort 3



(a)



(b)



(c)

Figure C.2: Clusters of K-means for cohort 3. (a) year 1; (b) year 2; (c) year 3.
To visualize K-means clustering, data points are projected on two axes; the values on the x and y axes are connected to these two PCs and do not exhibit a distinctive variable.

Comparison of k-means clustering with latent class analysis(LCA)

a. Cohort 2

1. Year 1 Cohort 2

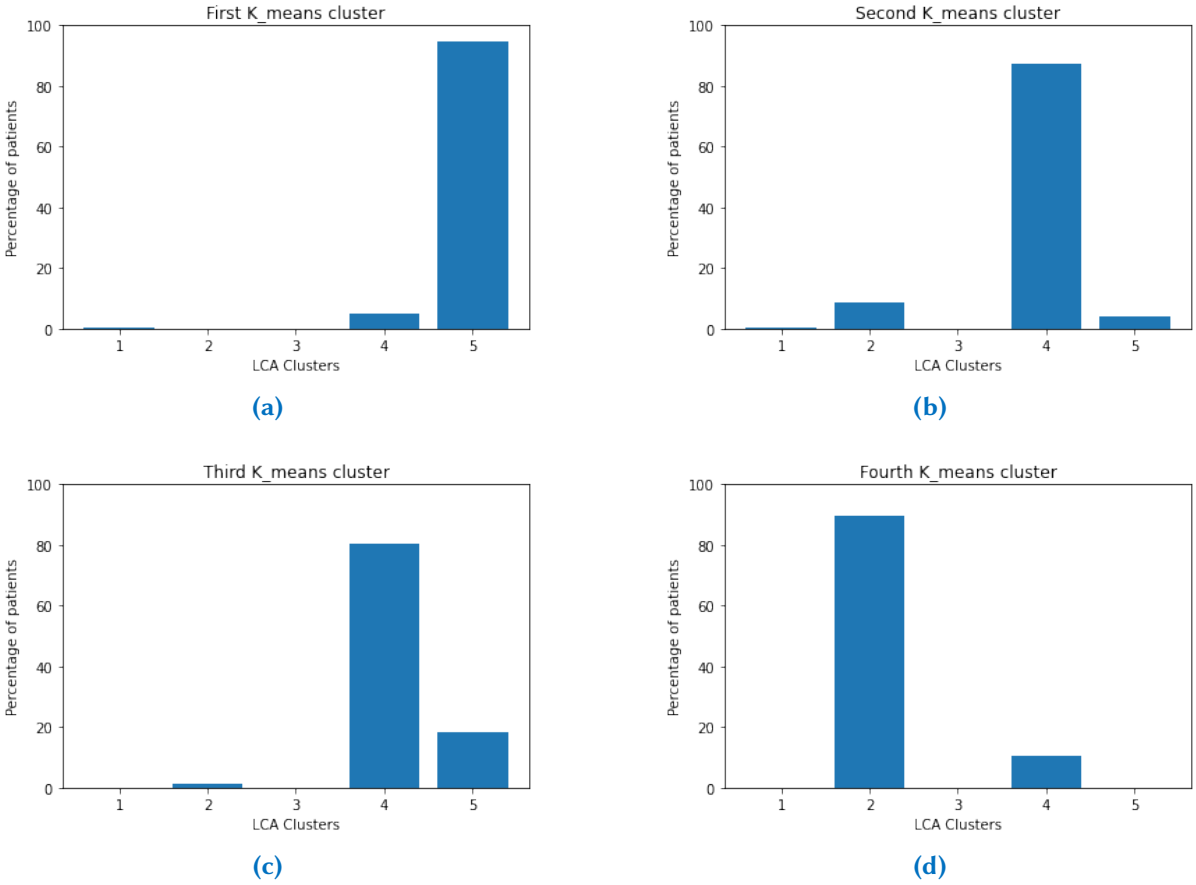


Figure D.1: Comparison of K-means clustering with latent class analysis(LCA) in year 1 cohort 2

2. Year 2 Cohort 2

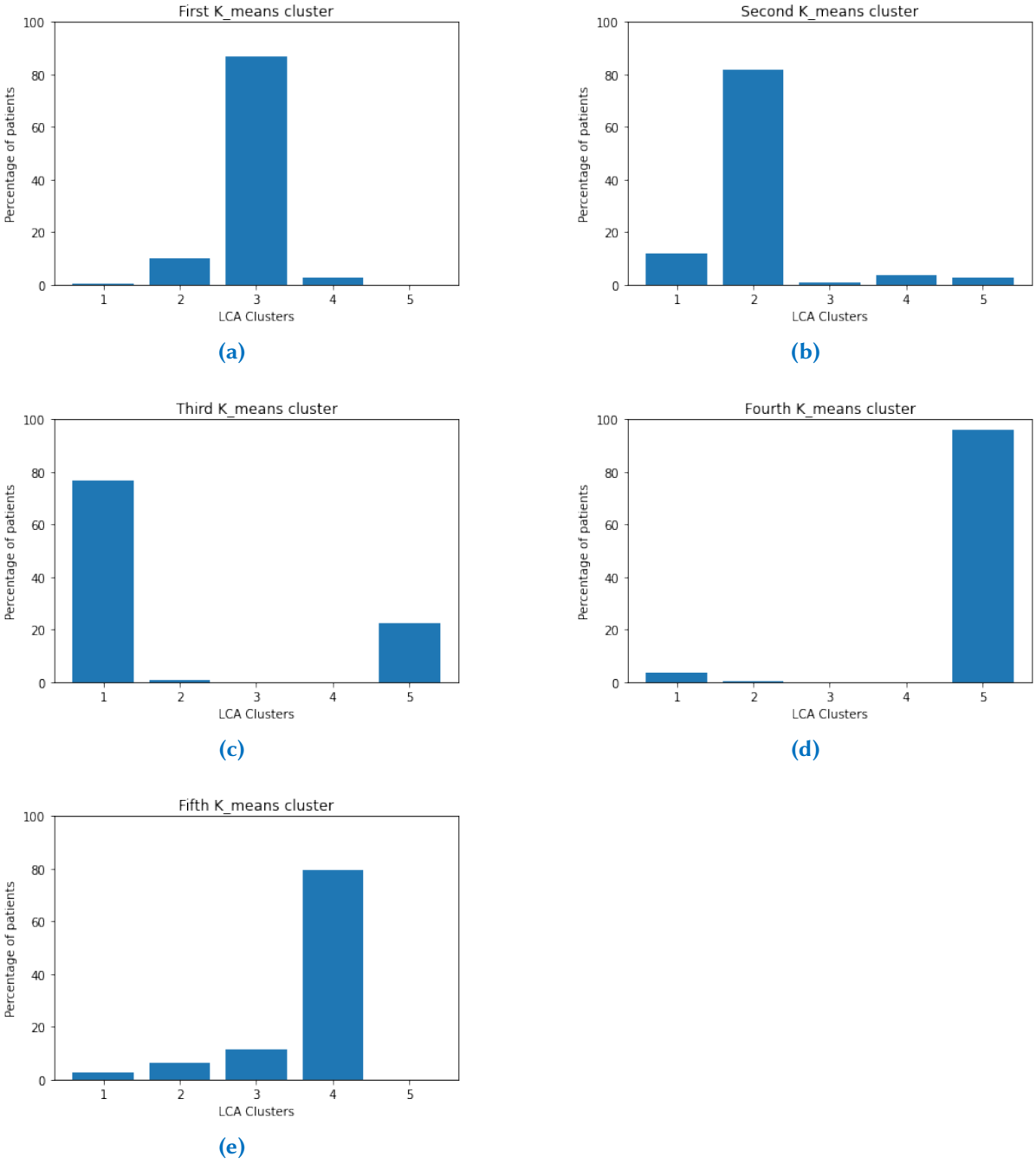


Figure D.2: Comparison of K-means clustering with latent class analysis(LCA) in year 2 cohort 2

3. Year 3 Cohort 2

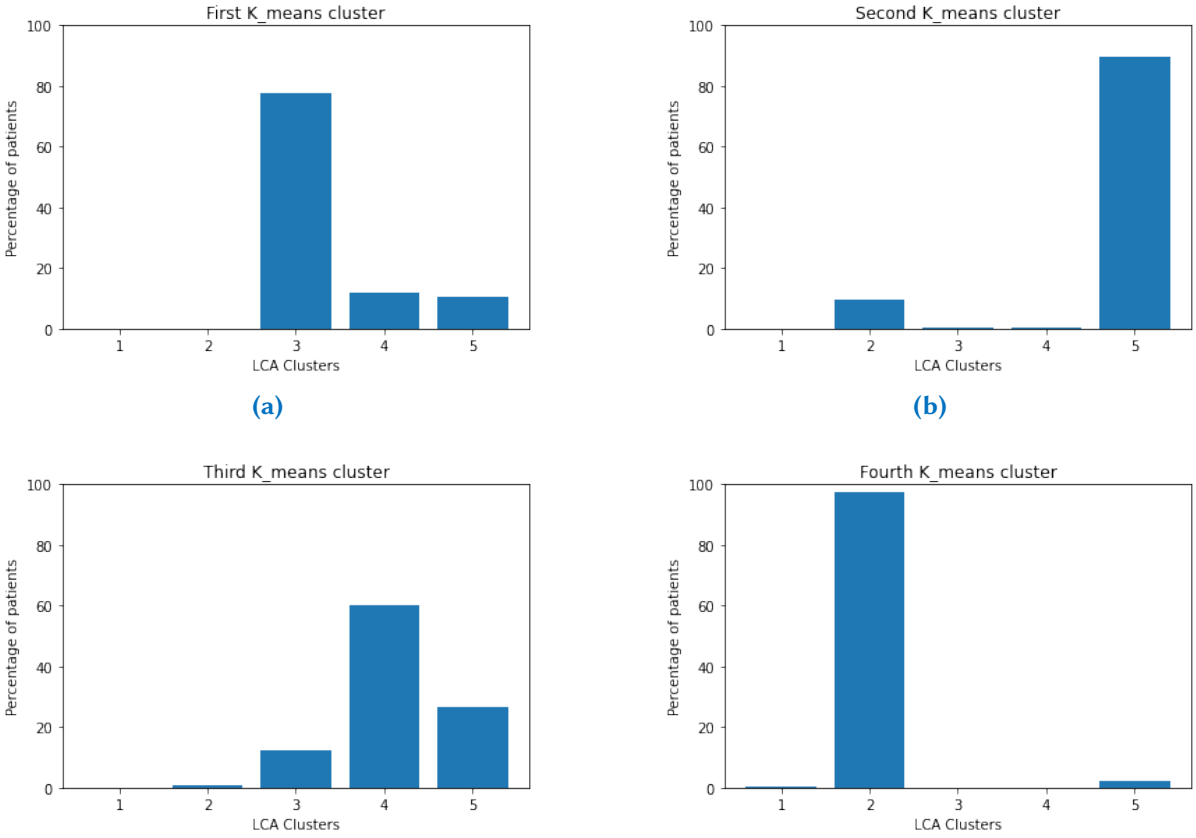


Figure D.3: Comparison of K-means clustering with latent class analysis(LCA) in year 3 cohort 2

b. Cohort 3

1. Year 1 Cohort 3

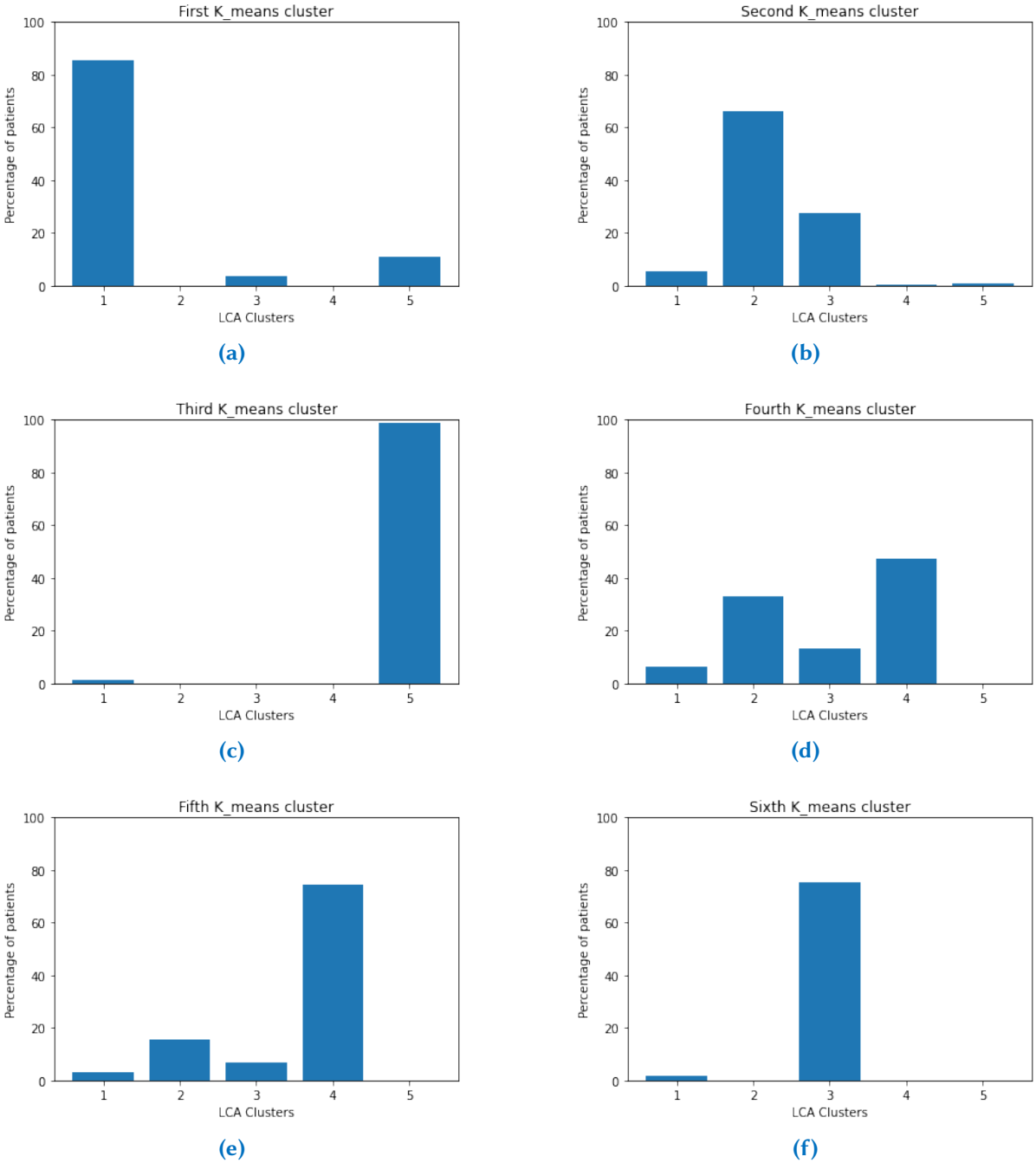
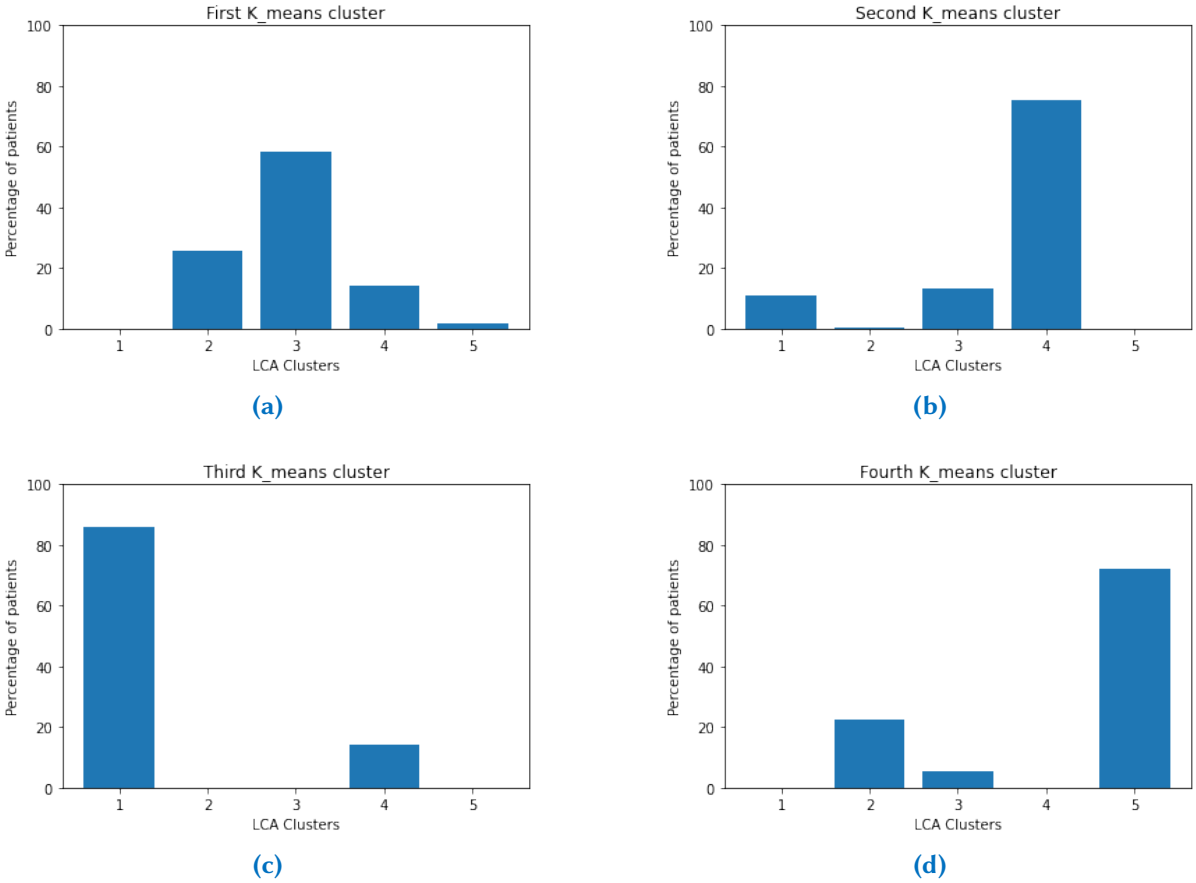


Figure D.4: Comparison of K-means clustering with latent class analysis(LCA) in year 1 cohort 3

2. Year 2 Cohort 3



(a) (b) (c) (d)

Figure D.5: Comparison of K-means clustering with latent class analysis(LCA) in year 2 cohort 3

3. Year 3 Cohort 3

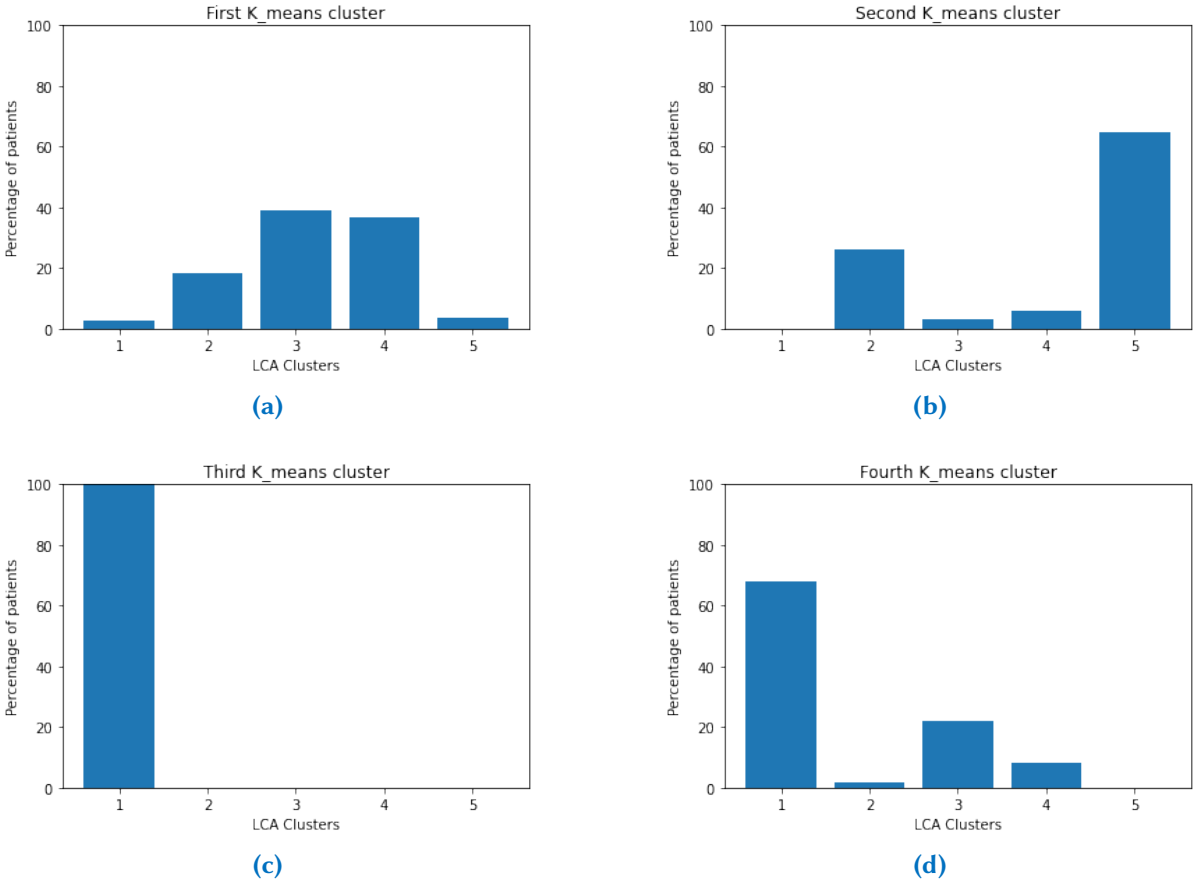
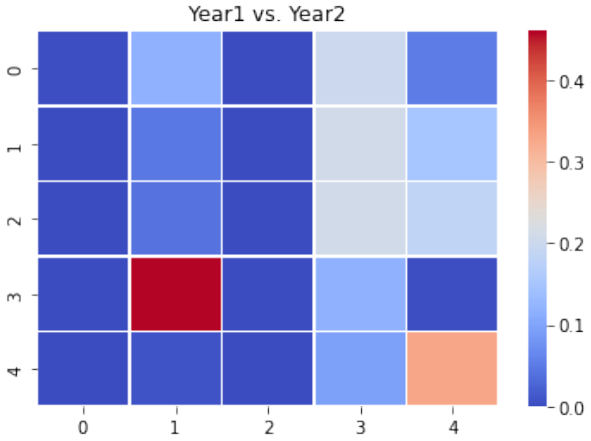


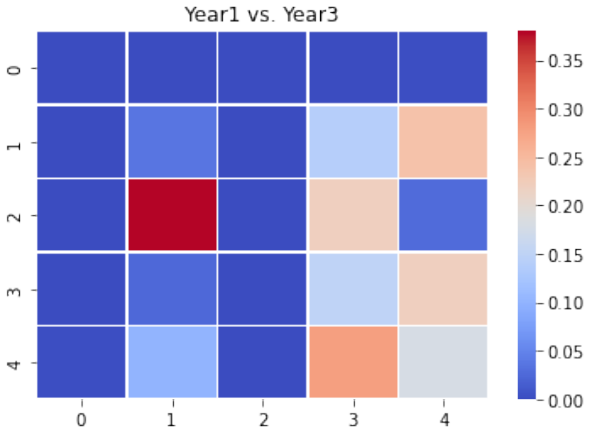
Figure D.6: Comparison of K-means clustering with latent class analysis(LCA) in year 3 cohort 3

Jaccard similarity index

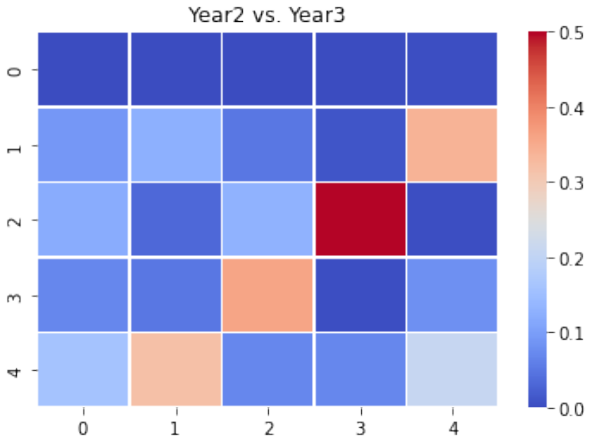
a. Cohort 2



(a)



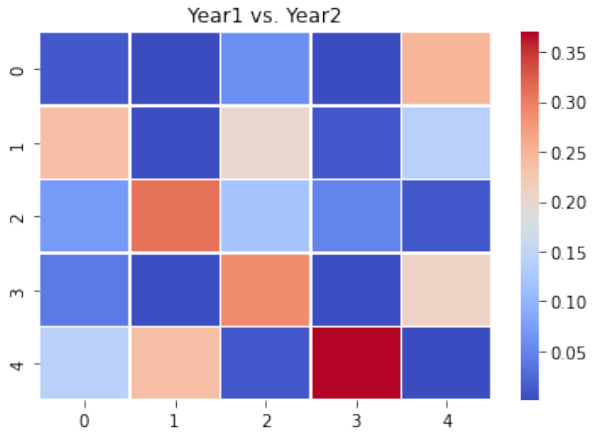
(b)



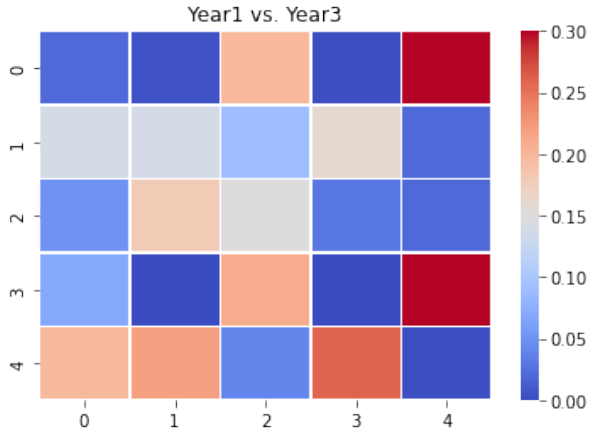
(c)

Figure E.1: Jaccard similarity index for cohort 2. (a) year 1; (b) year 2; (c) year 3 .

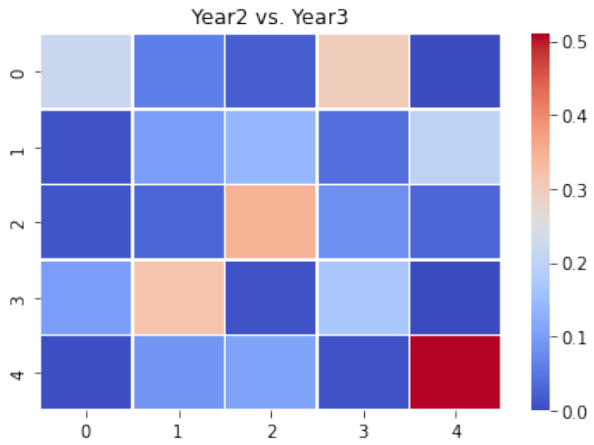
b. Cohort 3



(a)



(b)



(c)

Figure E.2: Jaccard similarity index for cohort 3. (a) year 1; (b) year 2; (c) year 3 .

Important variables based on MDA

a. Diagnosis

1. Year 2

Cohort1	Cohort2	Cohort3
Other long term (current) drug therapy	Hypertensive diseases	Abnormalities of breathing
Long term (current) use of anticoagulants	Other long term (current) drug therapy	Abnormal electrocardiogram[ECG] [EKG]
Long term (current)use of aspirin	Long term (current) use of aspirin	Hypertensive diseases
Chronic ischemic heart disease	Encounter for other preprocedural examination	Disorders of lipoprotein metabolism and other lipidemias
Abnormalities of breathing	Low back pain	Other long term (current) drug therapy
Osteoarthritis	Atrial fibrillation and flutter	Disease of the blood and blood forming organs
Abnormal electrocardiogram [ECG] [EKG]	Atherosclerotic heart disease of native coronary artery without angina pectoris	Cardiomegaly
Other specified postprocedural states	Bradycardia, unspecified	Weakness
Major depressive disorder, single episode, unspecified	Disorders of lipoprotein metabolism and other lipidemias	Atherosclerotic heart disease of native coronary artery without angina pectoris
Encounter for other preprocedural examination	Abnormal electrocardiogram [ECG] [EKG]	Long term (current) use of anticoagulants
Disorders of lipoprotein metabolism and other lipidemias	Cardiomegaly	Encounter for immunization
Hypertensive diseases	Type 2 diabetes mellitus	Encounter for general adult medical examination without abnormal findings
Other fatigue	Weakness	Hypothyroidism
Encounter for general adult medical examination without abnormal findings	Encounter for immunization	Type 2 diabetes mellitus
Encounter for screening mammogram for malignant neoplasm of breast	Chest pain, unspecified	Other chronic pain
Encounter for immunization	Hyperkalemia	Other fatigue
Encounter for screening for malignant neoplasm of colon	Pleural effusion	Overweight and obesity
Overweight and obesity	Overweight and obesity	Encounter for screening for malignant neoplasm of colon
Hypothyroidism	Encounter for general adult medical examination without abnormal findings	Vitamin D deficiency, unspecified
Unspecified abdominal pain	Disease of the blood and blood forming organs	Cough
Chest pain, unspecified	Major depressive disorder,single episode, unspecified	Encounter for other preprocedural examination
Atrial fibrillation and flutter	Other chronic pain	Other specified postprocedural states
Hypokalemia	Chronic obstructive pulmonary disease	Low back pain
Localized edema	Long term (current) use of anticoagulants	Encounter for screening mammogram for malignant neoplasm of breast
Dorsalgia, unspecified	Anxiety disorders	
Diarrhea, unspecified	Constipation, unspecified	
Encounter for follow-up examination after completed treatment for conditions other than malignant neoplasm	Abnormalities of breathing	
Edema, unspecified	Other specified postprocedural states	
Other specified soft tissue disorders	Hypokalemia	
Other chronic pain		
Atherosclerotic heart disease of native coronary artery without angina pectoris		
Cardiomegaly		

Table F.1: Year 2 diagnosis comparison

2. Year 3

Cohort1	Cohort2	Cohort3
Disorders of lipoprotein metabolism and other lipidemias	Hypertensive diseases	Other long term (current) drug therapy
Hypertensive diseases	Bradycardia, unspecified	Hypertensive diseases
Other long term (current) drug therapy	Abnormal electrocardiogram [ECG] [EKG]	Disorders of lipoprotein metabolism and other lipidemias
Encounter for immunization	Other long term (current) drug therapy	Abnormalities of breathing
Long term (current) use of aspirin	Disorders of lipoprotein metabolism and other lipidemias	Anxiety disorders
Type 2 diabetes mellitus	Abnormalities of breathing	Abnormal electrocardiogram [ECG] [EKG]
Urinary tract infection, site not specified	Chest pain, unspecified	Overweight and obesity
Overweight and obesity	Disease of the blood and blood forming organs	Long term (current) use of anticoagulants
Vitamin D deficiency, unspecified	Long term (current) use of anticoagulants	Encounter for other preprocedural examination
Osteoarthritis	Major depressive disorder, single episode, unspecified	Atherosclerotic heart disease of native coronary artery without angina pectoris
Atherosclerotic heart disease of native coronary artery without angina pectoris	Long term (current) use of aspirin	Major depressive disorder, single episode, unspecified
Hypothyroidism	Type 2 diabetes mellitus	Encounter for general adult medical examination without abnormal findings
Encounter for general adult medical examination without abnormal findings	Chronic obstructive pulmonary disease	Encounter for immunization
Disease of the blood and blood forming organs	Encounter for immunization	Hypothyroidism
Hypokalemia	Overweight and obesity	Type 2 diabetes mellitus
Nonrheumatic mitral (valve) insufficiency	Atherosclerotic heart disease of native coronary artery without angina pectoris	Encounter for screening mammogram for malignant neoplasm of breast
Cardiomegaly	Other specified postprocedural states	Cough
Abnormal electrocardiogram [ECG] [EKG]	Osteoarthritis	Vitamin D deficiency, unspecified
Diarrhea, unspecified	Unspecified abdominal pain	Other fatigue
Chronic ischemic heart disease	Pleural effusion	Disease of the blood and blood forming organs
Encounter for other preprocedural examination	Weakness	Encounter for screening for malignant neoplasm of colon
Other specified postprocedural states	Hyperkalemia	
Weakness	Nonrheumatic mitral (valve) insufficiency	
Major depressive disorder, single episode, unspecified	Hypokalemia	
Abnormalities of breathing		
Chronic obstructive pulmonary disease		
Chest pain, unspecified		
Anxiety disorders		

Table F.2: Year 3 diagnosis comparison

b. Procedures

1. Year 2

Cohort1	Cohort2	Cohort3
Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	Initial hospital care, per day, for the e valuation and management of a patient
Measurement of albumin in plasma	Measurement of Troponin	Radiologic examination, chest; single view, frontal.
Drug test(s), definitive,	Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral	Measurement of magnesium
Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral	Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count	Measurement of inorganic phosphorus (phosphate)
Comprehensive metabolic panel	Radiologic examination, chest; single view, frontal.	Subsequent hospital care, per day, for the evaluation and management of a patient
Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	Prothrombin time test	Prothrombin time test
Measurement of total calcium	Drug test(s), definitive,	Non-covered item or service
Therapeutic procedure, one or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	Non-covered item or service	Partial thromboplastin time test on blood
Automated urinalysis using dip stick and microscopy of urine	Comprehensive metabolic panel	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report
Measurement of Troponin	Partial thromboplastin time test on blood	Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral
Collection of venous blood by venipuncture	Collection of venous blood by venipuncture	Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count
Direct measurement of high density cholesterol	Subsequent hospital care, per day, for the evaluation and management of a patient	Collection of venous blood by venipuncture
Level 4 outpatient visit for established patient with problem of moderate to high severity 25 minutes	Initial hospital care, per day, for the evaluation and management of a patient	Drug test(s), definitive,
Measurement of thyroid stimulating hormone (TSH)	Aerobic and anaerobic bacterial culture of blood	Comprehensive metabolic panel
Non-covered item or service	Measurement of magnesium	Direct measurement of high density cholesterol
Measurement of total bilirubin	Measurement of lactate (lactic acid)	Measurement of thyroid stimulating hormone (TSH)
Measurement of vitamin D 25 hydroxy	Level 4 outpatient visit for established patient with problem of moderate to high severity 25 minutes	Measurement of glycosylated hemoglobin (HbA1C)
Measurement of magnesium	Level 3 outpatient visit for established patient with problem of low to moderate severity 15 minutes	Measurement of albumin in plasma
Prothrombin time test	Direct measurement of high density cholesterol	Automated urinalysis using dip stick and microscopy of urine
Partial thromboplastin time test on blood	Measurement of glycosylated hemoglobin (HbA1C)	Influenza immunization administered
Measurement of lactate (lactic acid)	Measurement of total calcium	Level 4 outpatient visit for established patient with problem of moderate to high severity 25 minutes
Radiologic examination, chest; single view, frontal.	Measurement of inorganic phosphorus (phosphate)	Level 3 outpatient visit for established patient with problem of low to moderate severity 15 minutes
Analysis of CO2 of blood	Automated complete blood cell count	Measurement of Troponin
		Postoperative follow-up visit, included in global service.
		Measurement of total calcium
		Measurement of vitamin D 25 hydroxy

Table F.3: Year 2 procedure comparison

2. Year 3

Cohort1	Cohort2	Cohort3
Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count	Automated urinalysis using dip stick and microscopy of urine	Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral
Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral	Measurement of potassium in plasma	Non-covered item or service
Collection of venous blood by venipuncture	Hospital discharge day management; 30 minutes or less.	Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count
Direct measurement of high density cholesterol	Collection of venous blood by venipuncture	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report
Comprehensive metabolic panel	Radiologic examination, chest; single view, frontal.	Drug test(s), definitive,
Drug test(s), definitive,	Measurement of direct bilirubin	Partial thromboplastin time test on blood
Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	Measurement of lactate (lactic acid)	Measurement of Troponin
Non-covered item or service	Measurement of magnesium	Collection of venous blood by venipuncture
Measurement of total calcium	Measurement of alkaline phosphatase	Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only.
Prothrombin time test	Measurement of aspartate amino transferase (AST) (SGOT)	Automated urinalysis using dip stick and microscopy of urine
Measurement of glycosylated hemoglobin (HbA1C)	Measurement of inorganic phosphorus (phosphate)	Direct measurement of high density cholesterol
Level 3 outpatient visit for established patient with problem of low to moderate severity 15 minutes	Non-covered item or service	Prothrombin time test
Measurement of inorganic phosphorus (phosphate)	Prothrombin time test	Measurement of thyroid stimulating hormone (TSH)
Initial hospital care, per day, for the evaluation and management of a patient	Drug test(s), definitive,	Comprehensive metabolic panel
Measurement of lactate (lactic acid)	Subsequent hospital care, per day, for the evaluation and management of a patient	Influenza immunization administered
Analysis of CO2 of blood	Initial hospital care, per day, for the evaluation and management of a patient	Level 3 outpatient visit for established patient with problem of low to moderate severity 15 minutes
Subsequent hospital care, per day, for the evaluation and management of a patient	Long description: Prednisone, oral, per 5 mg Short description: Prednisone oral	Measurement of glycosylated hemoglobin (HbA1C)
Measurement of chloride in blood	Automated complete blood cell count (CBC) with automated differential leukocyte (WBC) count	Subsequent hospital care, per day, for the evaluation and management of a patient
Subsequent hospital care, per day, for the evaluation and management of a patient,	Comprehensive metabolic panel	Hospital discharge day management; 30 minutes or less.
Measurement of sodium in plasma	Measurement of glycosylated hemoglobin (HbA1C)	Measurement of albumin in plasma
Aerobic and anaerobic bacterial culture of blood	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	Measurement of creatinine in blood
Automated urinalysis using dip stick and microscopy of urine	Measurement of thyroid stimulating hormone (TSH)	Measurement of vitamin D 25 hydroxy
Measurement of total bilirubin	Level 4 outpatient visit for established patient with problem of moderate to high severity 25 minutes	
Measurement of magnesium	Level 3 outpatient visit for established patient with problem of low to moderate severity 15 minutes	
Measurement of Troponin	Level 2 outpatient visit for evaluation a established patient 10 minutes or less	
Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only.	Subsequent hospital care, per day, for the evaluation and management of a patient,	
	Measurement of Calcium; ionized	
	Continuous positive airway pressure ventilation (CPAP), initiation and management	

Table F.4: Year 3 procedure comparison

c. Medications

1. Year 2

Cohort1	Cohort2	Cohort3
Ceftriaxone (as Ceftriaxone Sodium)	Rocuronium Bromide	Heparin Sodium
Fentanyl Citrate	Lidocaine Hydrochloride	Ondansetron Hydrochloride [Ondansetron]
Ondansetron Hydrochloride [Ondansetron]	50 ml Albumin Human	Sodium Chloride
Sodium Chloride	Hydromorphone HCL	Fentanyl Citrate
Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water] #4	2 ml Sugammadex	Docusate Dodium
Sodium Chloride [Veterinary Lactated]	Dexamethasone Sodium Phosphate	Polyethylene Glycol
Docusate Sodium	Acetaminophen	Enoxaparin Sodium
Furosemide	Ephedrine Sulfate [Premierpro RX Ephedrine Sulfate]	Docusate Sodium
Heparin Sodium	Sodium Chloride[Veterinary Lactated]	Furosemide
Enoxaparin Sodium	Robinul	Ceftriaxone (as Ceftriaxone Sodium)
Polyethylene Glycol	Fentanyl Citrate	Sodium Chloride [Veterinary Lactated]
Ipratropium Bromide	Sodium Chloride	Ipratropium Bromide
Dextrose Monohydrate	Ondansetron Hydrochloride [Ondansetron]	Famotidine
Nitroglycerin	Donepezil Hydrochloride [Donepezil Hydrochloride]	Phillips Milk of Magnesia
Epinephrine;isopropyl Alcohol [Epinephrinesnap-v]	Oxycodone Hydrochloride	Calcium Chloride
Chlorhexidine Gluconate	Hydrocodone Bitartrate	Acetaminophen
Product Containing Precisely Oxycodone Hydrochloride (Clinical Drug)	Ceftriaxone (as Ceftriaxone Sodium)	Dexamethasone Sodium Phosphate
50 ml Albumin Human	Omeprazole	0.5 ml Hydromorphone Hydrochloride
Pregabalin	Metformin Hydrochloride	Hydrocodone Bitartrate
Calcium Chloride	Meclizine Hydrochloride [Meclizine Hydrochloride]	Lidocaine Hydrochloride
Levalbuterol Hydrochloride	Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water] #4	
Lidocaine	Zocor	
Lidocaine Hydrochloride	Tetracaine Hydrochloride	
Rocuronium Bromide	Furosemide	
	Heparin Sodium	
	Docusate Sodium	
	Morphine Sulfate	
	Enoxaparin Sodium	

Table F.5: Year 2 medication comparison

2. Year 3

Cohort1	Cohort2	Cohort3
Sodium Chloride	Atropine Sulfate	Ondansetron Hydrochloride
Ondansetron Hydrochloride	Buspirone Hydrochloride	Fentanyl Citrate
Fentanyl Citrate	Ondansetron	Sodium Chloride
Sodium Chloride	Montelukast	Ceftriaxone (as Ceftriaxone Sodium)
Hydrocodone Bitartrate	Phenol	Docusate Sodium
Nitroglycerin	Flagyl	Rocuronium Bromide
Furosemide	Sucralfate [Carafate]	Docusate Sodium
Heparin Sodium	Remeron	Sodium Chloride
Ceftriaxone (as Ceftriaxone Sodium)	Pregabalin	Dexamethasone Sodium Phosphate
Ipratropium Bromide	Esomeprazole Magnesium	Lidocaine Hydrochloride
Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water]_#4	Sodium Chloride	Enoxaparin Sodium
Polyethylene Glycol	Docusate Sodium	Acetaminophen
Phillips Milk of Magnesia	Ondansetron Hydrochloride [Ondansetron]	Phillips Milk of Magnesia
Piperacillin Sodium	Fentanyl Citrate	Bisacodyl
Lidocaine Hydrochloride	Morphine Sulfate	Famotidine
Dexamethasone Sodium Phosphate	Sodium Chloride [Veterinary Lactated]	Morphine Sulfate
Enoxaparin Sodium	Docusate Sodium	Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water] #4
Docusate Sodium	Famotidine	Hydralazine Hydrochloride
	Enoxaparin Sodium	Heparin Sodium
	Rocuronium Bromide	Hydrocodone Bitartrate
	Ceftriaxone (as Ceftriaxone Sodium)	Lorazepam
	Lidocaine Hydrochloride	
	Dexamethasone Sodium Phosphate	
	Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water] _#4	
	Ergocalciferol	
	Ropinirole	
	Methylprednisolone Acetate	
	Piperacillin Sodium	
	Lamictal	
	Tizanidine [Zanaflex]	
	Magnesium Sulfate Heptahydrate [Magnesium Sulfate in Water] _#3	
	Acetaminophen	

Table F.6: Year 3 medication comparison

d. Labs

1. Year 2

Cohort1	Cohort2	Cohort3
Platelets	Platelets	Hemoglobin A1c Hemoglobin total in Blood
Hepatitis B virus surface Ab Units volume in Serum	Hemoglobin A1c Hemoglobin total in Blood	Bilirubin.total Mass volume in Serum Plasma or Blood
Protein Mass volume in Serum or Plasma	Protein Mass volume in Serum or Plasma	Protein Mass volume in Serum or Plasma
Glucose	Hepatitis B virus surface Ab Units volume in Serum	Hepatitis B virus surface Ab Units volume in Serum
Bilirubin total Mass volume in Serum Plasma or Blood	Folate Mass volume in Serum Plasma or Blood	Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood
White blood cell Leukocytes volume in Blood	Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood	Urea nitrogen Mass volume in Serum Plasma or Blood
Hemoglobin A1c Hemoglobin total in Blood	Urea nitrogen Mass volume in Serum Plasma or Blood	Hematocrit
Urea nitrogen Mass volume in Serum Plasma or Blood	Hematocrit	Glucose
Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood	White blood cell Leukocytes volume in Blood	Platelets
Folate Mass volume in Serum Plasma or Blood	Glucose	White blood cell Leukocytes volume in Blood
Hematocrit		

Table F.7: Year 2 lab comparison

2. Year 3

Cohort1	Cohort2	Cohort3
Urea nitrogen Mass volume in Serum Plasma or Blood	Hematocrit	Hematocrit
ematocrit	Platelets	Folate Mass volume in Serum Plasma or Blood
Hematocrit	White blood cell Leukocytes volume in Blood	
White blood cell Leukocytes volume in Blood		White blood cell Leukocytes volume in Blood
Glucose	Glucose	Urea nitrogen Mass volume in Serum Plasma or Blood
Bilirubin total Mass volume in Serum Plasma or Blood	Hemoglobin A1c Hemoglobin total in Blood	Platelets
	Hepatitis B virus surface Ab Units volume in Serum	
Hepatitis B virus surface Ab Units volume in Serum	Protein Mass volume in Serum or Plasma	Protein Mass volume in Serum or Plasma
		Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood
Platelets	Urea nitrogen Mass volume in Serum Plasma or Blood	Hemoglobin A1c Hemoglobin total in Blood
Hemoglobin A1c Hemoglobin total in Blood	Bilirubin total Mass volume in Serum Plasma or Blood	
Alkaline phosphatase Enzymatic activity volume in Serum Plasma or Blood		
Protein Mass volume in Serum or Plasma		
Folate Mass volume in Serum Plasma or Blood		

Table F.8: Year 3 lab comparison

e. Vital signs

1. Year 2

Cohort1	Cohort2	Cohort3
Heart rate	BMI	Body temperature
Blood Pressure Diastolic	Blood Pressure Diastolic	BMI
Body temperature	Blood Pressure Systolic	Heart rate
BMI	Heart rate	Blood Pressure Diastolic
Blood Pressure Systolic	Body temperature	Blood Pressure Systolic

Table F.9: Year 2 vital sign comparison

2. Year 3

Cohort1	Cohort2	Cohort3
Blood Pressure Diastolic	Body temperature	Body temperature
Heart rate	BMI	Heart rate
Blood Pressure Systolic	Heart rate	Blood Pressure Systolic
Body temperature	Body temperature	BMI
BMI	Blood Pressure Diastolic	Blood Pressure Systolic

Table F.10: Year 3 vital sign comparison