Electronic Health Record Based Disease Mapping Using Data Analytics

Adama Science & Technology University

Adama, Ethiopia

August 2018

Electronic Health Record Based Disease Mapping Using Data Analytics

Principal Investigator

Elias Lemuye (MSc.)

Research Members

Prof. I. Achim (PhD)

Endale Aregu (MSc.)

Gedamu Alemu (MSc.)

Acknowledgements

This research has been realized with the financial support and guidance of Adama Science and Technology University. Moreover, Addis Ababa Health Bureau, Diredawa Health Bureau, Hawassa and other hospitals also support us in providing row data and documents. Hence the research members are happy when they express their gratitude to those who have played significant role for the accomplishment of this study.

Contents

List of Tables

Lists of Figures

Lists of Acronyms & Abbreviations

CDC - Centers for Disease Control and Prevention

DHS-Demographic Health Survey

HDD- Health Data Depot

eCHIS - Electronic Community Health Information System

eHMIS- Electronic Health Management Information Systems

EHR- Electronic Health Records

ERM- Electronic Record Management

EPI -Expanded Program of Immunizations Coverage Survey

FMoH- Federal Ministry of Health

GIS – Geographic Information Systems

HSTP – Health Sector Transformation Program

ROE -Return on Effort

SARA -Service Availability and Readiness Assessment

SPA+ -Service Provision Assessment Plus

WHO- World Health Organization

WSS- within-cluster sum of square

Abstract

While wonderful new medical discoveries and innovations are in the news every day, uncertainties and unanswered healthcare questions are a daily reality for the decision makers who provide care. Measuring & visualizing the magnitude of disease burden on population have been considered as one of the necessary input for tailored healthcare program design and taking action.

The overall purpose of this study is to detect disease burden and make cluster analysis by applying data analytics tools and techniques on EHR.

To meet the stated objective it has been attempted to include sources of the data, methods of data collection, techniques and tools of analysis. Particularly cluster analysis using hierarchical and kmeans algorithms are applied.

The results of the study revealed, EHRs very important mainly in saving costs, infrastructure and skilled manpower are available that can support the proper handling of EHR systems. The reporting system is established from bottom up with the support of guidelines and training. However, it has been seen on different national health policies and our observation, there is less use of the system especially the ERM by the physicians and other health experts.

Moreover, data analytics tools and techniques that can support EHR data analytics, especially for clustering, was also identified Most of the software on data clustering is open-source software, which is freely available. On the other hand, most of the commercial software comprises implementations of classical algorithms such as k-means or agglomerative clustering.

Finally, with the application of the selected clustering techniques, metrics, tools and dataset, it has been attempted to successfully detect optimal number of disease clusters and met the objectives of the study.

CHAPTER ONE

1. Introduction

This chapter presents introductory part comprising of the background of the study, rationale for the study and statement of the problem, basic research questions, objectives of the study, and significance of the study.

1.1.Background

Disease Burden in Ethiopia

While wonderful new medical discoveries and innovations are in the news every day, uncertainties and unanswered heathcare questions are a daily reality for the decision makers who provide care (Nair, Hsu, & Celi, 2016). Community and population health is one of the areas in which healthcare analytics are used. The Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO), for example, are engaged in monitoring and forecasting disease incidence and prevalence with the aim of better understanding relationships between health, health-related behaviors, sanitation and water quality, and other factors (Steele, Chandler, & Reddy, 2016).

The term disease broadly refers to any condition that impairs the normal functioning of the body. Commonly, disease is used to refer specifically to infectious diseases, which are clinically evident diseases that result from the presence of pathogenic microbial agents, including viruses, bacteria, fungi, protozoa, multi cellular organisms, and aberrant proteins. (https://en.wikipedia.org, n.d.) In Ethiopia about 80% of diseases are attributable to preventable conditions that are related to personal and environmental hygiene, infectious diseases and malnutrition. Environmental risk factors alone account for 31% of the total disease burden in the country (WHO, 2013).

Disease Cluster Detection

The objects are clustered or grouped based on the principle of *maximizing the intraclass similarity and minimizing the interclass similarity*. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

There are three main situations in which the statistical analysis of disease clustering is important in: (Lawson, 1999)

- epidemiological research when trying to study the aetiology of a disease
- public health as part of geographical disease surveillance: and
- response to disease cluster alarms to evaluate whether thorough epidemiological investigations are warranted.

Electronic Health Records Data Analytics

Major electronic health based data analytics include (IBM, 2013):

- Clinical analytics bring together clinical, financial and operational data to answer questions and perform retrospective analysis about how healthcare organizations are running, the state of their patient populations and the effectiveness of programs.
- Advanced analytics, which are more predictive and forward– looking in nature, often focus on making predictions regarding at-risk patients. For example, they might help providers identify which patients require immediate intervention or additional treatment, or which patients could most benefit from particular wellness programs.

One of the strategic objectives of HSTP is to enhance use of technology and innovation. For the achievement of this objective several initiatives and their deliverables are designed. From these initiatives enhancing the health information systems through use of existing and new information technology is the one which includes the proper use of EMR/EHR. Ensuring Electronic / web-based health related information system is expected to be the

deliverable of this initiative which in turn is one of the means to enhance use of technology and innovation. (FMoH, 2014)

1.2.Statement of the Problem

Measuring & visualizing the magnitude of disease burden on population have been considered as one of the necessary input for tailored healthcare program design and taking action.

Spatial clusters, or 'hotspots,' of disease and health-related behaviors have long been of interest to public health researchers and policymakers. Defined as an unusual number of cases within a population, place, and time period, a disease hotspot is a geographical construct that can be identified, visualized and explored using GIS and spatial analysis methods (Mclafferty, 2015).

The issues raised by geographers and statisticians reveals; whether disease occurrences and their spatial spread exhibit patterns of some sort (clustered, dispersed, or random. (Lai, So, & Chan, 2009). How to analyze disease incidence or prevalence when we have geographical information. (Lawson, 2008)

The methods for detecting disease hotspots have advanced significantly to incorporate more robust statistical formulations and spatial search processes that are important for accurately detecting disease hotspots; however, everyday mobility is an important element in disease processes that has not been well incorporated in spatial cluster detection methods.

One of the solutions for the aforementioned problem is that since much of the patient information collected for clinical trials already exists in the patient record, clinical researchers can quickly import such information from the existing practice record into the research record. In doing so, they can save both time and money. EMR systems could speed data acquisition and searching, allow mass computing and sampling, and provide the research community access to a broader and more diverse patient population**.**

Despite plenty of benefits that can be obtained from EHR analytics, sadly, many in the medical community remain in the dark ages by not applying healthcare analytics to determine the needs of patients and their community. This is a limitation in the current medical facility, resulting in poor treatment options for patients. EHR data analytics obtained directly from a community is capable of identifying key needs local citizens have such as what their most common ailments are and services they require. It is essential for a healthcare provider to tailor its services towards the community and effective use of analytics turns data into actionable information.

1.3. Research Questions

- Given the importance of EHR system, available infrastructure and skill, how is the level of acceptance &utilization at point of clinical care and for research?
- Which data analytics tools & techniques best applied on EHR so as to detect disease burden?
- Given EHR dataset and cluster analysis techniques, what is the optimal detection disease clusters?

1.4. Objectives of the Study

1.4.1. General Objective

The overall purpose of this study is to detect disease burden and make cluster analysis by applying data analytics tools and techniques on EHR.

1.4.2. Specific Objectives

- State-of-the-art in EHR acceptance & utilization at hospitals
- To explore data sources, methods & tools that support data analytics
- To develop EHR data analytics that detect disease burden in the patient population
- Optimize clustering quality based on empirical metrics
- To evaluate the result and arrive at concluding remarks

1.5. Significance of the Study

Effective use of analytics in the healthcare industry can improve current care as well as can facilitate preventive care. Generally, the findings of this research are meant for both practitioners and researchers. For researchers, hotspots can provide clues to disease risk behaviors, suggesting local environmental or social characteristics that promote increased risk. For policy makers and planners, selectively targeting interventions to hotspot areas can be an effective public health intervention strategy (Mclafferty, 2015).

CHAPTER TWO

2. Literature Review

This chapter presents the review of literatures that attempted to see the sources of electronic health data, level of utilization of EHR, its benefits and major barriers of adopting EHR. Moreover, EHR data analytics and different types of clustering techniques are reviewed.

2.1.Sources of Electronic Health Data

The rapid growth of novel technologies has led to a significant increase of digital health data in recent years. More medical discoveries and new technologies such as mobile apps, capturing devices, novel sensors, and wearable technology have contributed to additional data sources. Therefore, the healthcare industry produces a huge amount of digital data by utilizing information from all sources of healthcare data (Fang, Pouyanfar, Yang, Chen, & Iyengar, 2016).

Healthcare data sources also indicate the different types of data that we can gather for the purpose of analysis. These include routine administrative sources, such as the HMIS; household surveys, such as the Demographic Health Survey (DHS) and Expanded Program of Immunizations Coverage Survey (EPI); health facility surveys, such as the Service Provision Assessment Plus (SPA+) and the Service Availability and Readiness Assessment (SARA); disease and behavioral surveillance; civil registration and vital statistics; financial and management information; censuses; and research studies (FMoH, 2016).

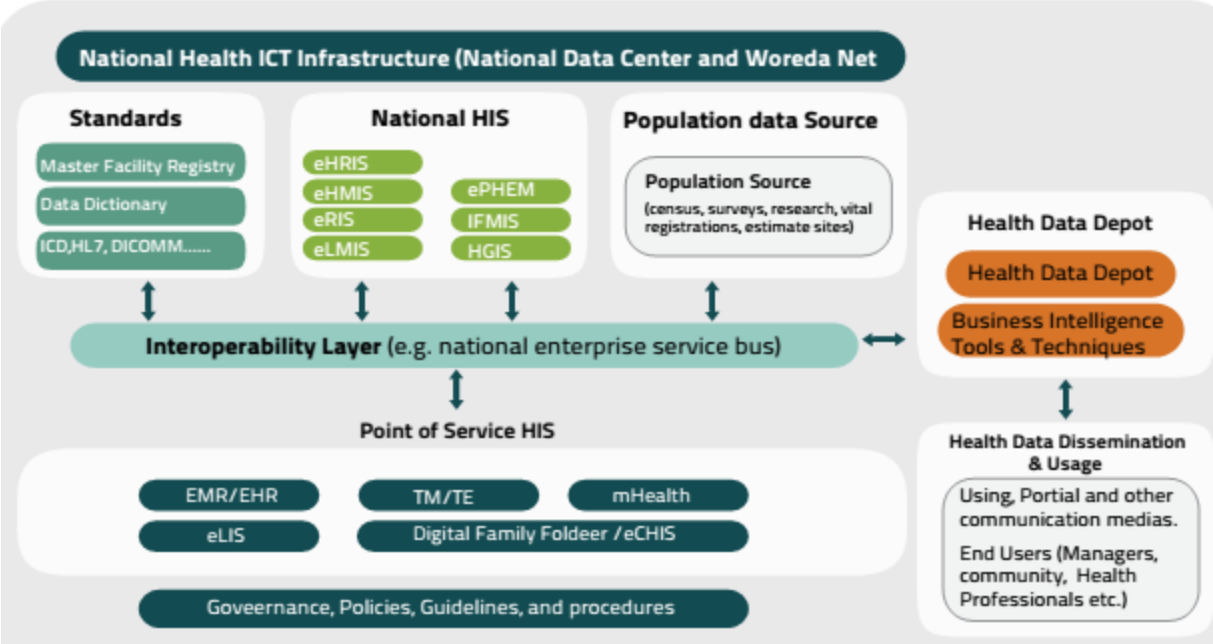


Figure 2.1: Ethiopia National Health Information Enterprise Architecture

Source: (FMoH, 2016)

As indicated on the figure, National Data Centre is a centralized storage facility that helps to store and provide adequate space to the data that is coming from various government sectors and agencies. The major applications that are hosted in the national data center are the national Ethiopia government portal (www.ethiopia.gov.et), Electronic Health Management Information System (eHMIS), Electronic Community Health Information System (eCHIS), and IFMIS. Therefore, this data center will also support the health sector to store health data and make it accessible to the responsible bodies when needed. The national level infrastructure also includes a Health Data Depot (HDD) that will serve as a data warehouse for aggregation and reporting of health data across many sources (FMoH, 2016).

2.2.EHR Acceptance & Utilization

Despite benefits associated with the use of electronic health records (EHRs), one major barrier to adoption is the concern that EHRs may take longer for physicians to use than paper-based systems. EHRs can save lives and save millions of dollars. So why have not EHR become the way most physicians practice medicine? There are many reasons; however, here are the four major reasons (Reddy & Aggarwal, 2015):

- High cost
- Lack of understanding (by all parties) about workflow considerations
- Lack of understanding that Physician Adoption is a change management process
- Inability to establish and follow a Timeline of Reasonable Goals including a real Re turn on Effort (ROE) for Physicians

Modern healthcare informatics generates and stores immense amounts of detailed patient and clinical process data. Very little real-world patient data have been used to further advance the field of health care. One large barrier to the utilization of these data is inaccessibility to researchers. Making these databases easier to access as well as integrating the data would allow more researchers to answer fundamental questions of clinical care (Nair, Hsu, & Celi, 2016).

Benefits of Electronic Health Records

Benefits of EHR	Description
Enhances Revenue	It decreases billing errors, provides a better documentation opportunity for these services that can be used to resolve financial disputes.
Avert Costs	Reduced paper and supply cost Improved utilization of tests Improved coordination of care & clinician satisfaction Improved accuracy of diagnosis & reliability Improved quality and convenience of care Improved aggregation of data and interoperability Improved legal and regulatory compliance

Table 2.1: Benefits of EHR

Source: (Reddy & Aggarwal, 2015)

Major Barriers to Adopting Electronic Health Records

Barriers	Description
Financial barriers	• start-up costs include purchasing hardware and software • Long-term costs include monitoring, modifying, and upgrading the system as well as storage and maintenance of health records.
Physician's resistance	• Physicians and staffs might not have sufficient technical knowledge to deal with EHRs, which leads them to think EHR systems are overly complex.
Loss of productivity	• Adoption of an EHR system requires a notable amount of time to select, purchase, and implement the system into clinical practice. During this period physicians have to work at a reduced capacity.
Usability issues	• The interface of software workflow has to be intuitive enough. In terms of usability, a comprehensive EHR system may be more complex than expected
Lack of standards	• Lack of uniform and consistent standards hinders the EHR adoption. This makes the data exchange difficult between the systems.
Privacy & security concerns	• EHR system may be subjected to attack since it contains personal and sensitive health records.

Table 2.2: Barriers to Adopting EHR

Source :(Reddy & Aggarwal, 2015)

Challenges of Using Electronic Health Data

Using the EHR data, we can conduct both patient-oriented and public health research. EHR data can be used for the early detection of epidemics and spread of diseases, environmental hazards, promotes healthy behaviors, and policy development. Big data success stories are becoming more common, but the challenges are no less daunting than they were in the past, and perhaps have become even more demanding as the field of data analytics in healthcare takes off (Nair, Hsu, & Celi, 2016).

The integration of genetic data with EHRs can open even wider horizons. As pooled EHRs achieve greater scale, researchers and other interested parties expect that the costs of hosting, sorting, formatting and analyzing these records are spread among a greater number of stakeholders, reducing the costs of pooled EHR analysis for all involved (Nair, Hsu, & Celi, 2016). But the data does not automatically provide us the knowledge. The quality and accuracy of the data is an issue to be taken care of.

2.3.EHR Data Analytics

The amount of data in healthcare is increasing at an astonishing rate. However, in general, the industry has not deployed the level of data management and analysis necessary to make use of those data. As a result, healthcare executives face the risk of being overwhelmed by a flood of unusable data. In Ethiopia, One of the strategic objectives of HSTP is to enhance use of technology and innovation. For the achievement of this objective several initiatives and their deliverables are designed. From these initiatives enhancing the health information systems through use of existing and new information technology is the one which includes the proper use of EMR/EHR. Ensuring Electronic / web-based health related information system is expected to be the deliverable of this initiative which in turn is one of the means to enhance use of technology and innovation. (FMoH, 2015)

Healthcare analytics refers to data analytic methods applied in the healthcare domain. Healthcare analytics is becoming a prominent data science domain because of the societal and economic burden of disease and the opportunities to better understand the healthcare system through the analysis of data (Steele, Chandler, & Reddy, 2016). The two major categories of EHR data analytics are clinical analytics which bring together clinical, financial and operational data to answer questions and perform retrospective analysis about how healthcare organizations are running, the state of their patient populations and the effectiveness of programs. The second one is advanced analytics, which are more predictive and forward– looking in nature, often focus on making predictions regarding at-risk patients. For example, they might help providers identify which patients require

immediate intervention or additional treatment, or which patients could most benefit from particular wellness programs (IBM, 2013).

2.4. Cluster Analysis

Clustering is one of the most widely used techniques for exploratory data analysis. Across all disciplines, from social sciences to biology to computer science, people try to get a first intuition about their data by identifying meaningful groups among the data points. Data clustering is one of the most popular data labeling techniques. In data clustering, we are given unlabeled data and are to put similar samples in one pile, called a cluster, and the dissimilar samples should be in different clusters (Aggarwal & Reddy, 2014).

Clustering is useful in several machine learning and data mining tasks including image segmentation, information retrieval, pattern recognition, pattern classification, network analysis, and so on. It can be seen as either an exploratory task or preprocessing step. If the goal is to explore and reveal the hidden patterns in the data, clustering becomes a stand-alone exploratory task by itself. However, if the generated clusters are going to be used to facilitate another data mining or machine learning task, clustering will be a preprocessing step (Aggarwal & Reddy, 2014).

Cluster analysis is a collection of methods for the task of forming groups where none exist. Sometimes it is possible to divide a collection of observations into distinct subgroups based on nothing more than the observation attributes. If this can be done, then understanding the population or process generating the observations becomes easier. The intent of cluster analysis is to carry out a division of a data set into clusters of observations that are more alike within cluster than between clusters (Steele, Chandler, & Reddy, 2016). Motivation for clustering in general including hierarchical clustering and applications encompass: analysis of data and pattern recognition, storage, search, and retrieval (Hennig, Meila, Murtagh, & Rocci, 2016).

Clusters are formed either by aggregating observations or dividing a single glob of observations into a collection of smaller sets. The process of cluster formation involves two varieties of algorithms. The first shuffles observations between a fixed numbers of

clusters to maximize within-cluster similarity. The second process begins with singleton clusters and recursively merges the clusters. Alternatively, we may begin with one cluster and recursively split off new clusters. In this study, we discuss two popular cluster analysis algorithms: the *k*-means algorithm and hierarchical agglomerative clustering (Steele, Chandler, & Reddy, 2016).

2.5. Hierarchical Cluster Analysis

Hierarchical cluster analysis (HCA) is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. It is most useful when you want to cluster a small number (less than a few hundred) of objects. The objects in hierarchical cluster analysis can be cases or variables, depending on whether you want to classify cases or examine relationships between the variables.

Unlike kmeans clustering, hierarchical clustering doesn't require us to specify the number of clusters beforehand. It is an alternative approach which builds a hierarchy from the bottom-up.

Strategies for hierarchical clustering fall into two types:

Agglomerative: where we start out with each document in its own cluster. The algorithm iteratively merges documents or clusters that are closest to each other until the entire corpus forms a single cluster. Each merge happens at a different (increasing) distance.

Divisive: where we start out with the entire set of documents in a single cluster. At each step the algorithm splits the cluster recursively until each document is in its own cluster. This is basically the inverse of an agglomerative strategy.

Hierarchical Clustering Algorithm:

1. find the closest two things
2. Put them together
3. Find the next closest

Once this is done, it is usually represented by a dendrogram like structure.

And it requires two arguments: a defined distance (Similarity) and a merging approach.

How do we define close? Defining closeness is a key aspect of defining a clustering method. Ultimately, the old rule of "garbage in, garbage out" applies. If you don't use a distance metric that makes sense for your data, then you won't get any useful information out of the clustering.

There are a number of commonly used metrics for characterizing distance or its inverse, similarity. For this particular study Euclidean distance was selected.

- **Euclidean distance**: A continuous metric which can be thought of in geometric terms as the "straight-line" distance between two points.

The important thing is to always pick a distance or similarity metric that makes sense for your problem.

In general the formula for Euclidean distance between point

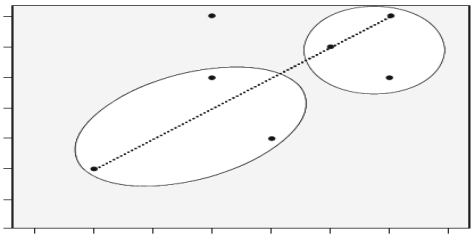
$A=(A1,A2,\ldots,An)$ and $B=(B1,B2,\ldots,Bn)$ is

$$Distance=((A1-B1)^2+(A2-B2)^2+\cdots+(An-Bn)^2)^{(1/2)}$$

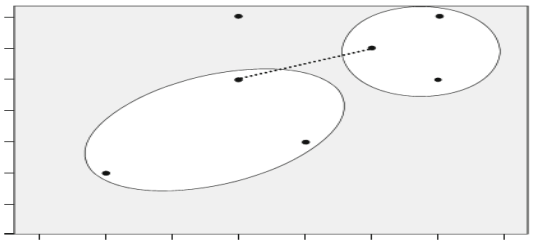
Hierarchical Cluster Analysis Method

The different hierarchical clustering methods are also considered as Optimizing Agglomerative Clustering Methods. There are a few ways to determine how close two clusters are (Aldenderfer & R. , 1984):

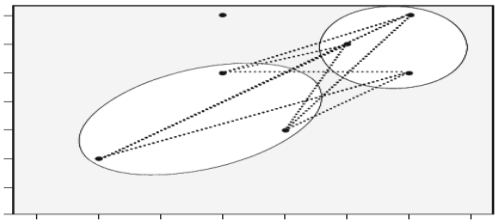
- **Complete linkage clustering**: Find the maximum possible distance between points belonging to two different clusters. This is the oppositional approach to single linkage which assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters.



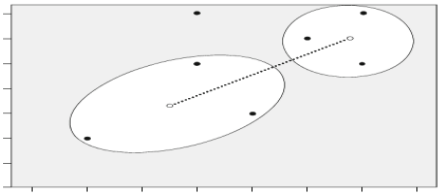
- **Single linkage clustering**: Find the minimum possible distance between points belonging to two different clusters. That means the distance between two clusters corresponds to the shortest distance between any two members in the two clusters.



- **Mean (Average) linkage clustering**: Find all possible pairwise distances for points belonging to two different clusters and then calculate the average.



- **Centroid linkage clustering**: Find the centroid of each cluster and calculate the distance between centroids of two clusters. The distance between the two clusters equals the distance between the two centroids.



Complete linkage and mean linkage clustering are the ones used most often.

2.6.K-Means Cluster Analysis

K-means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of *k* groups (i.e. *k* clusters), where *k* represents the number of groups pre-specified by the analyst. It classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e., high intra-class similarity), whereas objects from different clusters are as dissimilar as possible (i.e., low inter-class similarity). In k-means clustering, each cluster is represented by its center (i.e, centroid) which corresponds to the mean of points assigned to the cluster. K-means cluster analysis is a tool designed to assign cases to a fixed number of groups whose characteristics are not yet known but are based on a set of specified variables. It is most useful when you want to classify a large number of cases. It defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid (Hartigan, 1975):

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

• *xi* design a data point belonging to the cluster *Ck*

• *μk* is the mean value of the points assigned to the cluster *Ck*

Each observation (*xi*) is assigned to a given cluster such that the sum of squares (SS) distance of the observation to their assigned cluster centers *μk* is a minimum. We define the total within-cluster variation as follow:

$$tot.withinss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

The *total within-cluster sum of square* measures the compactness (i.e *goodness*) of the clustering and we want it to be as small as possible.

K-means algorithm

K-means algorithm can be summarized as follow (Kassambara, 2017):

1) Specify the number of clusters (K) to be created (by the analyst)
2) Select randomly k objects from the data set as the initial cluster centers or means
3) Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
4) For each of the k clusters update the *cluster centroid* by calculating the new mean values of all the data points in the cluster. The centroid of a *Kth* cluster is a vector of length *p* containing the means of all variables for the observations in the *kth* cluster; *p* is the number of variables.
5) Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations.

Factors Affecting Kmeans:

The major factors that can impact the performance of the K-means algorithm are the following (Aggarwal & Reddy, 2014):

1. choosing the initial centroids.

2. Estimating the number of clusters K.

K-means initializes the cluster means by randomly generating k points in the data space. This is typically done by generating a value uniformly at random within the range for each dimension. Each iteration of K-means consists of two steps: (1) cluster assignment, and (2) centroid update. Given the k cluster means, in the cluster assignment step, each point $x_j \in D$ is assigned to the closest mean, which induces a clustering, with each cluster C_i comprising points that are closer to μ_i than any other cluster mean (Zaki & Meira, 2014). Consider the one-dimensional data shown in Figure a. Assume that we want to cluster the data into k = 2 groups. Let the initial centroids be $\mu_1 = 2$ and $\mu_2 = 4$. In the first iteration, we first compute the clusters, assigning each point to the closest mean, to obtain

$C_1 = \{2,3\}$ $C_2 = \{4,10,11,12,20,25,30\}$

We next update the means as follows:

$\mu_1 = 2+3/2 = 2.5$ $\mu_2 = 4+10+11+12+20+25+30/7 = 16$

The new centroids and clusters after the first iteration are shown in Figure b. For the second step, we repeat the cluster assignment and centroid update steps, as shown in Figure c, to obtain the new clusters:

$C_1 = \{2,3,4\}$ $C_2 = \{10,11,12,20,25,30\}$

and the new means:

$\mu_1 = 2+3+4/3 = 3$ $\mu_2 = 10+11+12+20+25+30/6 = 18$

The complete process until convergence is illustrated in Figure. The final clusters are given as $C_1 = \{2,3,4,10,11,12\}$ $C_2 = \{20,25,30\}$ with representatives $\mu1 = 7$ and $\mu2 = 25$.

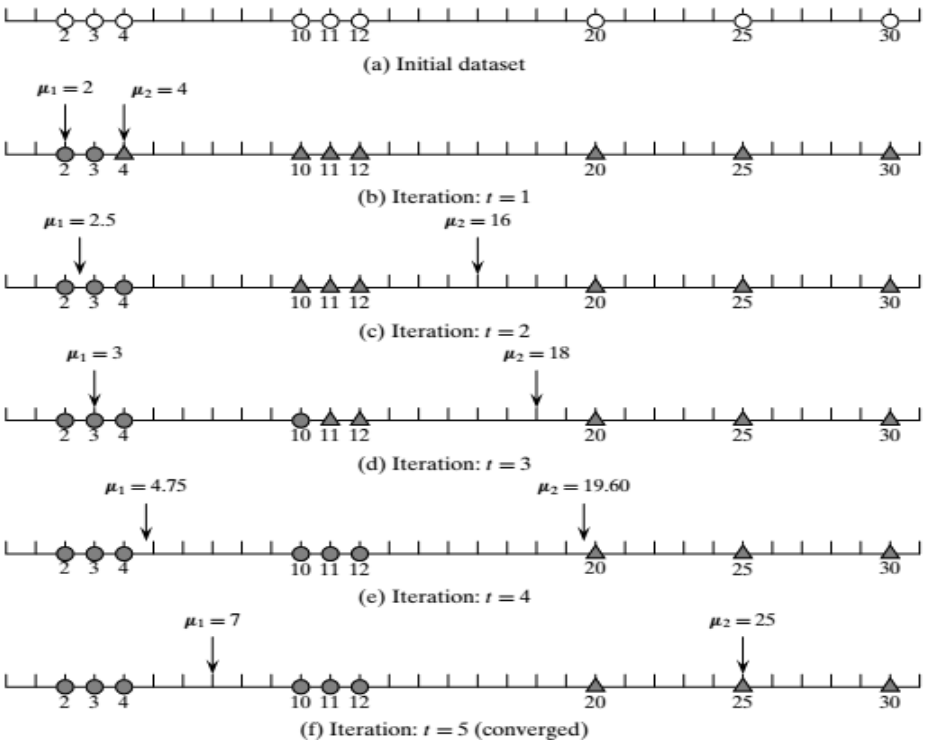


Figure 2.2: K-means in One Dimension

Source: (Zaki & Meira, 2014).

Methods of Cluster Optimization:

The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. A simple and popular solution consists of inspecting the dendrogram produced using hierarchical clustering to

see if it suggests a particular number of clusters. The other commonly used methods include direct methods and statistical testing (Kassambara, 2017):
1. Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named *elbow* and *silhouette* methods, respectively.
2. Statistical testing methods: consists of comparing evidence against null hypothesis. An example is the *gap statistic*.

For this particular study the directed methods have been reviewed:

Elbow method :

The Elbow method looks at the total intra-cluster variation or total within-cluster sum of square (WSS) as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS. The total WSS measures the compactness of the clustering and we want it to be as small as possible.
The optimal number of clusters can be defined as follow:
1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Average silhouette method:

Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k. It measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

The algorithm is similar to the elbow method and can be computed as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the average silhouette of observations (*avg.sil*).
3. Plot the curve of *avg.sil* according to the number of clusters k.
4. The location of the maximum is considered as the appropriate number of clusters.

2.7.Data Clustering Tools

Nowadays, significant amount of software is available for data clustering. Most of the software on data clustering is open-source software, which is freely available. On the other hand, most of the commercial software comprises implementations of classical algorithms such as k-means or agglomerative clustering. The free and Open-Source Software can be categorized as general and specialized clustering software (Aggarwal & Reddy, 2014)**:**

General Clustering Software:

- WEKA machine learning repository: contains software for clustering and other data mining related tasks such as data preprocessing, classification, and visualization.
- Spider: a widely used data mining software which contains the implementations of several popular clustering algorithms.
- Cluster: widely used open-source clustering software that contains several clustering and visualization algorithms.
- ELKI: suite of algorithms that include many classical partitioning algorithms, EM-based probabilistic algorithms, density-based algorithms, and subspace clustering algorithms.
- The KDnuggets web site: provides access to a significant number of open-source software sites for clustering and segmentation.

Specialized Clustering Software:

- OpenSubspace: contains the implementations of several subspace clustering algorithms.
- MALLET: performs clustering along with some statistical natural language processing and topic modeling.

- CLUTO: for clustering low- and high-dimensional data sets and for analyzing the characteristics of the various clusters.
- Gait-CAD : used for time-series data clustering

Commercial Packages :

- MATLAB: come with built-in methods for data clustering.
- The KDnuggets site provides a link to some of the more popular forms of software in this
  domain. This site provides pointers to software developed by other vendors such as IBM and SAS rather than only its own dedicated software.
- IBM SPSS: includes two step, k-means, and Kohonen clustering algorithms.
- SAS Enterprise Modeler: the clustering tool is available with other forms of visual and decision support.
- The NeuroXL Clusterizer : a commercial tool for clustering with neural networks.

CHAPTER THREE

3. Research Methodology

The research methodology section has attempted to incorporate how the detailed study is conducted. This include sources of the data, methods of data collection, techniques and tools of analysis.

3.1. Sources and Descriptions of the Data

Based on the plan disease count clustering analysis is conducted on the disease statistics data taken from selected health bureaus. The analysis is carried out on selected reportable diseases on specific time.

Sources of disease data	**Addis Ababa , Hawassa, Dire Dawa**
Selected Sources	Addis Ababa Health Bureau & Dire Dawa Health Bureau
Data Management tool	eHMIS
Selected data	2008 E.C data
Total Number of diseases	111- annexed
Selected Diseases	Top 20 diseases(based on frequency)- annexed

Table 3.1. Description of Sources of Disease Data

The eHMIS is an information system that enables health facilities, Woreda Health Offices (WorHO), Zonal Health Departments (ZHD), and RHBs to electronically compile weekly and immediately reportable diseases, out-patient department (OPD), inpatient-department (IPD), and service delivery data and electronically receive and submit them to the next level.

Core functionalities of eHMIS:

- Electronic Data Entry
    - Data Validation

    - Data Entry possible at all level in the health system
- Electronic Data Exchange
    - All kinds of reports can be generated and transmitted electronically to subsequent levels
- Easy and Quick Data/Indicator analysis
    - Charts can be produced, indicators can be calculated and compared instantly for analysis at all level,
    - Disaggregate data available for further analysis

3.2.Methods of Data Collection

Secondary data**:** using EHR dataset and document analysis. For the purpose of this study disease data was directly taken from Addis Ababa health bureau eHMIS systems and Dire Dawa Regional Health bureau eHMIS systems.

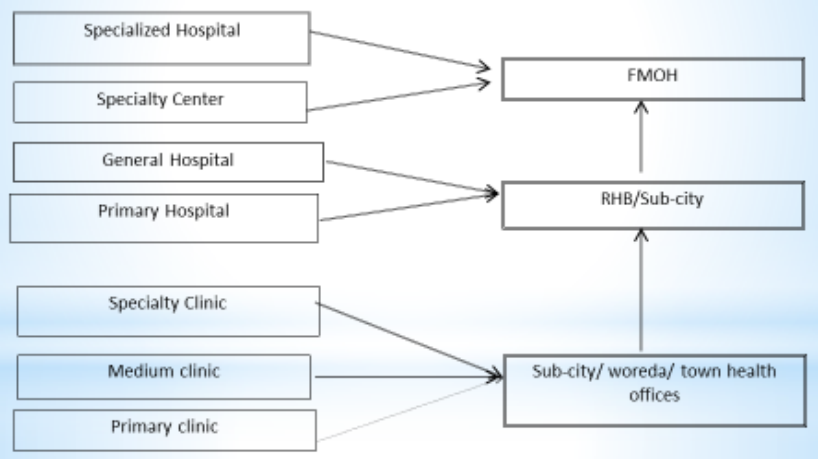


Figure 3.1: eHMIS Reporting Flow

Source: (FMoH, 2014)

- The reginal health bureau get the data from different clinics and hospitals on a monthly basis. Regional bureaus also report to federal on monthly, quarterly and annual basis.
- Primary data**:** observation and discussion with experts in the area during our visit for data collection is carried out.

3.3. Cluster Analysis Procedure

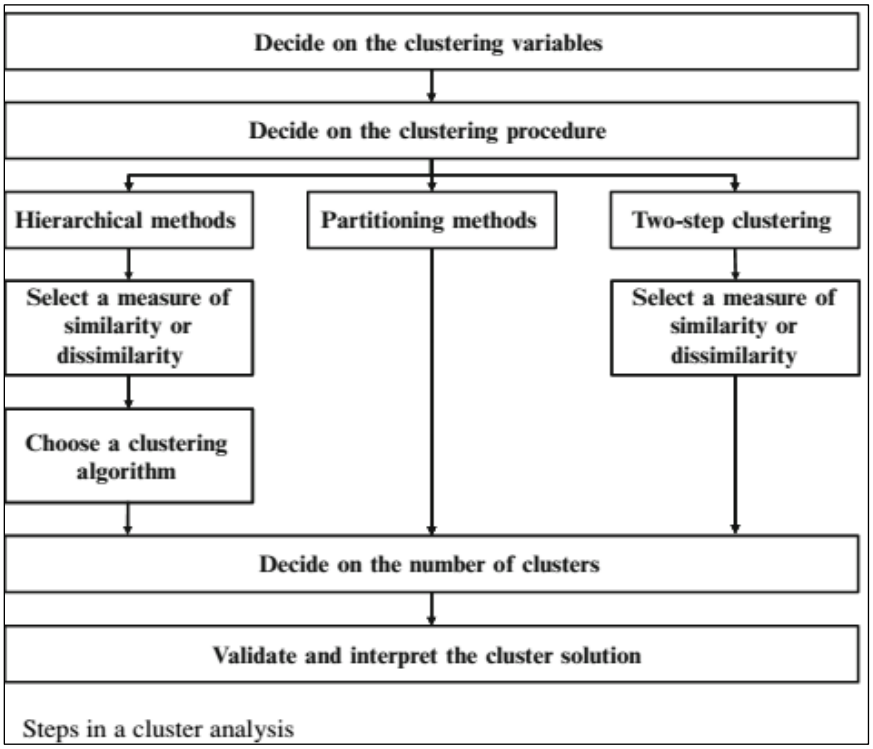


Figure 3.2: Steps in Cluster Analysis

Source: (Mooi & Sarstedt, 2011)

Decide on the Clustering Variables:

The types of variables used for cluster analysis provide different segments and, thereby, influence segment-targeting strategies.

Decide on the Clustering Procedure:

This always involves optimizing some kind of criterion, such as minimizing the within-cluster variance (i.e., the clustering variables' overall variance of objects in a specific cluster), or maximizing the distance between the objects or clusters.

There are many different clustering procedures and also many ways of classifying these A practical distinction is the differentiation between hierarchical and partitioning methods (most notably the k-means procedure)

3.4.Hierarchical Clustering Using R

Hierarchical cluster analysis (hclust) on a set of dissimilarities and methods for analyzing it.

Usage:
hclust(d, method = "complete", members = NULL)

plot(x, labels = NULL, main = "Cluster Dendrogram", sub = NULL, xlab = NULL, ylab = "Height")

Arguments

d a dissimilarity structure as produced by dist.
methodthe agglomeration method to be used. This include "single", "complete", "average".

members NULL or a vector with length size of d.

X an object of the type produced by hclust.

labels A character vector of labels for the leaves of the tree. By default the row names or row numbers of the original data are used. If labels = FALSE no labels at all are plotted.

main, sub, xlab, ylab character strings for title. sub and xlab have a non-NULL default when there's a tree$call.

We can perform a cluster analysis with the dist and hclust functions. The dist function calculates a distance matrix for your dataset, giving the Euclidean distance between any two observations. The hclust function performs hierarchical clustering on a distance matrix. So to perform a cluster analysis from raw data, we use these functions as shown below (http://www.instantr.com/2013/02/12/performing-a-cluster-analysis-in-r/).

> *modelname*<-hclust(dist(dataset))

The command saves the results of the analysis to an object named *modelname*.

The results of a cluster analysis are best represented by a dendrogram, which you can create with the plot function as shown.

> plot(modelname)

Be default, the row numbers or row names are used to label the observations. However you can use the labels argument to select a variable to use for the labels.

> plot(modelname, labels=dataset$variable)

To 'cut' the dendrogram to identify a given number of clusters, use the rect.hclust function immediately after the plot function as shown below:

```
> plot(modelname)
> rect.hclust(modelname, n)
```

where n is the number of clusters that you want to identify.

Alternatively you can cut the dendrogram at a specific height by adding the h argument.

```
> plot(modelname)
> rect.hclust(modelname, h=height)
```

3.5.K-Means Clustering in R

The k-means algorithm do not need to be dependent to an arbitrary variable unit, so we start by scaling the data using the R function *scale()* (Kassambara, 2017).

Required R packages and functions

The standard R function for k-means clustering is *kmeans*() , which simplified format is as follow (Kassambara, 2017):

kmeans(x, centers, iter.max = 10, nstart = 1) where :

- ***x**: numeric matrix, numeric data frame or a numeric vector*

- ***centers****: Possible values are the number of clusters (k) or a set of initial (distinct) cluster centers. If a number, a random set of (distinct) rows in x is chosen as the initial centers*
- ***iter.max****: The maximum number of iterations allowed. Default value is 10.*
- ***nstart****: The number of random starting partitions when centers is a number. Trying nstart > 1 is often recommended.*

To create a beautiful graph of the clusters generated with the *kmeans*() function, will use the *factoextra* package.

• Installing *factoextra* package as:

install.packages("factoextra")

• Loading *factoextra*:

library(factoextra)

Estimating the optimal number of clusters

The k-means clustering requires the users to specify the number of clusters to be generated. One fundamental question is: How to choose the right number of expected clusters (k)? Different methods can be applied. Here, we provide a simple solution. The idea is to compute k-means clustering using different values of clusters k. Next, the wss (within sum of square) is drawn according to the number of clusters. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. The R function *fviz_nbclust* () provides a convenient solution to estimate the optimal number of clusters (Kassambara, 2017).

```
library(factoextra)
fviz_nbclust(df, kmeans, method = "wss") +
    geom_vline(xintercept = 4, linetype = 2)
```

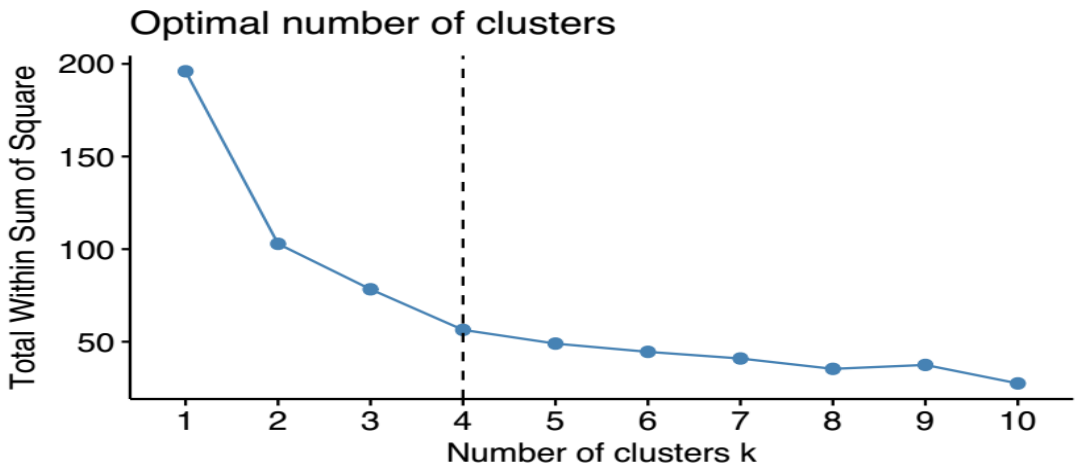


Figure 3.3: Optimal Number of Clusters

The plot above represents the variance within the clusters. It decreases as k increases, but it can be seen a bend (or "elbow") at k = 4. This bend indicates that additional clusters beyond the fourth have little value.

Computing k-means clustering

As k-means clustering algorithm starts with k randomly selected centroids, it's always recommended to use the *set.seed()* function in order to set a seed for *R's random number generator*.

For example, the R code below performs *k-means clustering* with k = 4:

```
# Compute k-means with k = 4
set.seed(123)
km.res <- kmeans(df, 4, nstart = 25)
```

As the final result of k-means clustering result is sensitive to the random starting assignments, we specify *nstart = 25*. This means that R will try 25 different random starting assignments and then select the best results corresponding to the one with the lowest within cluster variation. The default value of *nstart* in R is one. But, it's strongly recommended to compute *k-means clustering* with a large value of *nstart* such as 25 or 50, in order to have a more stable result (Kassambara, 2017).

Visualizing k-means clusters

It is a good idea to plot the cluster results. These can be used to assess the choice of the number of clusters as well as comparing two different cluster analyses. The function *fviz_cluster*() can be used to easily visualize k-means clusters (Kassambara, 2017).

CHAPTER FOUR

4. Results & Discussions

This chapter presents the generation of clusters of diseases using hierarchical and kmeans clustering algorithms. The hierarchical algorithm are tested and optimized with different linking methods and the kmeans algorithm is applied and it has been tried to identify the optimal clusters. Moreover, the results obtained from the each scenario is discussed.

Summary of the data & clustering techniques

Source Data	Addis Ababa Health Bureau
Data Management Tools	eHMIS
Year	2008 E.C
Programming Language used	R
Clustering Techniques used	Hierarchical clustering , Kmeans

Table 4.1: Data Source & Clustering Techniques

4.1. Hierarchical Cluster Analysis

In the dendrograms displayed below, each leaf corresponds to observation of each disease types reported for the year 2008 in Addis Ababa and Dire Dawa, for the top 20 reported diseases in terms of their disease counts. As we move up the tree in any dendrogram, observations that are similar to each other are combined into branches, which are themselves fused at a higher height. The height of the fusion, provided on the vertical axis, indicates the (dis)similarity between two observations. The higher the height of the fusion, the less similar the observations are. Conclusions about the proximity of two observations can be drawn only based on the height where branches containing those two observations first are fused. The height of the cut to the dendrogram determines the number of clusters obtained. It plays the same role as the number of clusters in k-means clustering. In order to identify clusters, we can cut the dendrogram with cutree() function.

- Import cluster library

```
library(cluster)
```

- Load the dataset

```
addis<-read.table("E:/corei5/diseaseData/DiseaseCluster/Addis2008Top20.txt",
header=TRUE)
```

- Scale the disease count dataset

```
skaddis=scale(addis)
```

- Computing dissimilarity matrix using Euclidean method

```
addisModel<-dist(skaddis, method="euclidean")
```

- Hierarchical clustering analysis using average Linkage

```
addisModelav<-hclust(addisModel, method="average")
```

- Plot the obtained dendrogram using average link

```
plot(addisModelav, main="Dendrogram of Top 20 disease counts")
```

- Cut the dendrogram to determine the number of clusters. The argument border is used to specify the border colors for the rectangles:

```
groups <- cutree(addisModelav, k=7) # cut tree into 9 clusters
rect.hclust(addisModelav, k=9, border=2:10)
```

- Hierarchical clustering analysis using single Linkage

```
addisModelsin<-hclust(addisModel, method="single")
```

- Plot the obtained dendrograms using single link

```
plot(addisModelsin, main="Dendrogram of Top 20 disease counts")
```

- Cut the dendrogram to determine the number of clusters. The argument border is used to specify the border colors for the rectangles:

```
groups <- cutree(addisModelsin, k=5) # cut tree into 7 clusters
rect.hclust(addisModelsin, k=7, border=2:8)
```

- Hierarchical clustering analysis using complete Linkage

```
addisModelcomp<-hclust(addisModel, method="complete")
```

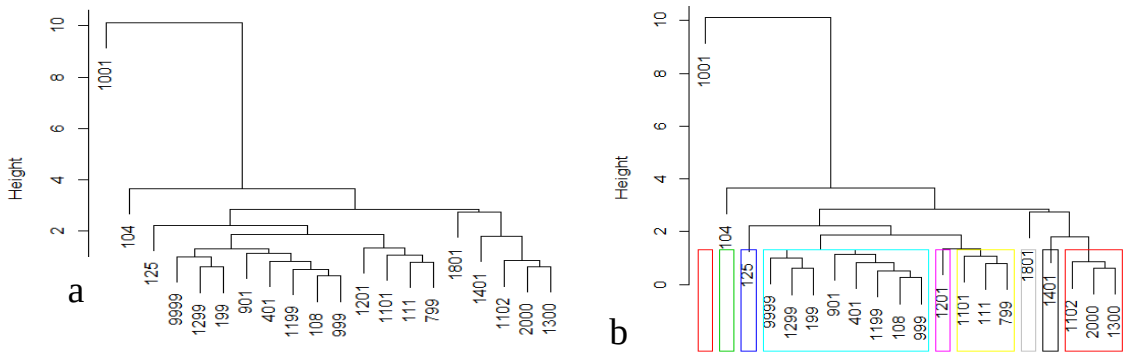
- Plot the obtained dendrograms using complete link

```
plot(addisModelcomp, main="Dendrogram of Top 20 disease counts")
```

- Cut the dendrogram to determine the number of clusters. The argument border is used to specify the border colors for the rectangles:

```
groups <- cutree(addisModelcomp, k=5) # cut tree into 7 clusters
rect.hclust(addisModelcomp, k=7, border=2:8)
```

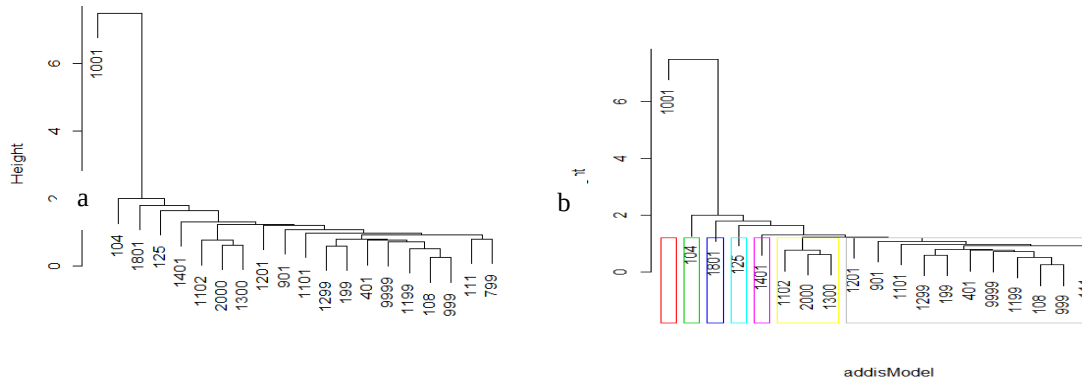
Hierarchical Clustering Using Average linkage method



Scenario 1: A dendrogram generated using Euclidean distance, hclust function and average linkage.

The dendrogram of scenario 1 reveals that disease code 1001, 104, and 125 are far apart from the other groups of disease codes. Dendrogram a and dendrogram b are the same, except that on dendrogram b cutree() function is applied to determine the number of available clusters. Here the first three clusters identified at the bottom of the dendrogram are by looking at which of the observed disease codes are fused first. Therefore, as we go from left to right, 1299 & 199 are fused and fall in one group, again 108 and 999 are also fused and fall in another cluster , the other which is also on the same level to fuse together are 2000 and 1300. As you can see from dendrogram b, if we put a box on some point between 0 and 2, we find about 3 clusters which relatively have larger number of diseases under each cluster. In these clusters of diseases, the number of diseases under each cluster ranges from three to eight diseases.

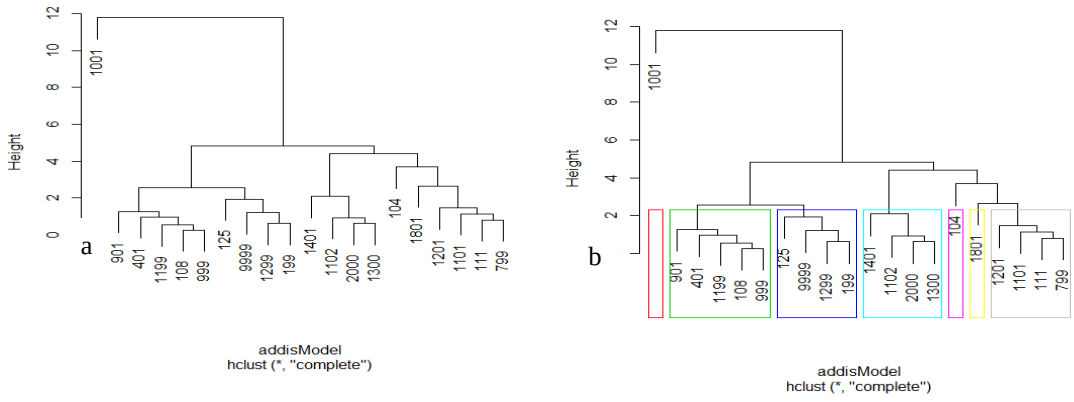
Hierarchical clustering analysis using single Linkage



Scenario 2: A dendrogram generated using Euclidean distance, hclust function and single linkage.

Here in scenario 2, disease code 1001 is too far from the other group of diseases. As we can see from the dendrogram here individual disease type join the tree step by step. This is often observed when we use single linking method called chaining. At the bottom of the dendrogram tree, when cutree() function is applied , two clusters of diseases are identified.

Hierarchical clustering analysis using complete Linkage



Scenario 3: A dendrogram generated using Euclidean distance, hclust function and complete linkage.

In scenario 3 we identified 4 clusters at the bottom of the tree using cutree() function with the smallest possible distance. So, based on observation of the dendrogram, the identified 4 clusters can be considered as the optimal number of clusters in this scenario. Moreover, as usual, disease code 1001 and 104 are far from the rest of the diseases, which indicates they are not similar with any other disease types and do not belong to any of the 4 clusters.

Summary of the Resulting Scenario

Scenario	Algorithm	Linkage Method	Number of Cluster
1	hclust	Average	3
2	hclust	Single	2
3	hclust	Complete	4

Table 4.2: Summary of the Resulting Scenario

As the summarized table 4.2 showed the application of hclust algorithm with varying linkage methods resulted in clusters ranging from 2 to 4. These resulting number of clusters may help to determine the number of k in the kmeans clustering method.

4.2.Kmeans Cluster Analysis

Here it has been attempted to generate optimal clusters using kmeans algorithm. In doing so the dataset are standardized, missing values are checked, distance calculation and optimization are carried out, and finally the results are displayed and explained.

```
library(cluster)
library(factoextra)
library(magrittr)
```

- load train data

  ```
  train <- read.csv("Addis2008Top20.csv")
  ```

- Standardize or scale the dataset

  ```
  Sktrain= scale(train)
  ```

- Check the missing values in complete dataset

table(is.na(Sktrain))

- Check missing values column wise

colSums(is.na(Sktrain))

- find distance

res.dist <- get_dist(Sktrain, stand = TRUE, method = "euclidean")

- Determining the optimal number of clusters

fviz_nbclust(Sktrain, kmeans, method = "wss")

- Compute and visualize k-means clustering

set.seed(123)
km.res <- kmeans(Sktrain, 3, nstart = 25)

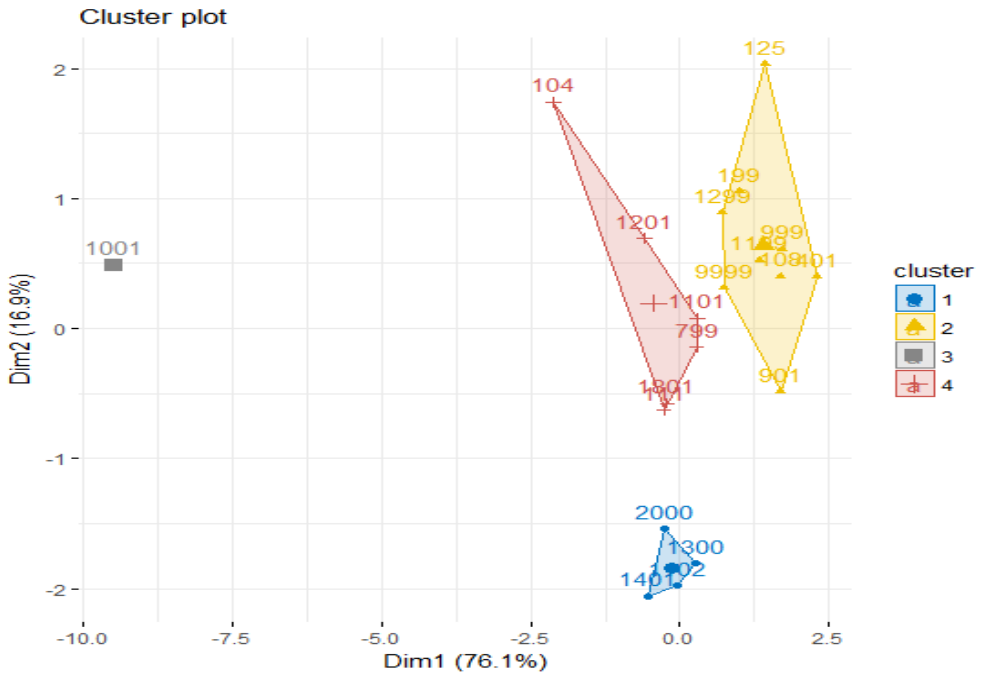
- Visualize

fviz_cluster(km.res, data = train, ellipse.type = "convex", palette = "jco", ggtheme = theme_minimal())

- Print the kmeans resuts

print(km.res)

The resulting cluster using Kmeans



Scenario 4: A disease cluster generated using kmeans

As shown in scenario 4, we find four clusters of diseases with the following detailed results.

K-means clustering with 4 clusters of sizes 4, 9, 1, 6

However, since the first cluster consists of only one disease it will not be considered as a cluster. The other three clusters relatively reveal groups of similar disease prevalence in the community.

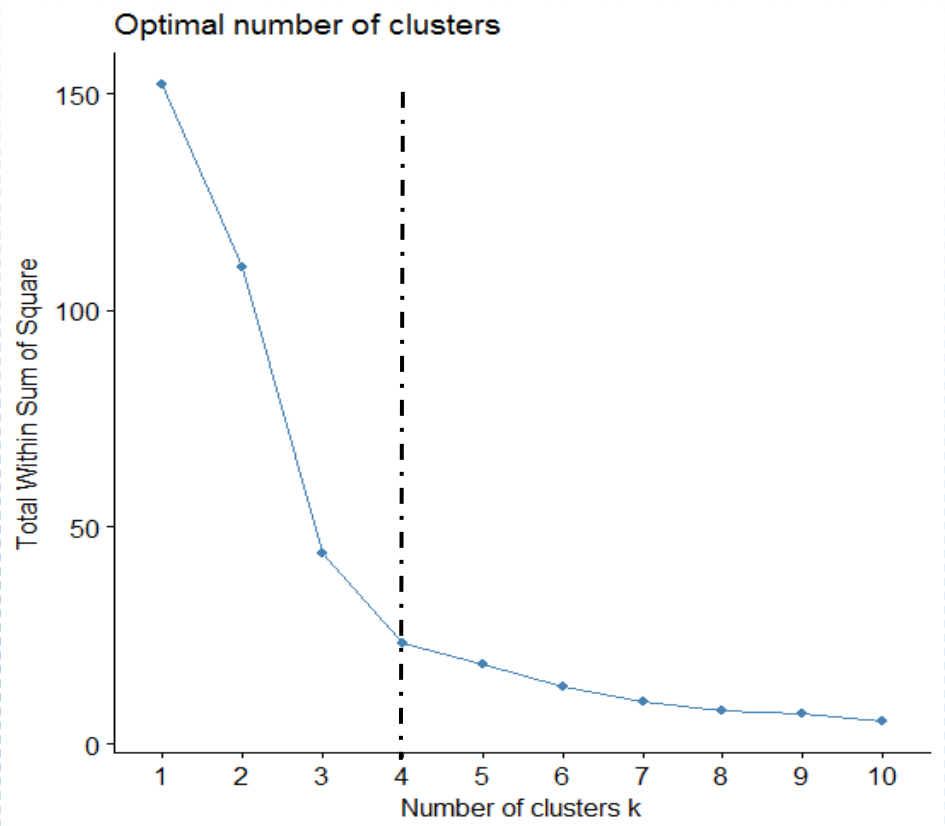
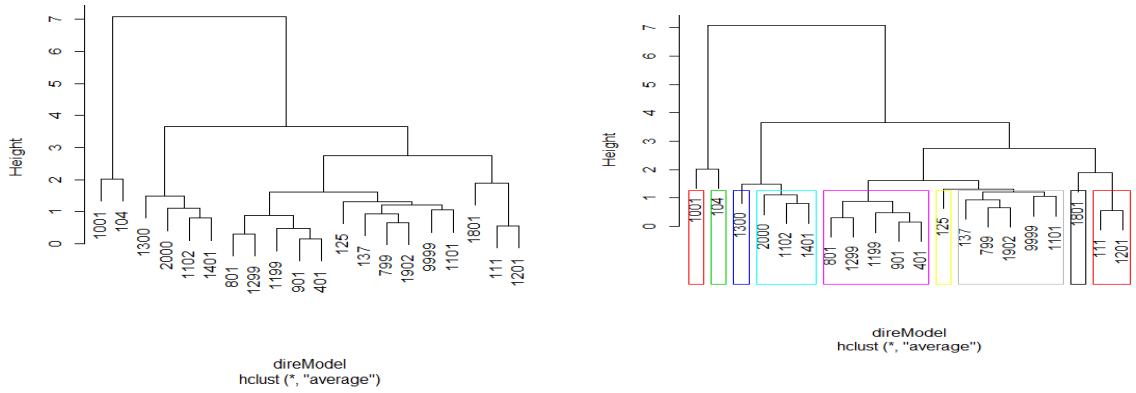


Figure 4.1: Optimal Number of Clusters

The plot in figure 4.1 represents the variance within the clusters. It decreases as k increases, but it can be seen a bend (or "elbow") at k = 4. This bend indicates that additional clusters beyond the fourth have little value.

Moreover, for the purpose of replication of the study, we performed the same techniques on other dataset which was obtained from Diredawa. In doing so , it has been checked the applicability of similar algorithms on different dataset and optimal number of cluster was also identified.

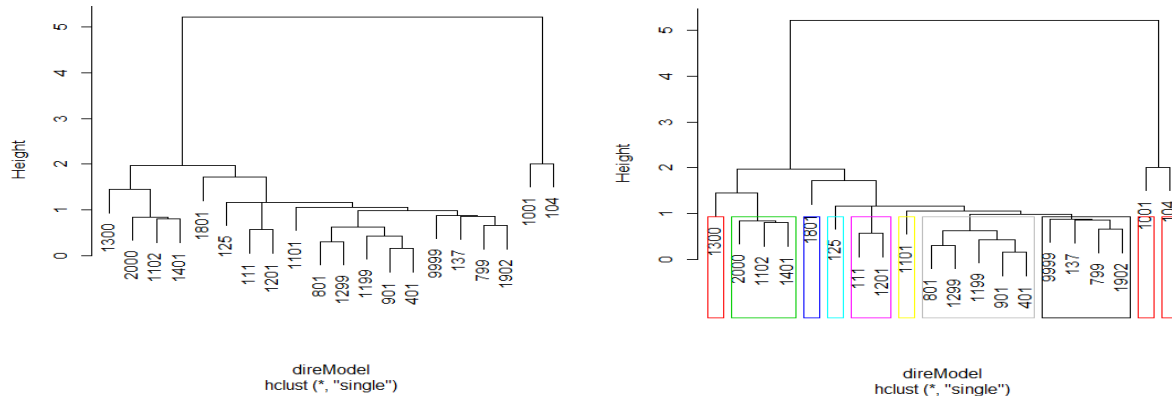
Hierarchical clustering analysis using average Linkage



Scenario 5: A dendrogram generated using Euclidean distance, hclust function and average linkage.

In scenario 5, after applying cutree() function , we identified four clusters of diseases. If we move a bit upper we may also find that disease code 1001 and 104 are fused and fall in the same cluster.

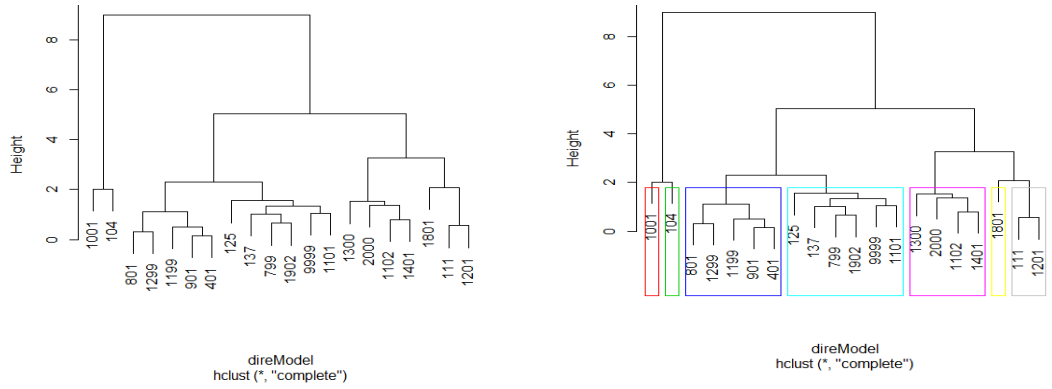
Hierarchical clustering analysis using single Linkage



Scenario 6: A dendrogram generated using Euclidean distance, hclust function and single linkage.

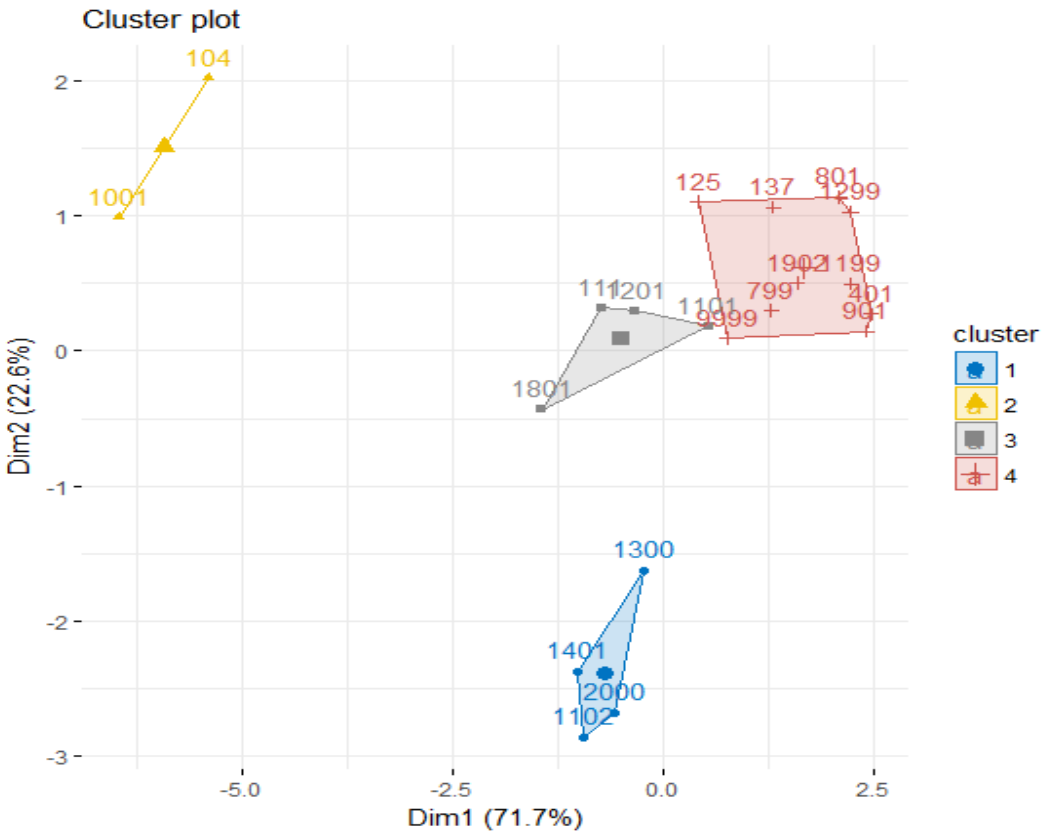
In this dendrogram, after applying cutree() function , we identified four clusters of diseases. Here, as we move a bit upper on the dendrogram tree, we may also find that disease code 1001 and 104 are fused and fall in the same cluster which most of the time these diseases were outliers.

Hierarchical clustering analysis using complete Linkage



Scenario 7: A dendrogram generated using Euclidean distance, hclust function and complete linkage.

The resulting Cluster using Kmeans



Scenario 8: A disease cluster generated using kmeans

As shown in scenario 8, we find four clusters of diseases with the following detailed results.

K-means clustering with 4 clusters of sizes 4, 2, 4, 10

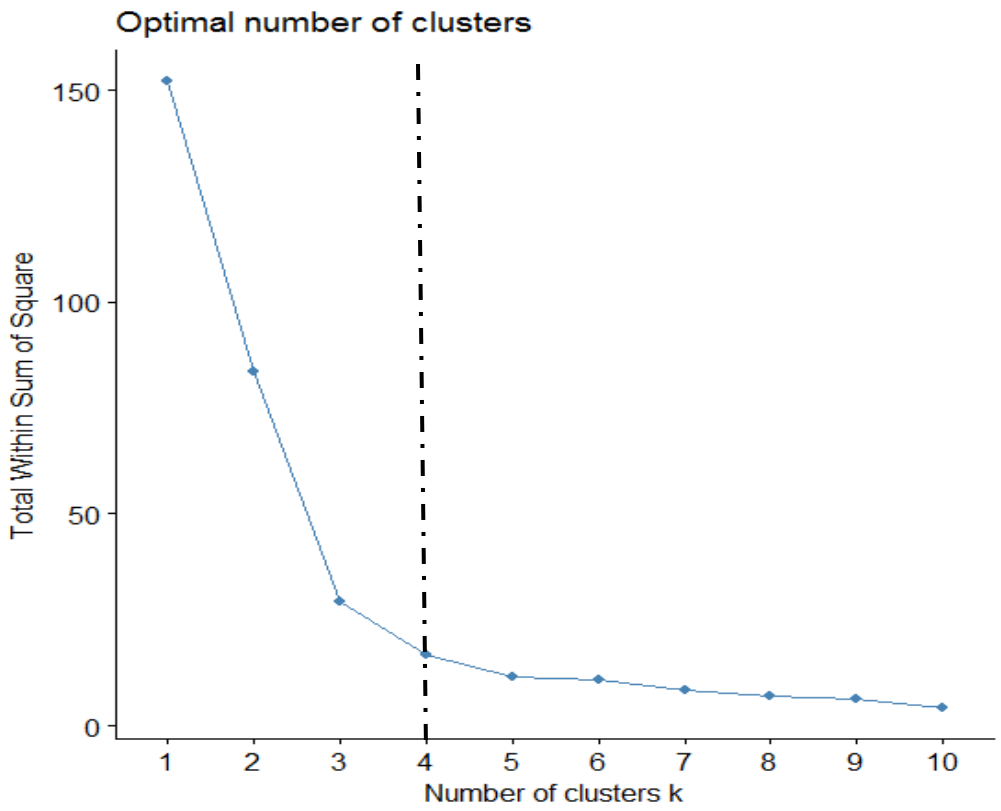


Figure 4.2: Optimal number of clusters

The plot in figure 4.2 represents the variance within the clusters. It decreases as k increases, but it can be seen a bend (or "elbow") at k = 4. This bend indicates that additional clusters beyond the fourth have little value.

4.3.Summary

To sum up, in this study it has been attempted to obtain answers for the following questions:

- Given the importance of EHR system, available infrastructure and skill, how is the level of acceptance &utilization at point of clinical care and for research?
- Which data analytics tools & techniques best applied on EHR so as to detect disease burden?
- Given EHR dataset and cluster analysis techniques, what is the optimal detection disease clusters?

The EHRs was seen as very important mainly in saving countless lives and millions of dollars. In the study area and may be in the regions equivalent to the study area, infrastructure and skilled manpower are available that can support the proper handling of EHR systems. All places we went for data collection have computer science graduates and ICT centers that support the functioning of the systems and the data processing tasks. The reporting system is stablished from bottom up with the support of guidelines and training. However, it has been seen on different national health policies and our observation, there is less use of the system especially the ERM by the physicians and other health experts.

Data analytics tools and techniques that can support EHR data analytics, especially for clustering, was also reviewed in this study. As the literature indicated, significant amount of software is available for data clustering. Most of the software on data clustering is open-source software, which is freely available. On the other hand, most of the commercial software comprises implementations of classical algorithms such as k-means or agglomerative clustering.

Finally, with the application of the selected clustering techniques, metrics, tools and dataset, it has been attempted to successfully detect optimal number of disease clusters and met the objectives of the study.

CHAPTER FIVE

5. Conclusion & Recommendation

This chapter presents the concluding remarks drawn as a results of experimentation and analysis and possible recommendations.

5.1. Conclusions

With advancement in data collection, storage and processing technology, researchers should have become access to more patient data than at any time. Nowadays, much of these data are not accessible as such and underutilized. The ability to harness the EHR would allow for continuous learning systems, and provide decision support for community health based on data from similar scenarios.

While EHRs have demonstrated potential to support public health practice like the pattern of the prevalence of disease in similar manner. However, there are limitations to more widespread public health use of EHR. With respect to data availability, EHRs found in Ethiopia are generally designed around the statistical reporting of disease counts, and do not include details of the patient population like psychosocial, behavioral, and environmental variables of interest to public health. Moreover, the data used for this study is aggregate lacking details of individual patient data at particular place, time and particular disease.

EHRs can support the practice of public health including public health surveillance, disease and injury investigation and control, decision making, quality assurance, and policy development.

The study was challenged in identifying to focus on selected disease categories, the availability of disease data in high level lacking details, and lack of EMR at hospital level

which can help to get detailed patient data. Moreover, it has been observed that EHR is not used for both clinical and research purpose as such.

However, in this particular study it has been tried to show the application of hierarchical and kmeans clustering algorithms provided that group of diseases that show similar prevalence in the population are detected. In the analysis of this study, it has been tried to identify from two up to four clusters of diseases and some of the diseases were observed to be outside of any other groups.

5.2.Recommendations

Researchers continue to ask fundamental questions of our health system, making use of the deluge of data generated by EHRs. Unfortunately, that huge amount is problematic. As research with EHRs continues to evolve more investment in research-friendly clinical databases as well as cross-institutional collaborations are required. It is time for healthcare to be rewarded from a rich data source that is already available.

With the increasing investment on Health Information Systems and development of supporting health management information systems policies in Ethiopia, the researcher strongly recommend the proper utilization of the available electronic health information systems both for research and clinical practice. Moreover, promoting the need to increase skills and proper utilization of electronic health systems should be carried out in parallel.

In spite of the challenges, availability of aggregate data and limited utilization of algorithms the results were encouraging. Hence, since the number of experiments and selected algorithms may not be exhaustive, we recommend interested researchers to investigate more in the area.

Finally, detecting cluster of diseases often may not necessarily lead to generalization. However, we recommend public health researchers to study why a certain group of diseases revealed similar behaviors at particular place and time in terms of prevalence.

Bibliography

Aggarwal, C. C., & Reddy, C. K. (2014). *Data Clustering Algorithms and Applications.* New York: Taylor & Francis Group.

Aldenderfer, M. S., & R. , K. B. (1984). *Cluster Analysis.* Newbury Park: Sage Publications.

Fang, R., Pouyanfar, S., Yang, Y., Chen, S.-C., & Iyengar, S. S. (2016). Computational Health Informatics in the Big Data Age: A Survey. *ACM Computing Surveys, Vol. 49, No. 1, Article 12*, 1-36.

FasterCures. (2005). *Think Research Using Electronic Medical Records to Bridge Patient Care and Research.* Washington: www.FasterCures.org.

FMoH. (2014). *eHMIS_PHEM training participant manua.* Addis Ababa: FMoH, CDC, Tulane University.

FMoH. (2014). *Health Sector Transformation Plan HSTP (2015/16 up to 2019/20).Draft v1.* . Addis Ababa.

FMoH. (2015). *Health Sector Transformation Plan HSTP (2015/16 up to 2019/20).* Addis Ababa: Federal Ministry of Health.

FMoH. (2016). *Information Revolution Roadmap.* Addis Ababa: Federal Ministry of Health.

Hartigan, J. A. (1975). *Clustering Algorithms.* New York: John Wiley and Sons.

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2016). *Handbook of Cluster Analysis.* New York: Taylor & Francis Group, LLC.

https://en.wikipedia.org. (n.d.). *wikipedia*. Retrieved December 26 , 2015, from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Disease_%28disambiguation%29

IBM. (2013). *Data-driven healthcare organizations use big data analytics for big gains.* New York: IBM Corporation.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1996). Data Clustering: A Review. *IEEE Computer Society*, 1-69.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Jensen, P. B., Jensen , L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *NATURE REVIEWS*.

Kassambara, A. (2017). *Multivariate Analysis I: Practical Guide To Cluster Analysis in R.* STHDA.

Lai, P.-C., So, F.-M., & Chan, K.-W. (2009). *Spatial epidemiological approaches in disease mapping and analysis.* USA: CRC Press.

Lawson, A. (1999). *Disease mapping and risk assessment for public health.* West Sussex,England: John Wiley & Sons Ltd.

Lawson, A. (2008). *Bayesian disease mapping : hierarchical modeling in spatial epidemiology.* Newyork: CRC press.

Mclafferty, S. (2015). Disease cluster detection methods: recent developments and public health implications. *Annals of GIS Vol. 21 No. 2*, 127–133.

Mooi, E., & Sarstedt, M. (2011). *A Concise Guide to Market Research.* Springer.

Nair, S., Hsu, D., & Celi, L. A. (2016). *Secondary Analysis of Electronic Health Records.* Cambridge: Springer Open.

Reddy, C. K., & Aggarwal, C. C. (2015). *Healthcare Data Analytics.* Boca Raton: CRC Press.

Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science.* Switzerland: Springer.

Weeger, A., & Gewald, H. (2014). Acceptance and use of electronic medical records: An exploratory study of hospital physicians' salient beliefs about HIT systems. *Health Systems*, 1–18.

WHO. (2013). *WHO Country Cooperation Strategy 2012-2015 Ethiopia. WHO Regional Office for Africa*. WHO.

Zaki, M. J., & Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms.* New York: Cambridge University Press .

Annex

Diseases with corresponding Code

The selected diseases for the study are highlighted.

Code	Disease_Facility
100	Priority infectious diseases
101	Malaria (clinical without laboratory confirmation)
102	Malaria (confirmed with P. falciparum)
103	Malaria (confirmed with species other than P. falciparum)
104	Diarrhea (non-bloody)
105	Diarrhea with dehydration
106	Diarrhea with blood (dysentery)
107	Meningitis
108	Typhoid fever
109	Relapsing fever
110	Epidemic typhus
111	Acute Febrile Illness (AFI)
112	Acute Poliomyelitis/Acute flaccid paralysis
113	Measles
114	Plague
115	Cholera
116	Yellow fever
117	Dracunculiasis
118	Neonatal tetanus
119	Viral hemorrhagic fevers
120	Avian Human Influenza
121	Rift Valley Fever (RVF)
122.1	Human immunodeficiency virus [HIV] disease
122.2	AIDS
123	Tuberculosis all forms
124	Leprosy
125	Pneumonia
126.1	Sexually transmitted infections: urethral discharge
126.2	Sexually transmitted infections: persistent / recurrent urethral discharge in men
126.3	Sexually transmitted infections: genital ulcer
126.4	Sexually transmitted infections: vaginal discharge syndrome
126.5	Sexually transmitted infections: lower abdominal pain syndrome (pelvic inflammatory disease PID)
126.6	Sexually transmitted infections: scrotal swelling syndrome
126.7	Sexually transmitted infections: inguinal bubo swelling (swollen glands)
126.8	Sexually transmitted infections: neonatal conjunctivitis
126.9	Sexually transmitted infections: neonatal herpes
127.1	Leishmaniasis (Visceral)
127.2	Leishmaniasis (Cutaneous and Mucocutaneous)
128	Onchocerciasis
129	Diphtheria
130	Pertussis
131	Tetanus (other than neonatal tetanus)

132	Trypanosomiasis
133	Schistosomiasis
134	Trachoma
135	Viral hepatitis
136	Rabies
137	Helminthiasis
199	Other or unspecified infectious and parasitic diseases
200	Neoplasms
300	Diseases of the blood and blood forming organs and certain disorders involving the immune mechanism
301	Anemias
399	Other or unspecified diseases of the blood
400	Endocrine, nutritional and metabolic diseases
401	Diabetes mellitus
402	Iodine-deficiency-related goiter
403	Moderate acute malnutrition
404	Severe acute malnutrition
499	Other or unspecified endocrine, nutritional, and metabolic diseases
500	Mental and behavioral disorders
600	Diseases of the nervous system
601	Epilepsy
699	Other or unspecified diseases of the nervous system
700	Diseases of the eye and adnexa
701	Cataract
702	Glaucoma
799	Other or unspecified diseases of the eye and adnexa
800	Diseases of the ear and mastoid process
801	Otitis
899	Other or unspecified diseases of the ear and mastoid process
900	Diseases of the circulatory system
901	Hypertension and related diseases
999	Other or unspecified diseases of the circulatory system
1000	Diseases of the respiratory system
1001	Acute upper respiratory infections
1002	Acute bronchitis
1003	Asthma
1004	Chronic obstructive pulmonary disease (COPD)
1099	Other or unspecified diseases of the respiratory system
1100	Diseases of the digestive system
1101	Dental and gum diseases
1102	Dyspepsia
1199	Other or unspecified diseases of the digestive system
1200	Diseases of the skin and subcutaneous tissue
1201	Infections of the skin and subcutaneous tissue
1299	Other or unspecified diseases of the skin and subcutaneous tissue
1300	Diseases of the musculoskeletal system and connective tissue
1400	Diseases of the genitourinary system
1401	Urinary tract infection
1499	Other or unspecified disorders of the genitourinary system
1500	Pregnancy, childbirth and the puerperium
1501	Medical abortion without complication (safe abortion)

1502	Other abortion (spontaneous, with complication etc.)
1598	Causes of abnormal pregnancy, childbirth and puerperium
1599	Other or unspecified obstetric conditions
1600	Certain conditions originating in the perinatal period
1700	Congenital malformations, deformations and chromosomal abnormalities
1800	Injury, poisoning and certain other consequences of external causes
1801	Trauma (injury, fracture etc.)
1802	Burns and corrosions
1803	Poisoning
1899	Other or unspecified effects of external causes
1900	External causes of morbidity and mortality
1901	Road traffic injuries
1902	Violence and other intentional injury
1999	Other or unspecified external causes of morbidity and mortality
2000	Factors influencing health status and contact with health services, including visit for examination and investigation (check-up, examination for driving license etc.)
9000	Other unclassified diseases
9001	District / region specific diseases - 1
9002	District / region specific diseases - 2
9999	Other or unspecified diseases

Disease Dataset Selected for the study- Addis Ababa 2008 E.C

These datasets are disease counts by sex and age group. The first column is disease code.

	Male0To4	Male5To14	Male>=15	MaleTot	Female0To4	Female5To14	Female>=15	FemaleTot
1001	28979	11257	31665	71901	24480	11287	43819	79586
104	14375	5069	13398	32842	11289	5176	15609	32074
1401	593	1436	16949	18978	965	2836	41962	45763
1102	129	1061	21197	22387	194	1739	34007	35940
2000	1931	1045	23653	26629	1957	1961	27681	31599
1300	203	710	22968	23881	257	1083	28746	30086
111	1993	2963	19292	24248	1650	3088	23459	28197
1201	5572	3701	13341	22614	5202	4454	18373	28029
1801	1723	3657	26954	32334	1186	2108	14106	17400
799	2542	2368	16660	21570	2251	2389	18965	23605
1101	627	3302	14096	18025	640	3985	19260	23885
9999	3872	1861	11669	17402	2911	1653	17931	22495
1299	3618	2354	9282	15254	3490	2994	14667	21151
901	45	134	13957	14136	93	243	18727	19063
1199	2243	1423	11422	15088	2031	1472	12983	16486
199	2187	2899	9128	14214	1814	3211	11708	16733
108	120	1389	10788	12297	152	1677	13100	14929
125	7551	1515	4956	14022	5972	1258	4944	12174
999	615	1242	10309	12166	747	1824	11349	13920
401	47	321	9732	10100	29	387	11564	11980

Disease Dataset Selected for the study- Dire Dawa 2008 E.C

These datasets are disease counts by sex and age group. The first column is disease code.

	Male0To4	Male5To14	Male>=15	MaleTot	Female0To4	Female5To14	Female>=15	FemaleTot
1001	1858	1068	2502	5428	1500	939	2775	5214
104	2287	815	1785	4887	1728	748	1960	4436
1102	22	153	2642	2817	13	178	3602	3793
1401	164	185	2133	2482	100	234	3658	3992
2000	91	118	2736	2945	53	48	3151	3252
1801	231	730	2376	3337	131	410	1693	2234
1300	33	138	2135	2306	36	280	2495	2811
111	452	444	1553	2449	351	427	1413	2191
1201	453	373	1404	2230	365	326	1389	2080
9999	232	255	890	1377	205	174	1541	1920
1101	57	377	1129	1563	37	344	1302	1683
125	557	274	778	1609	476	279	833	1588
799	171	170	1041	1382	137	159	917	1213
137	178	279	573	1030	135	306	598	1039
1902	14	265	937	1216	7	169	652	828
901	1	1	644	646	1	3	892	896
1199	67	58	694	819	45	64	580	689
401	2	8	596	606	0	14	783	797
801	188	163	308	659	124	141	403	668
1299	157	113	334	604	166	88	421	675