Hidden Pattern Identification Using Spatial Clustering In Centralized Health Care Architecture for Arrhythmia Detection

By

Jawairia Rasheed

2009-NUST-MSPHD- CSE (E)-14

MS-09 (CSE)

Submitted to the Department of Computer Engineering in fulfillment of the requirements for the degree of

MASTER OF SCIENCE
In
SOFTWARE ENGINEERING

Thesis Supervisor

Dr. Shoab Ahmed Khan

College of Electrical & Mechanical Engineering

National University of Sciences & Technology

2012

DECLARATION

I hereby declare that I have developed this thesis entirely on the basis of my personal efforts under the guidance of my supervisor Dr. Shoab Ahmed Khan. All the sources used in this thesis have been cited and the contents of this thesis have not been plagiarized. No portion of the work presented in this thesis has been submitted in support of any application for any other degree of qualification to this or any other university or institute of learning.

Jawairia Rasheed

ACKNOWLEDGMENT

First of all, I am thankful to Allah Almighty for His blessings and bounties at each and every step. It would have been impossible to achieve this milestone without His consent. who listened to my pray and written the degree from here in my fate.

I feel obligated to my Supervisor ***Dr. Shoab Ahmed Khan*** for his tremendous assistance and remarkable supreme guidance which has always been a direct source of inspiration and motivation for me. I am especially thankful for his sincerity and dedication for not only my thesis but also throughout the entire course of study period. He always provided me his kind help at every step whenever and wherever I needed. Thank you sir it would not have been possible without your support. I would like to thanks *Lecturer* ***Sajid Gul Khawaja*** for his assistance during thesis.

I would like to thanks my siblings and family for all their love and support. For my parents who believe in me and my efforts during my entire life. And most of my loving mother who financially and morally supported me throughout my life, who wants to see me successful and even getting more education. It’s all because of you.

To ALLAH…

ABSTRACT

The thesis aims to develop a way of finding out the trends that might exist between different cardiovascular diseases(CVD) and the socioeconomic, demographic and medical history of a patient, thus enabling the health sector to prevent large number of deaths caused by such diseases each year. To achieve this goal the current structure of healthcare sector needs to be somewhat changed and hence this thesis also suggests a new centralized healthcare structure. The thesis not only suggests ways of centralizing the healthcare structure of the country but also shows the ways in which it can be used in reducing CVD. A large portion of the paper deals with reading the acquired ECG of a patient the way Minnesota ECG Code elaborates and then using this data to find the possible arrhythmia the patient might be suffering from. After gathering this information data mining algorithms are used to find the possible trends between these diseases and other aspects of patients life. By identification of these patterns possible preventive measures and reorientation of health care facilities are suggested for the betterment of society.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF GRAPHS

Chapter 1

1.Introduction

According to 2007 World Health Organization estimation 17.5 million deaths occur every year across the world and one out of three in the world are caused due to cardiovascular diseases (CVD) such as stroke, heart attack or heart failure. The report also stated that not less than 20 million people have an acute atherosclerosis or Ischemia but survive. The occurrence of death due to CVDs is eighty percent and CVDs are burdening globally 86 percent in developing countries[1]. The leading cause of death in developing countries including Pakistan is CVD due to the prevalence of hypertension over the age of 50. One of the risk factor burdening the cardiovascular disease is lack of adequate data and pertinent disease history of the patient. Most of the primary physicians do not have recent knowledge about the managing of CVD. Public also don't have sufficient knowledge about the prevention of CVD even many people are unaware about it. There are no records among specialists about managing hypertension in Pakistan. Non-Communicable Diseases (NCDs) and injuries are thought as main causes of morbidity and mortality in Pakistan. Through estimates it has come to known that they are accounting for approximately 25% of the total deaths within the country. CVDs are significantly increasing the death ratio of individuals and regarding economic perspective burdening heavily societies, nations and health sectors. In majority of cases, economically productive workforce suffers from these diseases. Population based existing data on morbidity relating to CVDs in Pakistan indicates that over the age of 45 years one out of three adults suffers from high blood pressure. Similarly the survey conducted by the College of Physicians and Surgeons Pakistan give an even alarming situation. According to the survey in Pakistan more than 50% of deaths in the adult population are due to Cardiovascular Diseases and about 30% of all deaths are due to cardiovascular causes[1].

According to WHO report *Global atlas on cardiovascular disease prevention and control* leading causes of death and disability of the world are due to cardio vascular diseases. No doubt preventive measures can be adopted and can be saved from CVDs but preventive measures are inadequate [2].

Heart Disease spares no age. About 1% of all babies born in Pakistan have Congenital Heart Disease. About 5% of school age children have Rheumatic Heart Disease. About 14% adult population of Pakistan suffers from High Blood Pressure (Hypertension) and Coronary Artery Disease (Angina and Heart Attacks). These programs therefore acquire especial importance for the healthcare providers [3]. This condition is made even worse by an consistently ailing healthcare infrastructure of the country. According to 2007-2008 Economic Survey of Pakistan for every 1225 patients there was only one doctor to treat them. But in reality there is a much worse distribution of doctors between rural and urban patients. About 75% of Pakistan's population resides in rural areas. A travel of several kilometers and an expense of quite an amount is needed for a patient to visit the doctor , even for mere disease diagnosis, which delays a patient visit to doctor, leading to costly complications at a later stage resulting into hospitalizations. On the other hand the lack of centralization of healthcare system in Pakistan is causing great deal of hindrance if effectively fighting the disease and efficiently utilizing the available resources.

The gap between the doctors and patients is increasing on the other hand, availability of specialized and trained doctors and lack of awareness in the masses are major factors contributing towards increasing death rates due to diseases like hypertension. Every time a patients enters the hospital he has to go through a tough process of getting into medical records of the hospital before he can actually see the doctor. Even if the patients had been to the same hospital before, he has to get registered every time he comes to hospital and get what is call a "token" to get in the waiting list. On the other hand most of the hospitals take no pain of saving patients record and even if they do the records are never consulted if the patient returns to the hospital. This not only makes the process slow but also in many cases proves to be an important factor in the death of the patient. Unavailability of patient's complete medical history prevents the doctor from identifying the cause of disease and hence death may occur.

The health care sector of Pakistan has not been concerned with research for the most of its history. Even though the trends are changing now, no data about the patients exists, and even if it does it's scattered. There is no central repository from where researchers could extract data and use it for scholarly purposes. No record of the patients is present to extract trends of the prevalent diseases. Similarly the absence of centralization of healthcare infrastructure also is one of the main reason of the lack of consensus that exist between the leading hospitals of the country about the burden caused by the CVD on the country's socioeconomic condition. As the number of heart patients in Pakistan is increasing and the resulting mortality rate is going up. Cardiologist and other specialist doctors cannot be appointed to far flung areas due to low patient to doctor ratio. Hence an alternate mean of making doctors available to patients in such areas need to be devised. The importance of medical records of the patients cannot be neglected as it is important for not only helping treat the patients better but also in the prevention of disease. Centralizing the hospitals and healthcare units would not only serve the purpose of storing and the patient's medical record but it will also provide access to patients in the remote areas to specialist doctors hence eliminating the unequal distribution of doctors in the country.

The records would then be used for trend analysis to find the trends that might exist between diseases and different socioeconomic factors and hence provide a chance to fight them in much organized and effective way. This centralization would also prove helpful in carrying out population surveys and better estimating the distribution of different diseases in different population groups and the similarities and differences that might exist in these groups.

1.1 The Problem

The question then arises from where should we start? According to Vice Chancellor Dow University of Health Sciences (DUHS) Professor Masood Hameed Khan there is an immediate need to solve the problem caused by CVD in Pakistan, where cardiovascular disease are causing deaths of more than 50 percent of adults. This is very high alarming rate in Pakistan especially in the productive year of life. These deaths have strong socioeconomic consequences. Most of these deaths deprive the family of their sole bread earners and it is estimate that 72% of the deaths caused by CVD are of males. These numbers though shocking, but provide us with evidence that there lies clear trends between CVD and the medical and socioeconomic conditions of the patients and hence details study of the population can clearly outline these trends and hence provide us with a solution to fight these large number of deaths of the economic building force of the country.

1.2 The Solution Strategy

Today Pakistan has one of the largest healthcare infrastructure but due negligence of the concerned authorities the system is not being used efficiently. People are provided healthcare through a three tiered health care infrastructure and various facilities provided by health interventions. The first includes Basic Health Units and Rural Health Centers, providing the core of the basic healthcare structure. Secondary care provides not only basic facilities but also acute, ambulatory and inpatient care in district hospitals by tertiary care through teaching hospitals. The network of health services include 945 hospitals, 5349 basic health units and Sub Health Centers, 562 Rural Health Centers, 4755 Dispensaries, 903 MCH Centers and 290 TB Centers [4].

But all above mentioned work independently and no co-ordination exists between these units. Due to this not only patients have to suffer but the hospitals themselves have to go through a lot of undue trouble. In this age of technology and internet unfortunately the country's healthcare system still lingers in the 50s when it comes to communication and information sharing between hospitals. The need of taking the health care system into the new era of communication for fast and reliable information sharing is inevitable.

Though the detail structure and architecture of the centralized storage area network between hospitals and different units of the health care system will be explained in the coming chapters in detail, given below is architecture of for the sake of better understanding of the solution. Architecture of the complete system with remote patients connected with the central hospital via hybrid network of LAN, Dial-Up, GSM/GPRS, Radio and Satellite. National repository is a large storage that will store the raw data of all the patients using each patient's CNIC number as his or her unique ID. This will save the patients from waiting in long lines to get registered at hospitals every time he or she returns to the hospital. On the other hand the doctors will be able to retrieve the data of the patients including their complete medical history by simply using this unique ID for the desired patients.

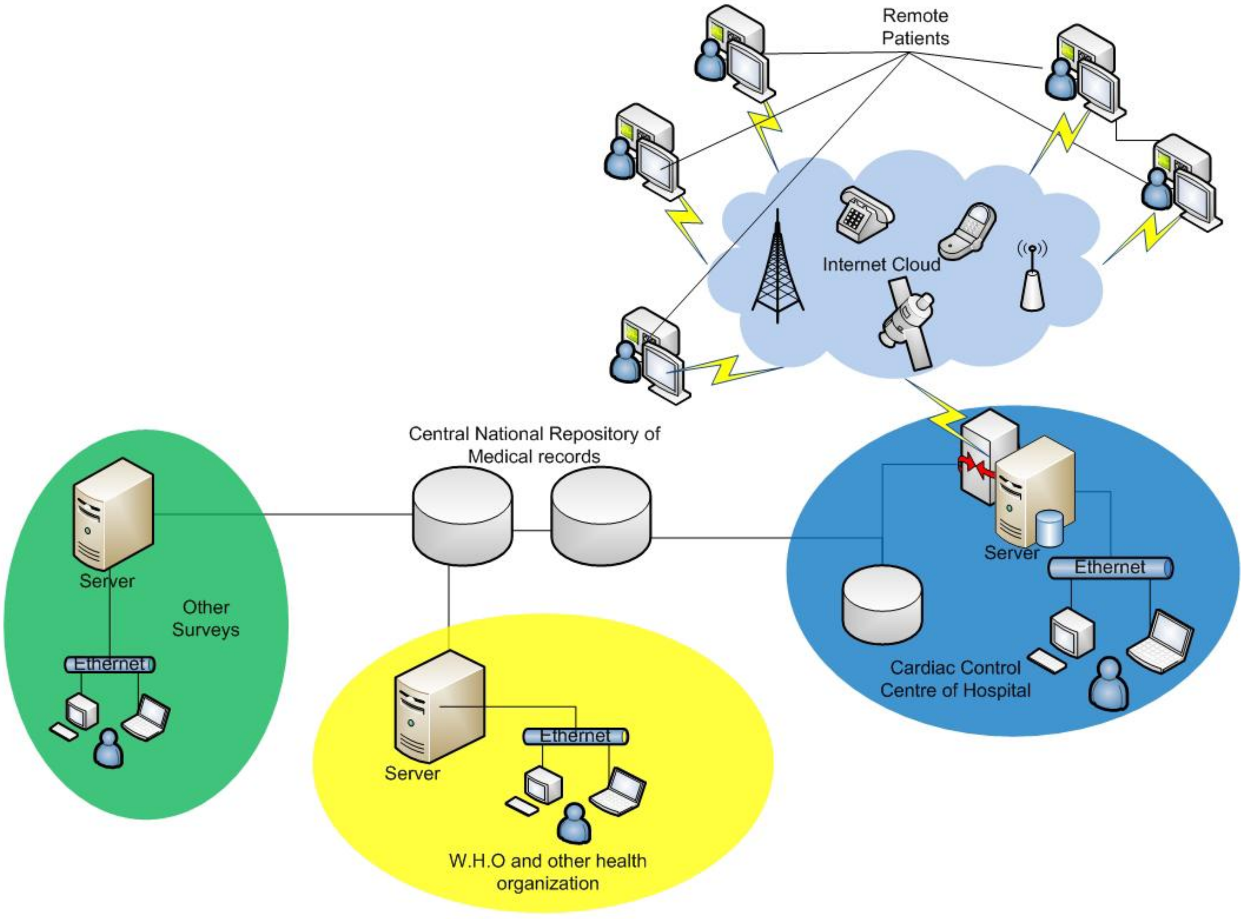


Figure 1: Architecture for Centralized Health Care System

But this in only one aspect of the architecture here it should be kept in mind that centralizing the hospitals is not the main objective rather using the data gathered by centralizing the health care system to help fight CVD is the main focus. But how will that be achieved? As it can be seen the gateway at the Hospital server receives physiological information from remote locations and generates an ECG work-list and makes automatic doctor assignments. Network Area Storage (NAS) keeps record of patient's data for future reference and data mining. This is the actual brains of the system. The data mining will be used to carry out trend analysis to find out hidden patterns that might exist between the physiological, demographical and socioeconomic information of a patient. Complete patient management System (PMS) and diagnostic software is based on HL7 standards.

One of the main apart and the core element in the system in the ECG and the data retrieved from it. This is the core element in identifying the nature of CVD that are prevalent in certain population. Hence the system must be able to store and communicate the ECG data. For this purpose understanding ECG is ever important and the book covers the main aspects of ECG in coming chapters. Other than ECG, a detail of patient medical, social, economical and demographical background is of utmost importance. As this is one of the first of its kind research in Pakistan no such form or study exist and hence the research requires making a form of its own that will be used for this research as a start and will be further developed with time. The form is provided in appendix A. The details about how the form was summed up will be provided later in the book.

1.3 The Path To Achieving Solution

The solution provided in previous section and the process of achieving it will be dealt in detail in the later chapters of the book. Here we provide a brief description the studies conducted and the process involved. The first and the foremost part of the research is to develop a form that is comprehensive in every aspect. That is it includes not only the medical and the ECG data related to a subject but also its Demographic, Social and Economic background. The main objective of the research is to find the hidden relation that might exist between CVD and these aspects of a subjects life style. Designing such form requires extensive study and research especially when it is being designed for a country for which no such study or form exists. Given below are brief introduction to the studies involved in designing of these forms. Later in this thesis each of these studies are discussed in detail.

1.3.1 Framingham Heart Studies

In 1948 identifying CVD as one of the main causes of death in America, the university of Boston and National Heart, Lung and Blood Institute started a revolutionary research in health research called the Framingham Heart Studies (FHS)[5]. The objective of the Framingham Heart Study was to identify the common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke [4a].

Hence FHS serves as a starting point for the development of a survey form that includes all the features needed to conduct a population study. FHS were carried out over a very long period of time and included three generation of participants. Though the methodology of this research may differ from that of the FHS but the objectives of both are very much the same; to identify the factors contributing towards CVD. And hence the types of survey used by FHS provide a good starting point for developing a survey form to conduct such studies in Pakistan but using rather latest techniques of data mining rather than observing three generations of patients with CVD history.

1.3.2 Minnesota ECG Code

Minnesota Code formulated by University of Minnesota is used mainly for population research and clinical trials and not for hospital practice. The code has been in use for more than 40 years, developed in the late 1950s by Dr. Henry Blackburn in response to the need for reporting ECG findings in uniform, clearly defined, and objective terms, is the most widely used ECG classification system in the world for clinical trials and epidemiologic studies. It incorporates ECG classification criteria that have been validated, widely employed, and accepted by clinicians. The Minnesota Code provides an objective ECG classification system free of impressionist physician bias, by which different studies can have a common standard to compare or pool ECG findings. The code is used for the population study in this research. The clusters for the data mining are formed on the basis of the results MN Code. The code is provided in appendix B. a detailed study of the code and its automation is provided in the coming chapter.

1.3.3 Accessing the Repository

The application to be developed for the National Repository can follow two models, either that of a client server application which will be useful if the application is to be used only over a single network. This model is most widely used in the world today. The client application on a workstation within the network accesses the functions of the National Health Repository by connecting and querying the server application running on the server within the network. In this case the server would be the Data Mining.

Engine which connects with the SAN, NAS or a SAN-NAS hybrid at the other end to access the health record. Another implementation of the client server model is possible. The client could connect to the server using the provided APIs and the web address of the server over the internet. This would allow the clients to be more remote and not necessarily attached in the network. However in order to access the National Health Repository the client end application would be necessary. The client application would handle all the user end processing and presentation of the data and would also be responsible for attaining the data from the Data Mining Engine using the web APIs built for this specific reason.

The most simple and practical application of the national health repository would be over the internet in the form of a website. This would unburden the clients from the use of a specific application and would enable multi platform access to the Health Repository without restriction to Operating System or Device from which the Repository may be accessed. The Data Mining Engine in this case would also be used as the Web Server and would handle all the processing and presentation as well as the Data Mining Operations. Though certain tradeoffs between cost, reliability and data security exists between these different ways of accessing the central repository. Each of the above technique is discussed in detail in later chapter and a conclusion is developed in light of all limitations and requirements.

1.4 Expected Improvements

Engineering and healthcare industries cannot be thought of separate fields anymore. The world identified the need of introducing latest engineering techniques to the healthcare industry and it's about time we start working on these lines too. The health infrastructure of Pakistan is not being used effectively and efficiently. The potential of this vast health care infrastructure is being wasted. But use of latest engineering techniques can help use the present facilities to their fullest. The thesis introduces the latest engineering techniques to the healthcare setup in Pakistan to improve the healthcare facilities and fully utilize the facilities present. At the mean time the thesis aims to open the gateway to medical research in Pakistan and finding the hidden factors behind the disease pertinent to different areas. So the preventive measures can be taken by the government in the reduction of disease and facilities for the cure of disease can be provided.

According to 2008 WHO survey, during the year of 2008 total of 57 million deaths occurred in the world ;among them 36 million that are about 63% of it were due to non communicable cardiovascular diseases (NCD). Approximately 80% of these NCD deaths (29 million) occurred in low and middle income countries. NCDs are the most frequent cause of deaths in most of the countries. The prevalence of NCDs is rapidly increasing and is researched to cause almost three quarters as many deaths as other diseases by 2020.

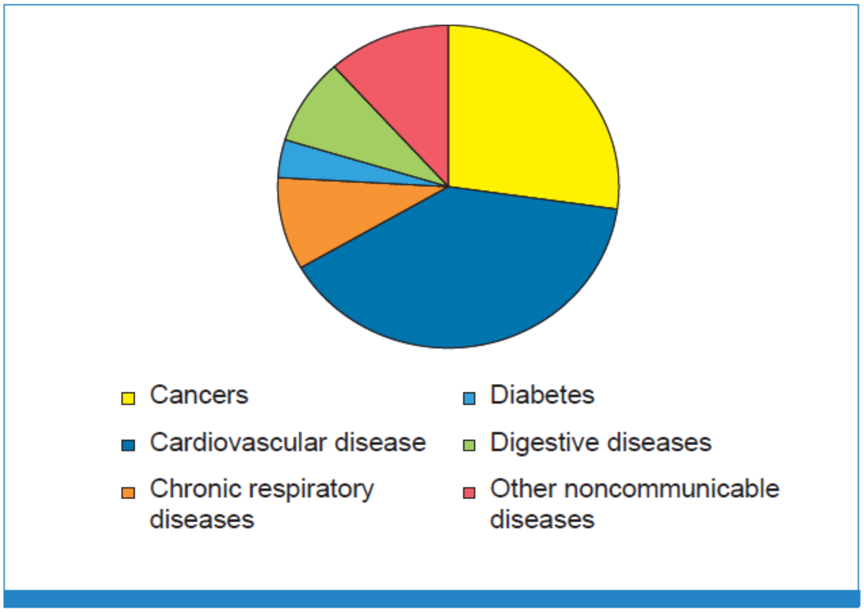


Figure 2. Global NCD death ratio under the age of 70 during 2008

There is an immediate need of reducing the number of these deaths and the only way to do this is by improving the healthcare setup of Pakistan and eradicating the underlying causes . 22% of the deaths caused by NCD are due to CVD and the what is more alarming is that this 22% is 72% of all the adults deaths in the country[6]. Thus CVD is eating up the nation building force of the country. These disease are not diseases for which there is no cure, these are simple disease like hypertension and high blood pressure, but they prove to be fatal due to lack of preventive and counter measures. The thesis aims to provide the health care sector indentify reasons that are major contributing factors toward the rise of deaths due to such diseases. This is the direct outcome of the thesis.

1.5 Risk factors

Most preventable risk factors associated with CVD are related to four particular behaviors: use of tobacco, being physically inactive, malnutrition and harmful use of alcohol. These behaviors are leading cause of four key metabolic/physiological changes: high blood pressure, obesity, high cholesterol level and hyper diabetes. In terms of attributable deaths, the leading CVD risk factor is raised blood pressure (which lead to 13% of global deaths), followed by tobacco use (9%), high blood glucose (6%), lack of physical activity (6%) and (5%) obesity [7]. some of the diseases are caused by the occurrence of other diseases or the abnormalities in the coronary arteries of heart[8].

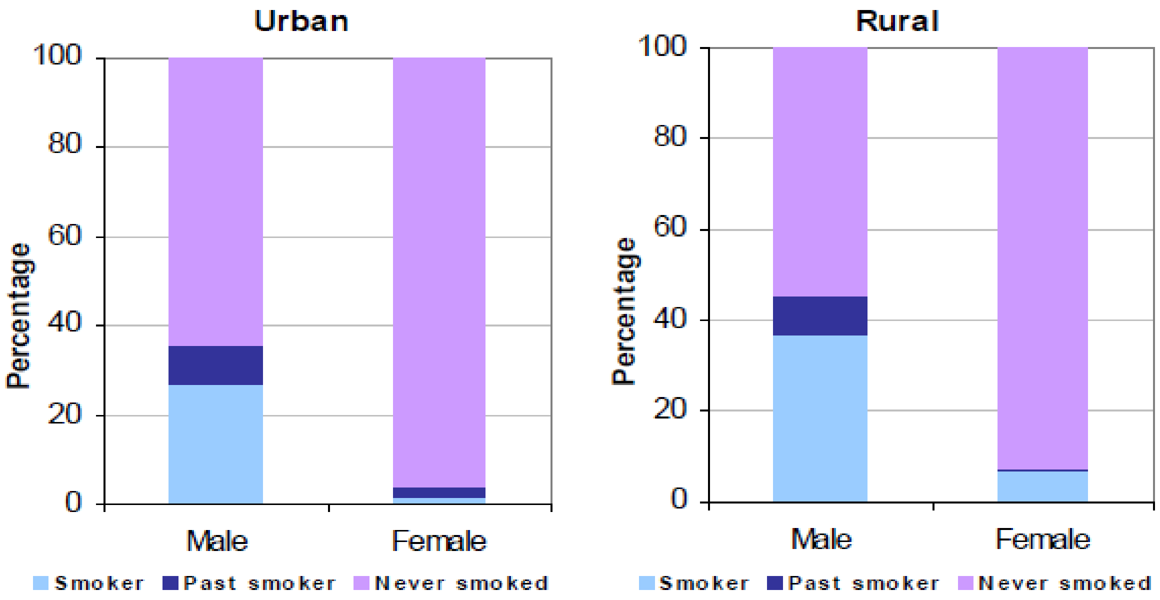
As hypertension and other socio-economic factors are also directly related to CVD, so combined analysis of behavioral risk factors and socio-economic factors relevant to geographical areas can help us finding different patterns of disease in different areas. As discussed earlier there is unequal distribution of medical facilities in Pakistan due to unavailability of doctors, so these patterns can be useful for not only providing the medical facilities but also highlighting the factors in front of government for eradicating the risk factors.

1.5.1 Smoking

Numerous surveys have been made at many times in Pakistan to estimate the prevalence of tobacco use part of other knowledge, attitude and practice. However, their results are not strictly comparable due to differences in evaluation methodologies, instruments, and the geographic boundaries of the surveys. For example, 10 years ago, National Health Survey revealed the consumption of tobacco 54% by men and 20% by women. More recently, the NAP-NCD First Round of Surveillance showed prevalence of 33% in men and 4.7% women. By comparing the first round and second round of surveillance it is concluded that there has been a decrease in prevalence of smoking. However, presumptions can be made according to the given the geographical variations in both surveys (the former being national and the latter one being regional), methodological differences and subtle incomparability’s in the assessment tools and definitions.

Prevalence of Smoking

Definition: Percentage of the population smoking tobacco (cigarettes, *beeri*, *hukka*) on a daily or occasional basis vis-à-vis those who smoked in the past and those who never smoked.



Graph1: NCD- Prevalence of smoking – by place of residence and gender

Smoking Status	Urban			Rural		
	Male	Female	Total	Male	Female	Total
Smoker	27.3	1.7	11.8	36.6	6.9	19.7
Past Smoker	8.4	2.2	4.6	9.1	0.2	4.1
Never Smoked	64.3	96.1	83.6	54.3	92.9	76.1

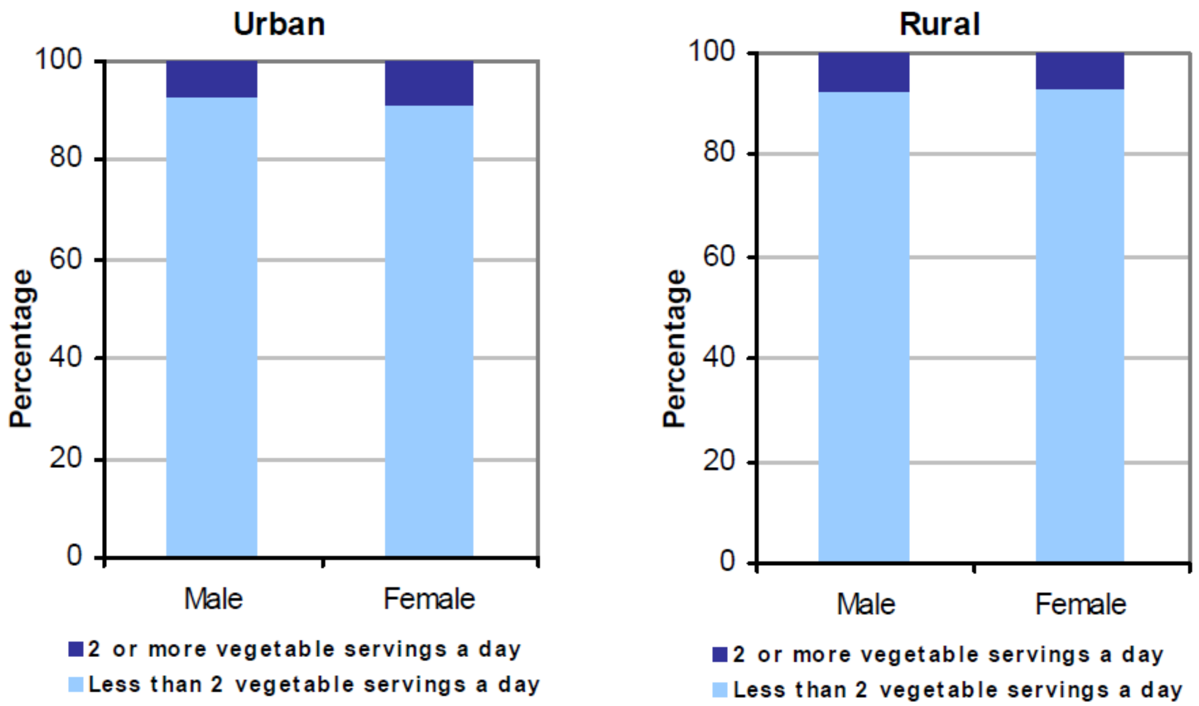
Table1. NCD Prevalence of smoking by place of residence and gender

1.5.2 Diet and physical activity

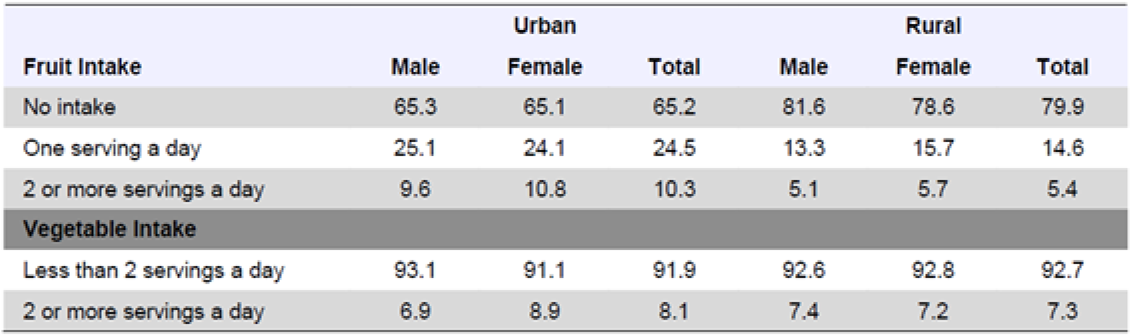
First Round of Surveillance of NAP-NCD was the survey first time conducted for accumulation of population based data on diet and physical activity using authentic tools according to parameters useful for measuring cardiac disease related risk behaviors. The data showed that more than 65% of the urban and 79% of the rural population take less than one serving of fruit a per day and 90% of Pakistani population uses less than two servings of vegetables per day. According to these trends clear instructions can be made for potential interventions in the area of NCD prevention in terms of behavior modification on one hand, and revise policies to make fruits and vegetables more affordable and accessible.

Diet Plans

The graph shows the number of vegetable servings consumed per day when considering by place of residence and gender



Graph 2. Percentage of the population consuming fruits and vegetables per day – by place of residence and gender.

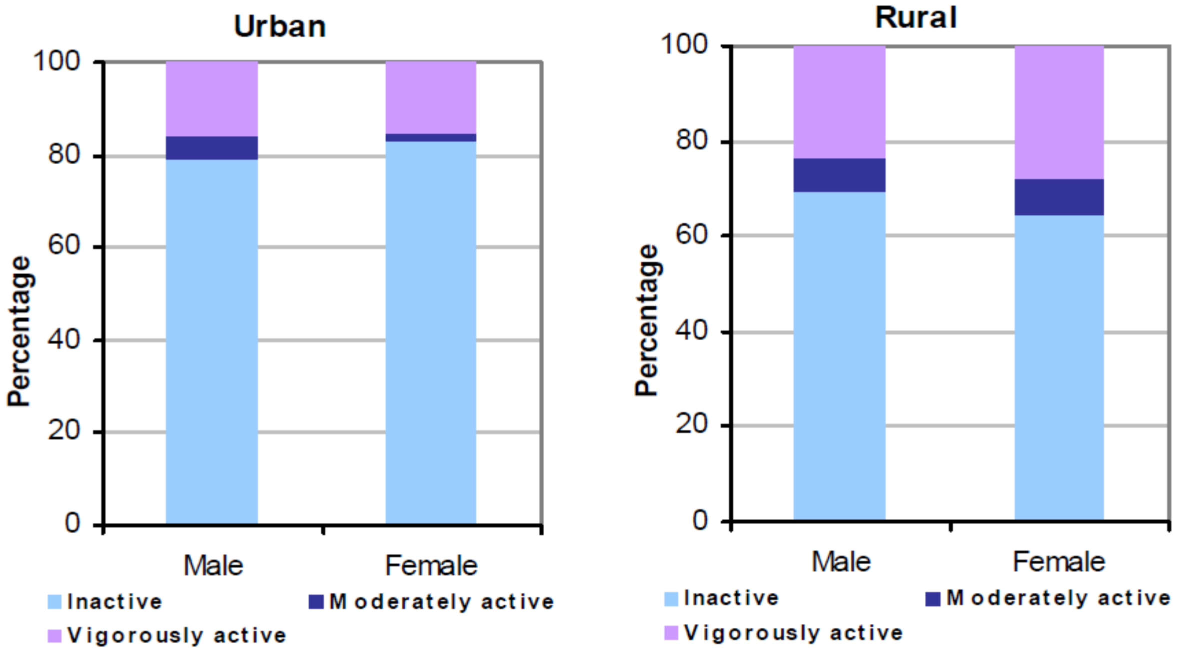


		Urban			Rural	
Fruit Intake	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
No intake	65.3	65.1	65.2	81.6	78.6	79.9
One serving a day	25.1	24.1	24.5	13.3	15.7	14.6
2 or more servings a day	9.6	10.8	10.3	5.1	5.7	5.4
Vegetable Intake						
Less than 2 servings a day	93.1	91.1	91.9	92.6	92.8	92.7
2 or more servings a day	6.9	8.9	8.1	7.4	7.2	7.3

Table 2. NCD Percentage of people utilizing fruits and vegetables per day by place of residence
. and gender

Physically active during work

Definition: Percentage of the population physically active during work, based on the GPAQ STEPS module



Graph 3. NCD-Level of physical activity during work - by place of residence

Degree of Physical Activity	Urban			Rural		
	Male	Female	Total	Male	Female	Total
Inactive	79.2	83.0	81.5	69.4	64.3	66.5
Moderately active	5.0	2.0	3.2	7.0	8.0	7.6
Vigorously active	15.8	14.9	15.3	23.6	27.7	25.9

Table 3. NCD- Level of physical activity during work, expressed in degree of activity

1.5.3 Obesity

From the NAP NCD First Round of Surveillance Obesity has been defined according to the WHO as well as the Asian criteria. According to this, in the district of Rawalpindi more than 28.4% of population living in urban areas and 23.3% of population living in rural areas was labeled as being overweight whereas 17 % and 7.9% in the rural and urban areas were found to be overweight. Thus, in the urban area a total of 45.8% and 31.2% in the rural areas were obese. According to Asian criteria, figures were higher than this: 63% city and 49% village population were found as being overweight. The same sources also provided data for central obesity. This is a major trend since central obesity is a more dominant risk factor for coronary heart disease than total body obesity, and has more close association with cardiovascular disease risk factors discussed than overall obesity as calculated by BMI in studies on the Pakistani population.

Prevalence of Hyperlipidemia

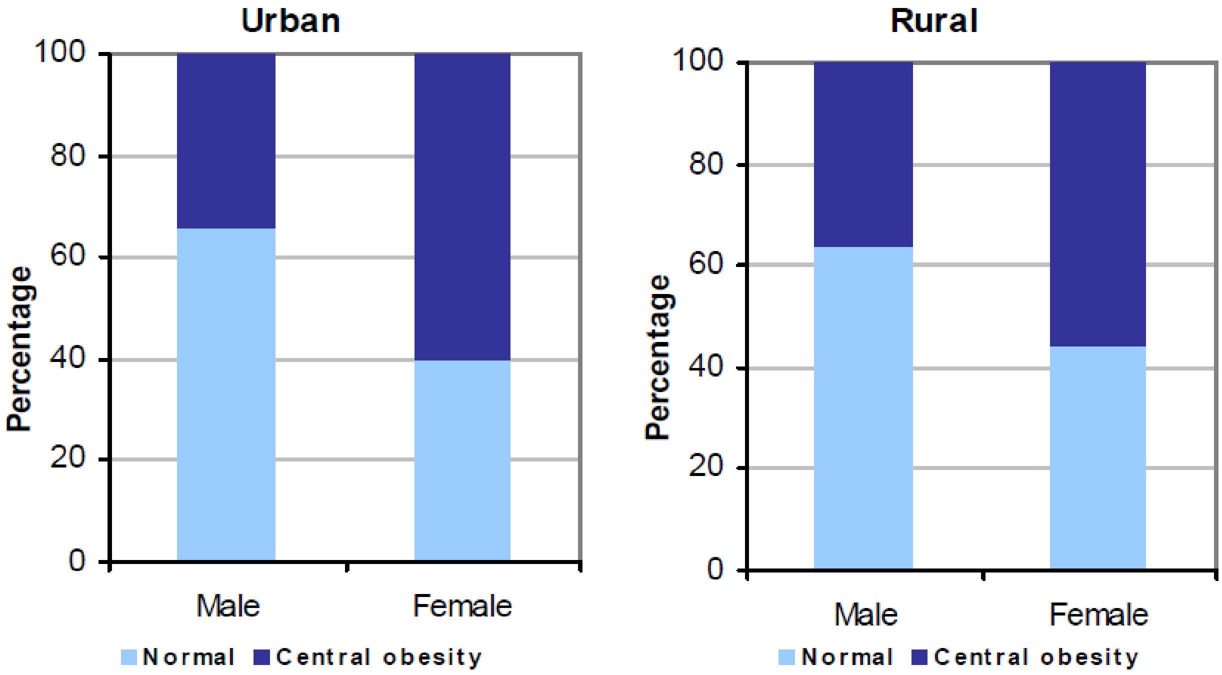
Definition: A fasting serum cholesterol level of greater than or equal to 200 mg/dl among males and females over 40 years of age.

	Male	Female	Total
Cholesterol ≥ 200mg/dl	31.3	38.1	34.7

Table 4: Cholesterol ratio in males and females

Prevalence Of Percentage of Population with Central Obesity

Definition: Waist circumference of greater than 80cm for females and greater than 90cm for males.



Graph 4. NCD- Percentage of adult population with central obesity

	Urban			Rural		
	Male	Female	Total	Male	Female	Total
Normal (%)	65.8	39.8	50.1	64.3	44.5	53.2
Central obesity (%)	34.2	60.2	49.9	35.7	55.5	46.8

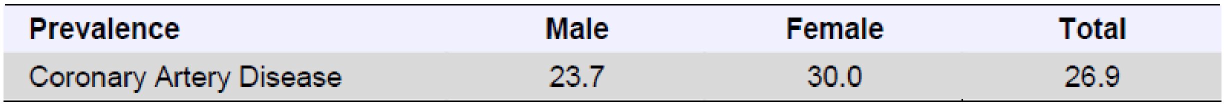
Table 5. Percentage of adult population with central obesity

1.5.4 Hypercholesterolemia

According to National Health Survey of Pakistan (1990-94) high cholesterol is defined as blood cholesterol of minimum 200 mg/dl, on the basis of this definition, it reported high cholesterol in more than 20% of the population over 15 years of age. This data is representing high cholesterol on national levels and is not being used for non fasting blood samples. Here cholesterol level is reported from a population based cross sectional survey on randomly selected persons over 40 years. Fasting serum cholesterol was used greater than 200 mg/dl as the definition of hypercholesterolemia for this survey and based on this, overall, 35% of the population was considered as hypercholesterolemia. Due to resource constraints biochemical analysis was not part of the NAP-NCD First Round of Surveillance, therefore, a mechanism for periodic reporting should be there on population sample for national representation. The cholesterol profiles were included in the Second National Health Survey of Pakistan, which recently is in planning phase, seems to be a viable option.

Prevalence Of Coronary Artery Disease

Definition: Coronary Artery Disease is defined as the composite outcome of abnormalities indicative of definite or probable CAD based on the Minnesota classification of ECG or past history of heart attack.

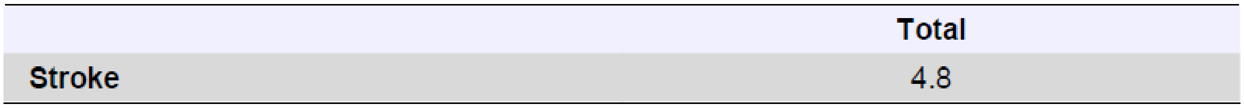


Prevalence	Male	Female	Total
Coronary Artery Disease	23.7	30.0	26.9

Table 6. NCD- Percentage of the population with Coronary Artery Disease

Prevalence Of Stroke

Definition: Stroke is defined as an affirmative answer to the following 'have you ever had a Stroke or Stroke-like illness in which part of your body was paralyzed for more than 24 hours?' Respondents were explained that paralysis refers to sudden weakness or numbness in any part of the body.



	Total
Stroke	4.8

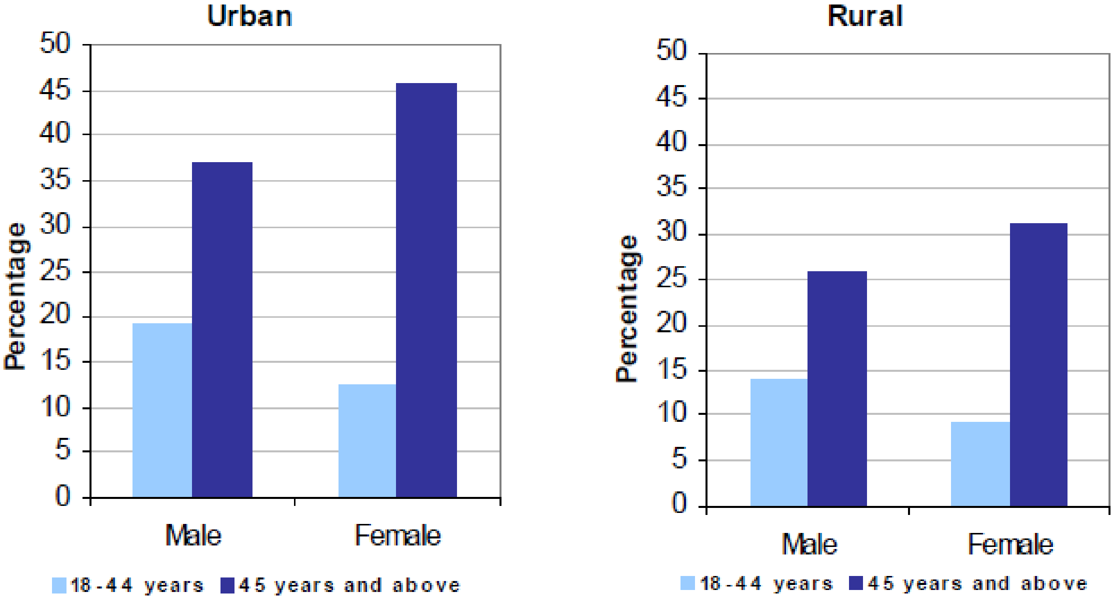
Table 7. NCD- Percentage of the population with Stroke

1.5.5 Hypertension

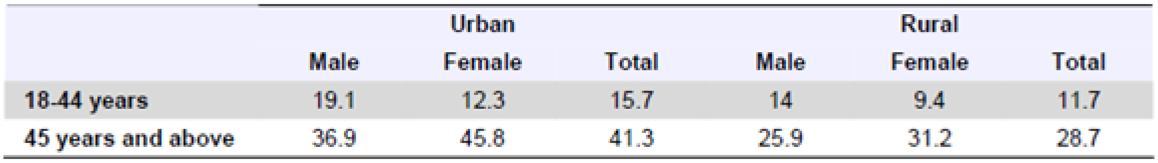
According to National Health Survey (1990- 1994) 18% of the population having age more than 15 years and 34% having age more than 45 years were labeled as hypertensive. Similar tendency was noted through another survey conducted in the Northern area of Pakistan[9]. According to recent NAP-NCD First Round of Surveillance high blood pressure is reported according to the new JNC 7 criteria. Even using the strict new criteria for blood pressure , 25% of the population over 18 years have high blood pressure in Rawalpindi district. High prevalence of high blood pressure is an important indicator of the burden of CVD in a population and there is consistent need to repeated consideration for current resource allocation behavior in public health in Pakistan[8].

1.5.6 Prevalence Of High Blood Pressure

Definition: Blood pressure greater than 140 mmHg systolic and/or 90 mmHg diastolic and/or taking anti hypertensive medication



Graph 5: NCD-Percentage of Population with High Blood Pressure



	Urban			Rural		
	Male	Female	Total	Male	Female	Total
18-44 years	19.1	12.3	15.7	14	9.4	11.7
45 years and above	36.9	45.8	41.3	25.9	31.2	28.7

Table 8: NCD-Percentage of Population with High Blood Pressure

1.6 Thesis Organization

In this section an overview of thesis is provided. This for the better understanding of the reader to so that the reader knows in advance what the chapter aims to conclude. Chapter 2 of thesis deals with the literature reviewed for this research. The research is one of its kind for the country like Pakistan and it is not possible to complete this research without looking into similar researches and studies conducted in the developed countries. Not only this but for efficient design and development of the system right tools must be selected and this can only be done by looking into similar researches and studies. Chapter 3 deals with the methodology of the complete structure of the system in the light of needs and limitations of the research. The chapter first looks into the current healthcare system of Pakistan and its shortcomings and then in the light of all these findings a solution to overcome these shortcomings is proposed. And finally complete design of the proposed solution is discussed. In chapter 4 implementation and results of the structure is elaborated. Here the design proposed in the previous chapter is developed and tools used for its development are examined. The structure is dealt in detail and its pros and cons and provided finally the complete designed structure is given by the end of the chapter.

Chapter 2

2. Literature Review

The chapter first discusses Framingham Heart Studies which is one of the main part of the research as it is used to identify common risk factors that participate in CVD and are accumulated for three generations over large population. Later Minnesota ECG code is discussed as this code serves as back bone of population survey .Finally the chapter deals with the data mining aspects and survey form developed for research as well as motivation for conducting research for the betterment of country.

2.1 Framingham Heart Studies

In 1948 CVD was identified as one of the main causes of death in America, the university of Boston and National Heart, Lung and Blood Institute started a revolutionary research in health research called the Framingham Heart Studies (FHS). The objective of the Framingham Heart Study was to identify the common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed symptoms of CVD or suffered a heart attack or stroke[10].

Hence FHS serves as a starting point for the development of a survey form that includes all the features needed to conduct a population study. FHS were carried out over a very long period of time and included three generation of participants. Though the methodology of the proposed work may differ from that of the FHS but the objectives of both are similar ; to identify the factors contributing towards CVD. And hence the types of survey used by FHS provide a good starting point for developing a survey form to conduct such studies in Pakistan but using rather latest techniques of data mining rather than observing three generations of patients with CVD history**.**

2.1.1 History Of Framingham

Cardiovascular disease (CVD) is one of the main cause of mortality and illness in United States. In 1948, the Framingham Heart Study in the supervision of National Heart Institute (known as National Heart, Lung, and Blood Institute NHLBI) contributed an ambitious embark in the field of health research. At that time people were unaware of general factors causing the heart diseases and strokes but CVDs were rapidly increasing the death rates and became American epidemic since the beginning of the century. The aim behind Framingham Heart Study was identification of common factors that lead towards CVDs by developing it over long period of time in large amount of people even though at first stage they don’t develop symptoms of disease or suffer from cardiac arrest. The researchers analyzed the life style and physical examination of 5209 men and women between the ages of 35 and 70 from Framingham town and conducted lifestyle interviews for analyzing common factors in the development of CVDs. By the detailed physical examination and medical background of patient some facts were recognized and in 1971 second generation of the participants were enrolled for similar analysis and examination. During 2002 they entered third phase for the enrollment of third generation of participants and more facts were gathered.

By careful studying the Framingham study of population researchers identified major risks for CVDs that were high blood cholesterol level, high blood pressure, smoking, diabetes, obesity, and being physically inactive and also a valuable information about HDL cholesterol levels, age, psychological issues and gender were recognized[11]. The concept about CVD risk parameters has become a part of the modern medical curriculum helped in the effective treatment and preventive measures. The important scientific contributions were made by Framingham study as well as enhancement in its research capabilities and inherent resources. New examination technologies of echocardiography , carotid artery ultrasound, MRI of brain and heart, have been integrated into past and ongoing protocols. Framingham researchers also collaborated with renowned researchers around the country and throughout the world about parameters of common diseases.

Need For Such Study

During 1930s to 1950s mankind overcome the infectious disease by controlling the risk factors. The infectious diseases were replaced by a mounting epidemic of cardiovascular disease (CVD) in the 1940s and 1950s. During World War II over the alarming rise of CVD became epidemic. Something had to be done. The secrets of the causes of CVD were not being revealed through simple laboratory and clinical research, so a continuous quest for techniques to treat and reverse the process was in order. Some suggested for primary preventative approach more reliable than search for cures [12]. Some of these prevention minded people occupied influential positions and translated their beliefs into actions. However, it was deemed that preventive approach delayed the disease[13]. The challenge was development of this preventive approach. The Framingham Study was given the challenge to recognize these modifiable parameters of patient and environment.

Research Design Aims And Hypotheses

The study was focused on and hypertensive CVD. Arteriosclerotic occurs by thickening and loss of elasticity of the walls of the arteries. It and hypertension were the most important of the CVDs and the least was known about them. A research plan began to emerge which consisted of the following. A randomly selected group of subjects in the age where these forms of CVD were known to develop would be selected. Based on a complete examination those subject free f definite signs of the disease would then be selected for reexamination at periodic internals and observation over a period of years. This would continue over a period of years until a sizable number were found to have acquired the disease. At that time a search would be made to identify the factors which influenced the development of the disease [14]. In the vocabulary of the 1980s the Framingham Heart Study as designed would be called a longitudinal cohort study. In the 1940s this terminology did not exist[15]. The attempt to label the Framingham Study and those studies that copied features of its design were important factors in generating it. The study as developed had one main aim to secure epidemiological data on arteriosclerotic and hypertensive CVD and two subsidiary aims . One to secure data on the prevalence of all forms of CVD in a representative population sample and second to test the efficiency of various diagnostic procedures [16].

These factors were generated by the Framingham investigators in consultation with an advisory committee composed of specialists representing several branches of medical sciences. The medical history and physical examination that would be obtained on the study subjects would generate data to test these hypotheses. The major hypotheses generated gave the following information that CVD increases with age, hypertension, high cholesterol level, use of tobacco and alcohol physical inactivity, high hemoglobin, body weight and diabetes. It occurs more frequently in males[17]. These hypotheses were directly used to determine the medical history obtained and the physical examination taken during the repeated tests of the study. There was continuing growth and modification in research hypotheses with the evolvement of medical science .

Uses Of The Framingham Heart Study

Since its inception statisticians have played a major role in the study. The major areas of activity at the beginning were development of the sampling plan, determination of the sample size and the length of the study, data processing and data analysis. The Framingham Study is well suited for the estimation of incidence rates. The major contribution of the study has come in detailing incidence rates, in particular relating risk factors and with is careful longitudinal follow-up it has near complete data on the development of diseases. The reviewed background of the Framingham Heart Study shows that it continued to yield valuable information for 40 years.

2.2 Development History of Minnesota ECG Code

During 1950s , studies of heart disease became systematic and tendency of keeping quantitative and standard records were used for comparison of disease rates. Diagnoses made from death certificates or by physicians of different training backgrounds can result in spurious differences[18].

Even special trained physicians can make highly variable independent diagnoses . The problem of comparing and assessing heart diseases was achieved by considering electrocardiogram (ECG). Because its results strongly represent heart disease patterns of great interest i.e., death of heart muscle (infarction), enlargement of heart size (hypertrophy), and irregularities of rhythm or conduction. The ECG looked efficient as it is a graphical record and standard procedure of collected data , measurements, and classification are amenable. It also looked ideal and acceptable because of its simplicity , painless and inexpensive record. This seemed very well until quick emergence of several records. There were large differences between physician’s impressionistic readings of ECGs even by the same cardiographer at different times because of the reason of non-existence of specific criteria for specific cardiac findings and physicians just relied on recognizing patterns. First the problem of standard criteria measurement to be solved before assessing and comparing endpoint cardiac “events” and population rates. Hence a need was greatly felt for a centralized ECG coding criteria.

Minnesota ECG Coding Center played the role of ECG visual reading service centre for clinical trials and epidemiological studies at nationally and internationally for more than 40 years. The objective of the centre is to record prognostic and diagnostic value of ECG for clinical trials and epidemiological studies. The Center has been at the top level in the process of development and evaluation of new criteria sets for incident cardiac events, including silent or symptomatic myocardial infarction, left ventricular hypertrophy, ischemia, heart rate variability , and QT dispersion [19].

The Minnesota Code classifies electrocardiogram for utilizing predefined measurement rules for assignment of specific codes pertinent to severity of ECG patterns Dr. Henry Blackburn developed the code for evaluation of ECG results in uniform, clear and meaningful terms for classifying ECG in the domain of epidemiologic studies and clinical trials. Criteria for ECG classification is incorporated and validated by clinicians. Through Minnesota Code meaningful ECG classification is done that is free of impressionist physician bias, and provide different studies with common standard for comparing or pool ECG predictions. It has now find out objective techniques for defining important parameter changes from serial ECG comparisons. Minnesota Coding techniques have been incorporated into a computer program for ECG comparisons and used by the researchers for devising new procedures.

Training and evaluation of technical staff for new studies is also provided by Minnesota ECG Coding Center and refresher training is also provided for further ongoing research studies. Researchers who want to use the Minnesota Code for coding their own ECGs can benefit by using the Minnesota Code interactive computer tutorial or can submit samples of their coded ECGs for evaluation. The complete code is provided in appendix A. The code plays a major role for this research work . As discussed Minnesota code don't have any standard in any hospital practice. Instead, it is a designed system for rigorous population researches and clinical trials. Hence it serves as the fundamental building block upon which the population survey in this research is conducted. The complete understanding of this code is beyond the scope of this research work . Here it is interesting to know that the research aims to automate the process of ECG coding and utilized the data generated for caring out data mining and trend analysis. Minnesota code gives the rules of ECG coding that need to be automated.

2.3 Data Mining

Data Mining is the process through which hidden patterns are identified among different fields in large amount of data sets[20]. Data mining plays an important role in this research to extract different patterns and relationships among demographics and arrhythmias. The patterns, once extracted, will allow us to develop a strategic approach to solve major medical problems.

Data, Information, And Knowledge

Database generated consists of different fields, ranging from demographics to ECG data. The raw data is the lower layer of the database. The overall database is two layered. The lower layer consists of the raw data. The second layer gives us field, the correlation of which with respect to other fields is to be developed. The patterns, associations, or relationships among demographics and ECG fields provide us information [21].

2.4 The Survey Form

The research aims to find the patterns that might exist between different CVD socioeconomic, demographic, medical and personal history of a patient. Hence the form that also serves as the basic design of the database must cover all the aforementioned aspect of a patient's life style. This not only make the form difficult to design but also makes it quite difficult to be handled. Framingham heart studies served as the guidance used for making of this survey form. Framingham studies are very detailed studies and certain aspect of the study are beyond the scope of this research. The studies are also designed keeping the American culture in mind. But for this research a survey form was to be designed keeping the living style of the people in this country. Hence Framingham though proved to be a good literature to start from, could not help in much in the compilation of form and hence local cardiologist were consulted for final development of the survey form.

Chapter 3

3. Methodology

3.1 Introduction

This chapter deals with the design of the complete structure of the system in the light of needs and limitations of the research. Though the main aim of the thesis is to find trends that exist between different CVD and patients but it cannot be achieved without the successful achievement of the secondary goal of centralizing the Healthcare Infrastructure of the country. Using the data gathered by centralizing the health care system to help fight CVD is the main focus. The chapter first looks into the new healthcare system of Pakistan proposed and implemented by [21] and then acquisition of ECG data that is used for the research and development activities and identifying the hidden patterns in different regions of Pakistan for across genders and socioeconomic groups and between the type of facilities where applicable, particularly in the case of mortality and morbidity data.

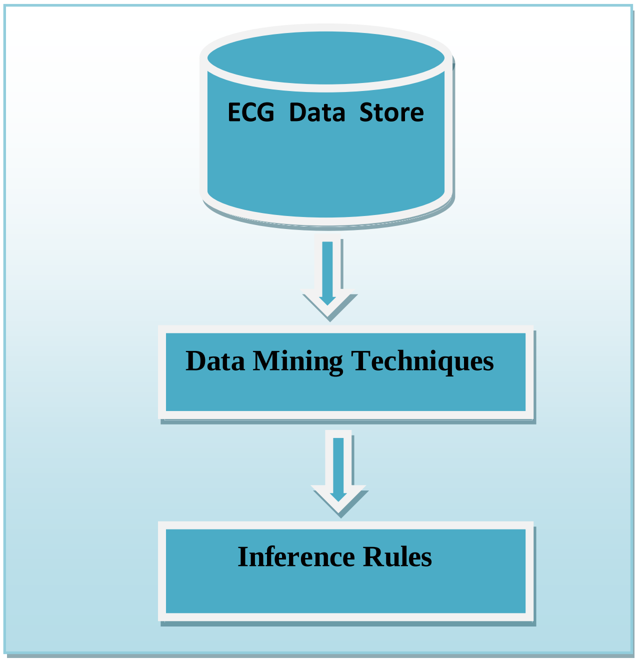


Figure 3 : Research Overview

3.2 A Prerequisite Step For Proposed Research

Though the main aim of the thesis is to find trends that exist between different CVD and patients but it cannot be achieved without the successful achievement of the secondary goal of centralizing the Healthcare Infrastructure of the country. Using the data gathered by centralizing the health care system to help fight CVD is the main focus. As described previously, specialist doctors cannot be assigned to rural areas due to low patient density. But we know that there are health centers in rural areas which as a start can serve as small hospitals. A mean of centralizing these health units need to be devised. The centralization of health network in the country would not only prove beneficial for patients but would also make the governance of these units much easier. This centralization will help remove the inequality that exist today in the distribution of health facilities between population of rural and urban areas. Figure 3.3 gives the proposed centralized Healthcare infrastructure. Cardiologist and other specialist doctors cannot be assigned to rural areas on regular bases. But we know that there are adequate number of health units in the rural areas to start with. Hence by providing diagnostic machines such as ECG and few technicians having operational knowledge of computers, to operate these machines and send data over internet to doctors and specialist in urban areas. We can fill the gap between the doctors and the patient in the rural areas. Such technicians are easy to train. And this kind of system would also open a new set of employment opportunities for the knowledge youth of rural areas.

Now let us take a look how this system proposed in figure 3.1 works. Every hospital in the country will have a Cardiac Control Centre (CCC) is connected to the central repository which hold the record of every patient that had been to any hospital in the country. CCC in turn is connected all small health units in the district. Thus the hospital has remote branches connected to hospital via high speed dedicated connection. When a patient comes to one of the health unit in the rural area the technician in the area performs the required tests and send it over high speed connection to the CCC of the hospital in that district. CCC enters the patient record in the central repository and also the specialist doctor in our case a cardiologist looks at the tests and makes a recommendation whether patient need to be moved to a hospital or nay medicine or test that might be needed.

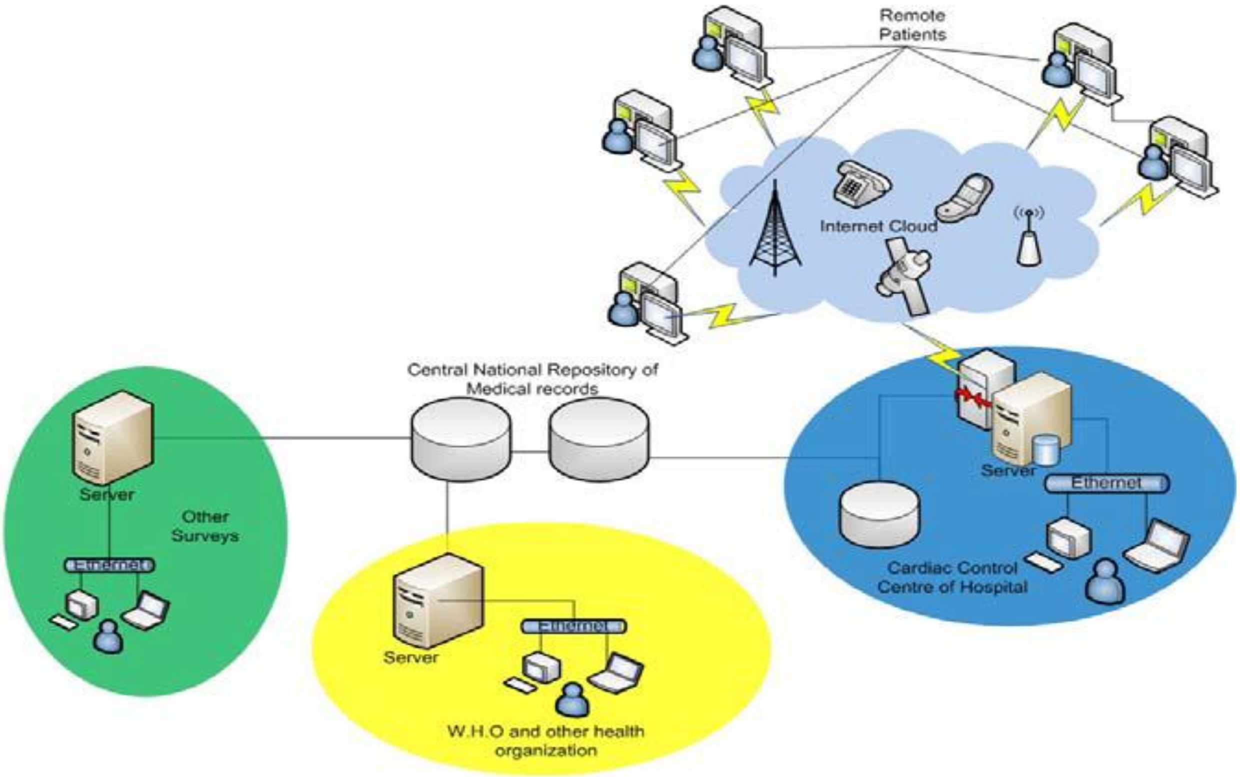

Remote Patients
Internet Cloud
Central National Repository of Medical records
Server
Other Surveys
Ethernet
Server
Ethernet
W.H.O and other health organization
Server
Ethernet
Cardiac Control Centre of Hospital

Fig 4: Centralized Health Care Architecture

Cardiac Control Center (CCC) would be located in the central hospital. Network diagram of the CCC is shown in Figure 3.2

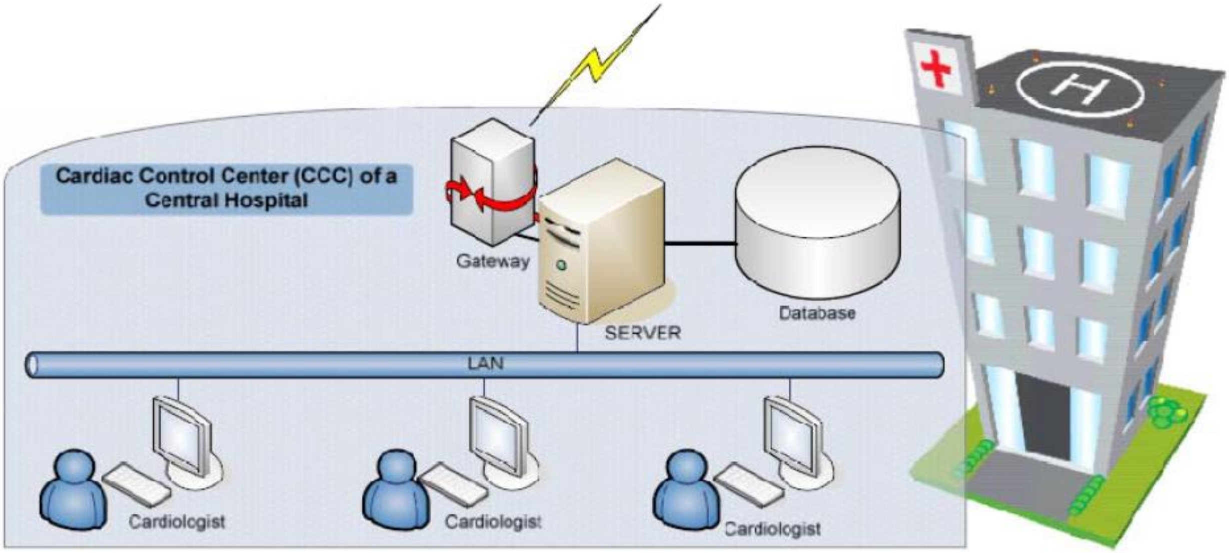


Figure: 4   Cardiac Control Centre located in District Hospital

First step is to acquire ECG data, patients' records and other physiological information from remote clients. For this purpose gateway would be developed at the Server which would receive ECG data from different networking technologies and store it locally. Figure 3.3 depicts the general architecture of the acquisition stage. Once the data is received, it is first organized into the database. Data organization would be carried out similar to a patient management system (PMS). It would include information about patients' history, demographic profile and priority level set by the software at the client end. ECG's detected with serious arrhythmia condition would automatically be assigned a high priority level and would be at the top of automatically generated ECG work-list.

Next, intelligent physician assignment would be undertaken. This depends upon following variables:

- Availability of the doctor
- Specialization of the doctor to deal with the case
- Call for second opinion
- Criticality of patients' health

Intelligent physician assignment would ensure that ECGs of patients are not left undiagnosed. If a doctor is not present, it will automatically allot another physician for diagnostics. It will also ensure that these ECGs are with the correct medical specialist and if any doctor has requested a second opinion on the case, it will automatically update its work list accordingly. Once a physician has submitted his/her final recommendation/diagnosis, server would transmit these recommendations to the remote patients in real-time. In the meanwhile, it will also store complete patient information and diagnostics in the local repository. ECG work list would be updated. This way the patient is made available to a specialist doctor without having moved to city. But this is only possible if a connection lies between the health unit and the hospital, as proposed in figure 3.1. Once a patient enters a hospital or any of its connected health units, he/she is allotted a unique ID which is his or her CNIC number. The data of the patient is stored by the CCC of the district hospital in the repository with CNIC as reference. Next time when the patient visits any hospital throughout the country the hospital the just needs to enter his/her CNIC to access his or her record.

The centralization will also provide with a better consultation facilities. Doctors from different hospitals would be able to share information about their patient very easily and hence consult each other easily. Thus overall all improving the healthcare facility in the country. The inequality of healthcare facility distribution would be reduced greatly and will overall improve the living standards in the country. The architecture in figure 3.1 proposes use of a dedicated high speed sanction between CCC and the repository. This is while taking into consideration that the patients records need to be kept secured against all kinds of threats and thus a dedicated connection can prove use full against hackers and other possible cyber criminals.

Similarly the centralization provides a way of better conducting health surveys in the country. These survey conductors would be made available data of patients from all over the country and thus would be able to better estimate the healthcare situation of the country. This would also prove beneficial from governance point of view. As the district hospital would be able to maintain the record of all the patients visited the hospital as well as the healthcare units in the district. Also district hospital would have records of all the health units and people employed in those units.

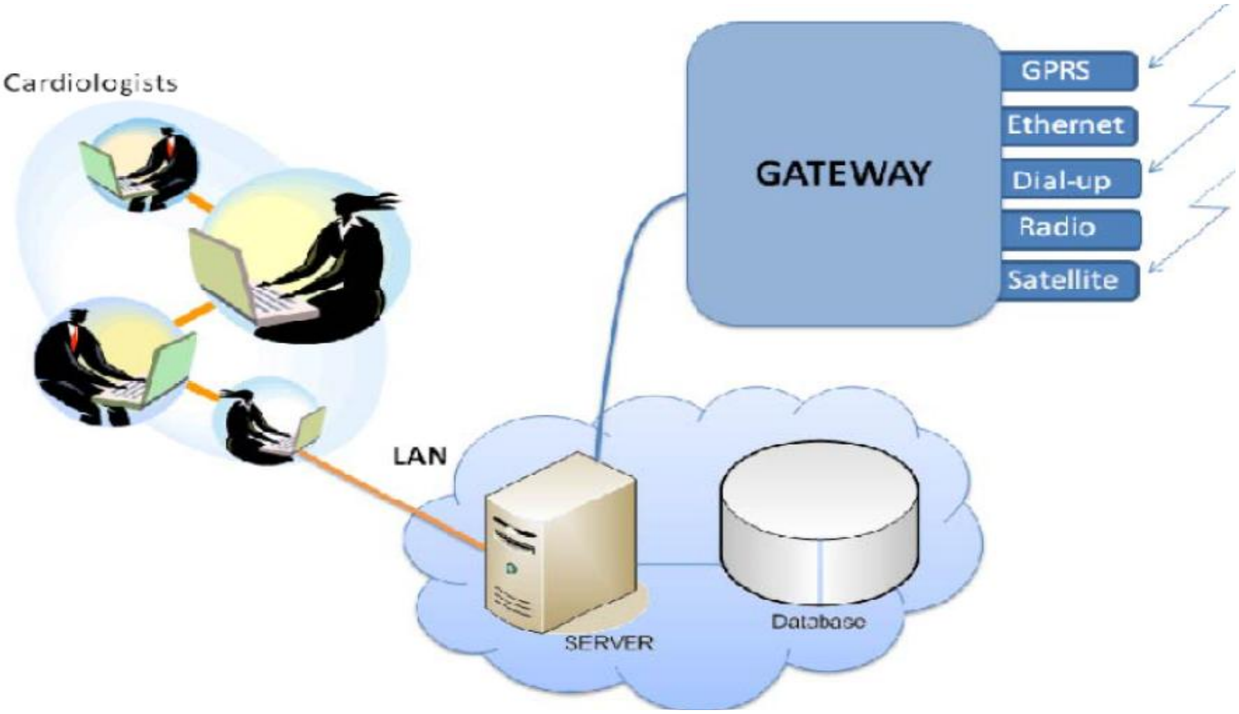


Figure 5 : Gateway Server

The introduction of IT in the healthcare in the proposed way will help not only improve the healthcare facilities but also open up new employment avenues for the youth of the rural areas and hence improve the overall living standard of the country. Hence it's about time we consider renewing the health infrastructure and take this industry in to new era. This centralization will also stimulate the health research in the country and hence prove beneficial for student and researchers of medicine.

3.2.1 Importance of Implementing Proposed Infrastructure Towards Achieving Goals of The Research

The research aims to use the modern techniques of data mining to find the trends that might exist between different CVD and other aspects of patient's life. As discussed in section 2.4, in order to achieve accurate results from a data mining algorithm we need to have large data sets. Thus one of the limitations to the research is the availability of large data sets. This can only be achieved by implementing the system proposed in figure 3.1. This system will centralize the whole medical industry in the country and hence all the medical records. Thus making it possible for us to accurately implement the data mining algorithms.

Though the main aim of the research is to find these trends but it cannot be achieved without the successful achievement of the secondary goal of centralizing the Healthcare Infrastructure of the country. Using the data gathered by centralizing the health care system to help fight CVD is the main focus. But how will that be achieved? As it can be seen the gateway at the Hospital server receives physiological information from remote locations and generates an ECG work-list and makes automatic doctor assignments. Network Area Storage (NAS) keeps record of patient's data for future reference and data mining. This is the actual brains of the system. The data mining will be used to carry out trend analysis to find out hidden patterns that might exist between the physiological, demographical and socioeconomic information of a patient. One of the main part and the core element in the system in the ECG and the data retrieved from it. This is the core element in identifying the nature of CVD that are prevalent in certain population. Hence the system must be able to store and communicate the ECG data.

3.3 ECG Storage

Previously in Pakistan ECG data of a patient was disposed off after the checkup and the patient on his next checkup had no history. In the proposed system ECG data will be temporarily stored in the database on the system which would then be used to populate the Minnesota entries in the central repository against patients CNIC . With the help of the data base the cardiologist will be able to retrieve the patient's history on his second or further checkups and the data will also be available to expert cardiologists around the country. The research aims to develop an ECG database which will include ECG data of patients as defined by the Minnesota ECG code. ECG database would greatly help in the evaluation and testing of the performance of new ECG analysis and data compression algorithms and hence play a key in the biomedical research. In Pakistan, there is a huge demand for such a database as the HRV patterns of the people in this region are very unique and different. Such a database would help researchers and developers to design systems and analysis techniques for cardiac patients particularly for Pakistan. ECG database Management System (EMS) which will perform the following tasks.

- Store and Manage ECG data
- Automate the ECG workflow with in a hospital
- Facilitate the ECG editing and confirmation process
- ECG Auditing and Tracking

The system's design is based on industry standard three-tier architecture and divided into three major components including the underlying layer of ECG database, the upper layer applications for ECG viewing, editing, data retrieval and data management, and the middle layer to adapt a flexibly configurable workflow according to the medical staff structure of a hospital.

3.4 Data Mining Techniques

This part of the research will be used to find these trends using the data gathered in the central repository. This is the main aim of our research, what we wish to accomplish and implement through it. Having accumulated all the medical records into a central repository, it is now viable to implement data mining algorithms onto the collected information to find out the required trends and discover the correlation between various factors which previously were not possible to discover.

Data mining algorithm onto the database will discover if confined spaces contributes in any way to the possibility of getting an arrhythmia. This would be done by comparing the data of people with the arrhythmia under consideration and the people living in highly populated areas. If the area of overlap between these two searches is significantly high, this would indicate that confined spaces do indeed cause the person to be more vulnerable to a certain kind of arrhythmia. Also If the area of overlap between the two searches is significantly less or negligible, then this would show that confined spaces actually improves health and decreases the vulnerability of people to the arrhythmia under consideration. However if the area of overlap between these two fields is just normal such that it appears to be random in nature, then no particular relation exists between the two variables.

3.4.1 K - Mean Clustering

K- mean clustering is a process of cluster analysis that divides n particles into k clusters in which each particle belongs to cluster having nearest mean[22].

$$arg_S \, min \sum_{i=1}^{k} \sum_{Xj \in S_i} \|x_j - \mu_i\|^2$$

Here (x1,x2,……xn) is a set of particles and *k*-means clustering divides the *n* particles into *k* sets ($k \leq n$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the sum of squares within-cluster. Where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

So k-means clustering , clusters the CVD in different regions . All the patients that belong to any one cluster will have some similarity of attributes. These attributes will then be extracted as these will be the attributes causing this abnormality. In this way the trends between CVD and different attributes will be extracted.

3.4.2 Pre-filtering

Data pre-filtering transforms the data into a meaningful format for processing of more easy and effective use. The objective of using pre-filtering for data acquisition and finding out the information about the most dominant factor makes the data more meaningful. The process of pre-filtering is similar to pre-processing step[23]. We will calculate mean and standard deviation across different attributes for CVD and will pick those attributes who are not frequently changing.

Mean

The mean is the arithmetic average of a set of attributes , or distribution. Here n represents the n number of attributes[24].

$$\bar{x} = \frac{1}{n} \cdot \sum_{n}^{1} x_i$$

So by taking the mean of values we will get those attributes who are most commonly found in the patients suffering from CVD.

Standard Deviation

Standard deviation denotes variation or dispersion existing from the average value or mean. The attributes having low standard deviation denotes that data points lie close to mean and high standard deviation denotes that points of data are spread out over large range of values.

$$S_N = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Figure 3.4 is showing a bell shaped curve and each have width of 1 standard deviation and also values in the range of dark blue portion are close to mean[25].

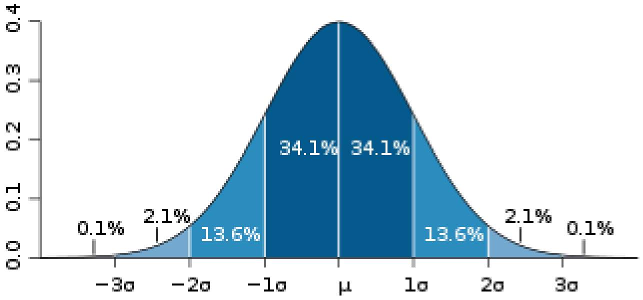


Figure 6 : A plot showing standard deviation

In my thesis I will choose a threshold and will apply threshold on the attributes of CVD and will pick those attributes of CVD whose standard deviation is less than threshold. In this way I will accurately find those attributes in specific region which are most common cause for CVD in specific region causing the dominant feature for morbidity and mortality in that region .

3.5 Inference Engine

Inferences are made on the basis of above data mining techniques. Inference rules are provided which are of great statistical significance based on the population data and provide comparisons between provinces and districts – geographically, between different areas of the country such as rural or urban, across socioeconomic groups and genders and between the type of facilities where available or not , particularly in the case of morbidity and mortality of data. However, paucity of data in many areas limits the ability for such comparisons.

Chapter 4

4. Implementation And Results

4.1   Acquisition Of ECG Data And Data Mining Techniques

ECG Data is accessed from the ECG data engine discussed in the above section 3.2 . ECG data shows the CVD diseases on the basis of diagnosis of the intelligent physician. Now in the database different features are also stored in the database about patients.

4.1.1 Clustering Results

ECG data are clustered by applying the k-means clustering in different regions according to some attributes of similarity. Here figure 4.1 shows the distribution of patients in different regions and represents a patient's map.

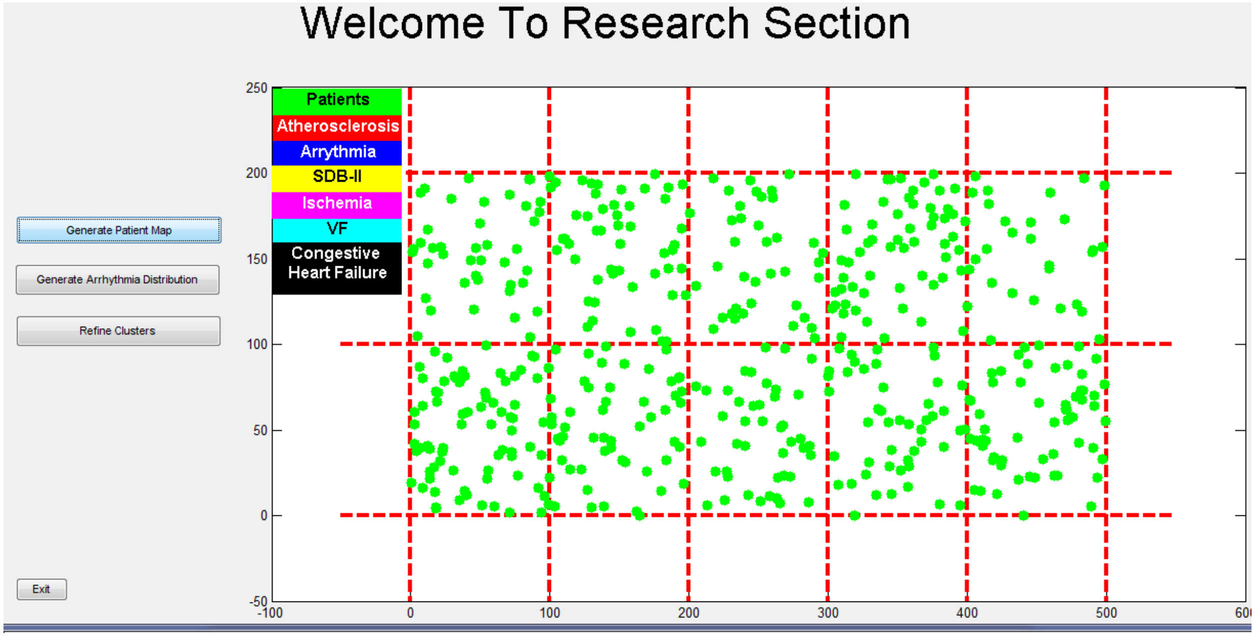


Figure 7:   Map of Patients distributed in different regions

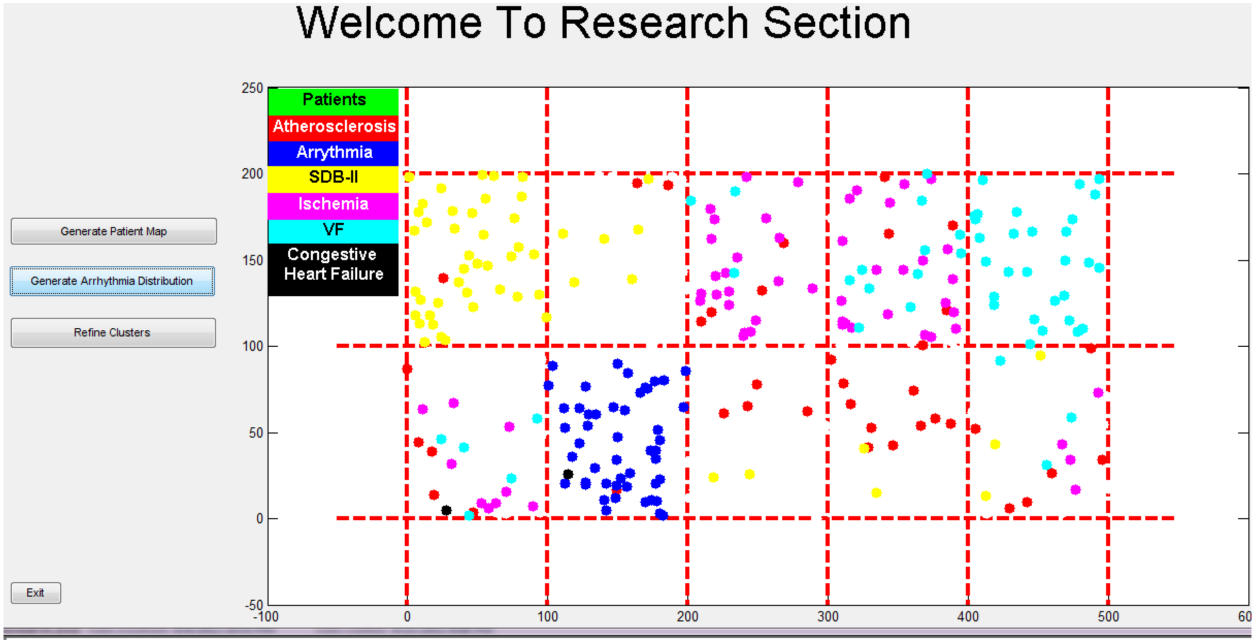


Figure 8: Distribution of patients having CVD in different regions

Figure 4.2 shows the distribution of patients having CVD in different regions. Different colors show different cardiovascular disease as shown in map. Ischemia disease is represented with pink color , SDB II with yellow, VF(ventricular fibrillation) with sky blue color , Congestive Heart Failure with black color, Atherosclerosis disease with red color. Now different regions show more than two or three diseases and some regions are dominant with some specific disease.

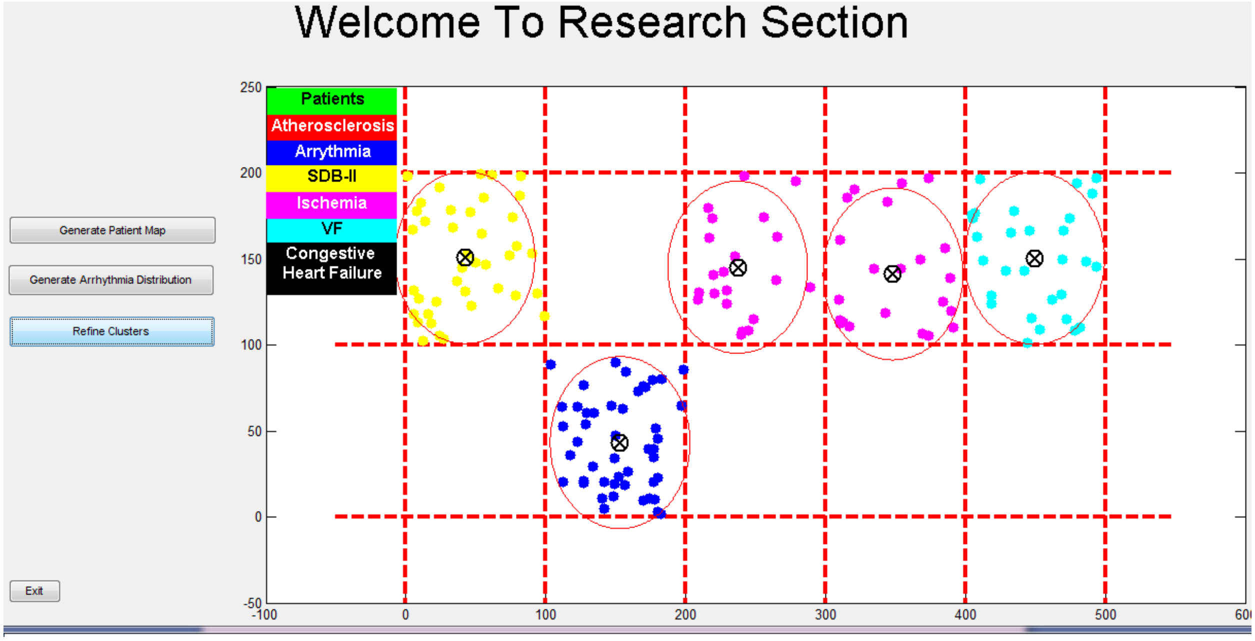


Figure 9: Refining the clusters of diseases in one region

Figure 4.3 shows the refinement of clusters and clusters with similarity of attributes are combined together with respect to some specific area. Disease common in region one is clustered together . Similarly there is formation of five different clusters showing different prominent diseases in different regions. Thus map shows the region specific diseases according to spatial coordinates.

4.2 Different Features of CVD

By studying the patients data and behavioral patterns of the patients there are some parameters which are directly related with CVD.

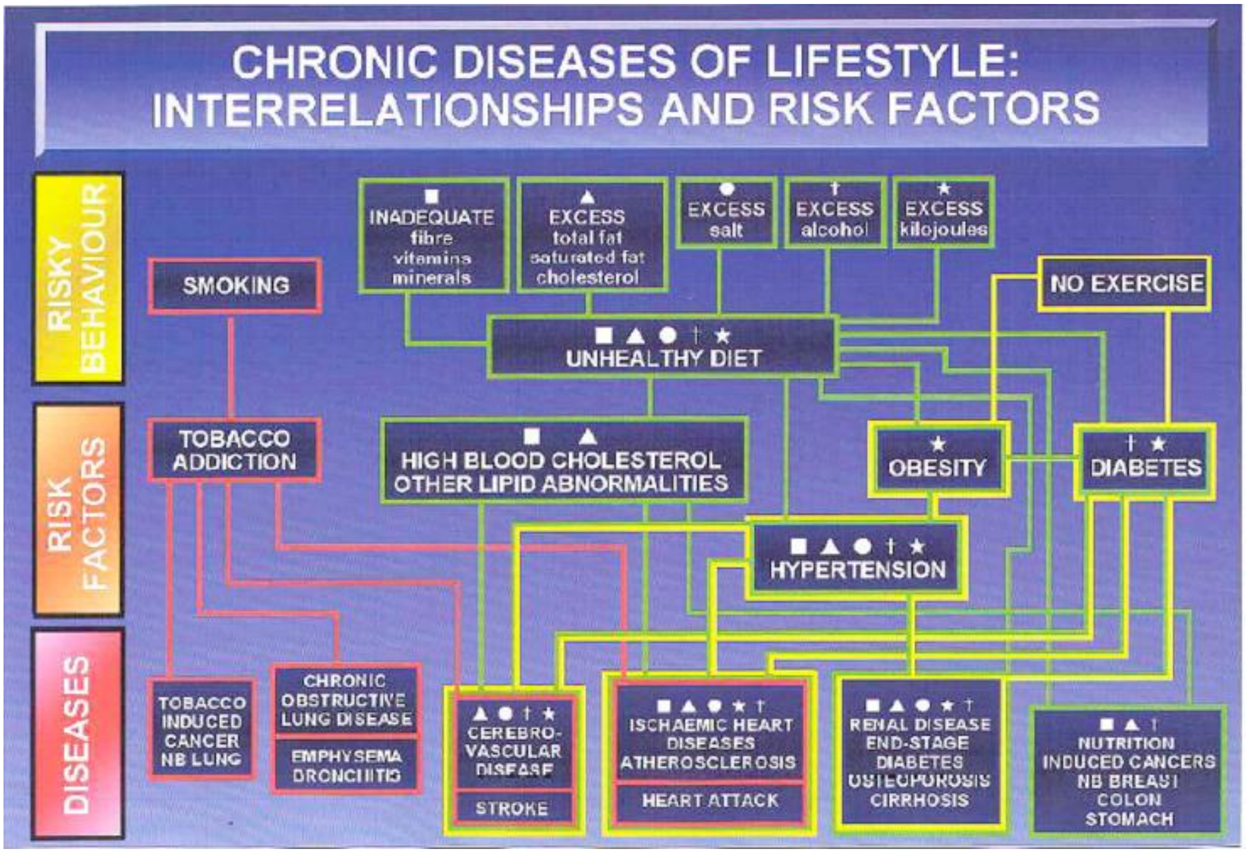

CHRONIC DISEASES OF LIFESTYLE:
INTERRELATIONSHIPS AND RISK FACTORS
RISKY BEHAVIOUR
RISK FACTORS
DISEASES
SMOKING
■ INADEQUATE fibre vitamins minerals
▲ EXCESS total fat saturated fat cholesterol
● EXCESS salt
† EXCESS alcohol
★ EXCESS kilojoules
NO EXERCISE
■ ▲ ● † ★ UNHEALTHY DIET
TOBACCO ADDICTION
■ ▲ HIGH BLOOD CHOLESTEROL OTHER LIPID ABNORMALITIES
★ OBESITY
† ★ DIABETES
■ ▲ ● † ★ HYPERTENSION
TOBACCO INDUCED CANCER NB LUNG
CHRONIC OBSTRUCTIVE LUNG DISEASE
EMPHYSEMA DRONCHITIS
▲ ● † ★ CEREBRO-VASCULAR DISEASE
STROKE
■ ▲ ● ★ † ISCHAEMIC HEART DISEASES ATHEROSCLEROSIS
HEART ATTACK
■ ▲ ● ★ † RENAL DISEASE END-STAGE DIABETES OSTEOPOROSIS CIRRHOSIS
■ ▲ † NUTRITION INDUCED CANCERS NB BREAST COLON STOMACH

Figure 10: Different risk factors related to CVD

These factors include

- High blood pressure
- High cholesterol level
- Use of Alcohol
- Smoking
- Drug Use

- Obesity
- Unhealthy Diet
- Age
- Other diseases such as Diabetes
- History of other heart diseases
- Hypertension
- Triggered activity due to some other disease
- Acute Infection
- Poor sleep
- Gender
- Medication Side effects
- Hereditary

These CVD occur due to above factors and are also influenced if some strategy is adopted to reduce the above factors. So the risk of these diseases can be reduced by the life style and behavioral changes[26]. There are socio-economic factors that are indirectly related to CVD[27] but their effect cannot be neglected and they effect equally as well as directly related factors. And these factors include

- Income
- Dependent Family Members
- Marital Status

4.3 Demographics Based On Health Issues

Pre-filtering is used for extracting the most prevalent factor in different areas. It is seen that different common factors are prevalent in different areas. We took sample data for six regions.

- Islamabad
- Rawalpindi
- Lahore
- Gujranwala
- Faisalabad

The result came with different common pairs. Some of the most common factors are as follows.

- Obesity
- High Blood Pressure
- High Cholesterol
- Smoking
- Hereditary
- Low Income
- Hypertension

4.4 Inference Rules

With the help of pre-filtering process we come across useful and relevant information which help us in making inferences and reiterating the process of providing facilities and resources in those areas where CVDs are more prevalent and increasing in the morbidity and mortality rate.

4.4.1 Analysis of CVD patterns in Islamabad Region

Islamabad is capital of Pakistan and developed with respect to other cities of Pakistan. People living in this city have good financial conditions and health care and pertinent medical facilities are easily accessible in this area[28]. According to the patients data following most dominant CVDs are found in Islamabad city.

1. Disease ***Atherosclerosis*** is more prevalent in Islamabad. The possible cause of the disease is obesity, High cholesterol level and High Blood Pressure.

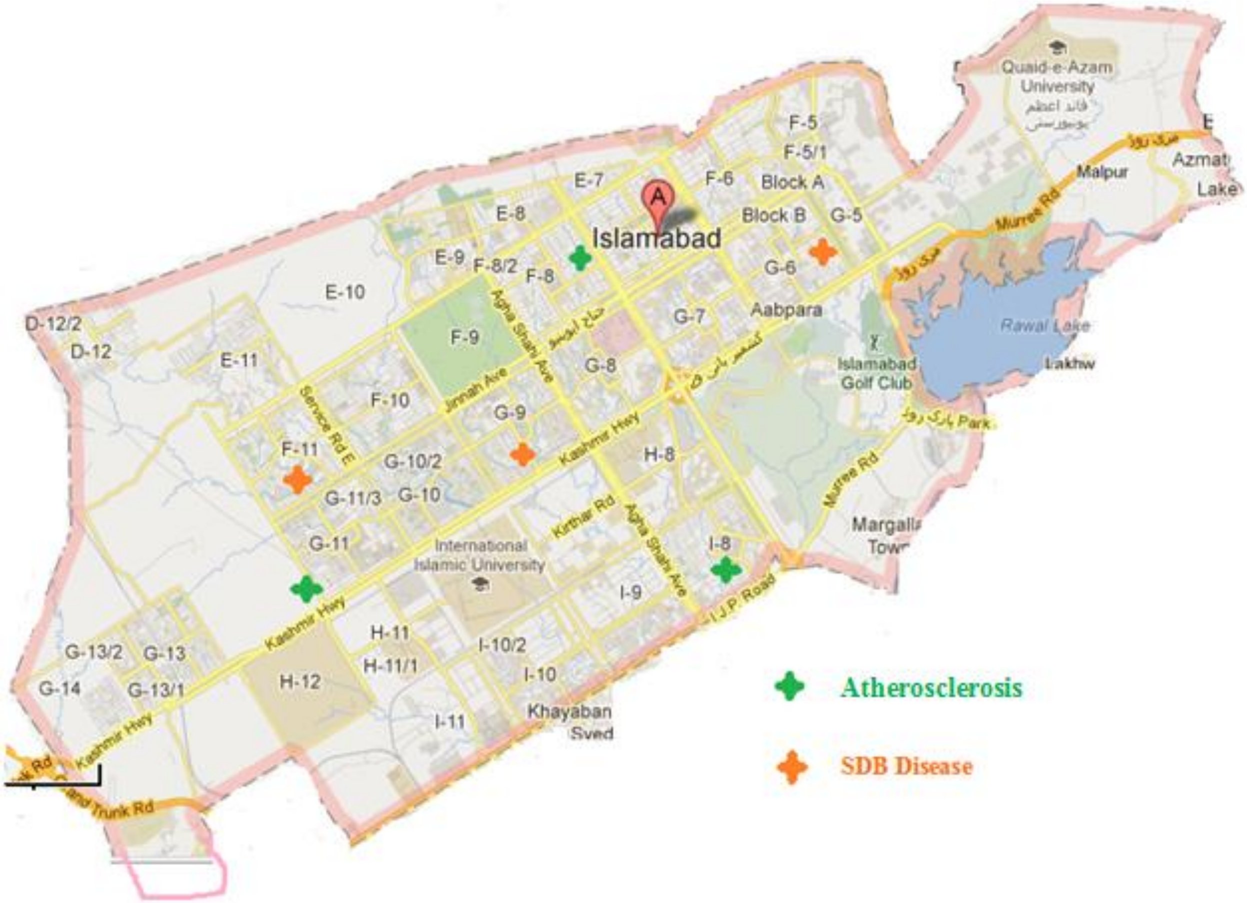


Figure 11: Islamabad Map showing prominent CVD diseases

2. SDB disease is another prevalent CVD disease in Islamabad. The possible cause of the disease is obesity, and age. Moreover disease is more common in males.

The causes of the both diseases shows that obesity is the major cause of the disease in this region. Behavioral and life style changes can help in reducing the disease in this region[29]. So people should use nutrient rich diet and low in fats. Moreover people should involve themselves in physical activity and exercise to reduce the disease.

4.4.2 Analysis of CVD patterns in Rawalpindi Region

Rawalpindi is a developed city of Pakistan, lies close to the capital and includes seven autonomous tehsils [30]. It is also developed city and there are abundant medical facilities and easy access to cardiac hospitals and general hospitals [31]. Population of Rawalpindi is divided in upper, middle and lower middle class. According to CVD patients profiles causes and factors for disease that came through

1. Ischemia disease was prominent disease in Rawalpindi region B. The possible cause of the disease was smoking, hypertension, and age in the range from 46 to 75 years. Moreover, people whose income was in the range 16000 to 20000.

2. Ischemia disease was prominent disease in Rawalpindi region A. The possible cause of the disease was High cholesterol level, Diabetes , and age in the range from 65 to 75 years. Moreover, people whose income was in the range 80000 to 100000.

3. Arrhythmia disease was prominent disease in Rawalpindi region A. The possible cause of the disease was history of heart disease, obesity and High Blood pressure. Moreover people whose income was in the range of 80000 to 100000.

4. SDB disease was prominent disease in Rawalpindi region B. The possible cause of the disease was history of CVD hypertension, poor sleep, age range of 45 to 75 years , having 4 to 5 dependent family members. Moreover people whose income was in the range of 16000 to 20000.

5. *Atherosclerosis* disease was prominent disease in Rawalpindi region C. The possible cause of the disease was medication , high blood pressure, hypertension, Triggered activity of some other disease. Moreover people whose income was in the range of 10000 to 14000 and dependent family members were in the range of 5 to 6.

6. Ventricular fibrillation disease was prominent disease in Rawalpindi region B. The possible cause of the disease was history of CVD hypertension, poor sleep, age range of 45 to 75 years , having 4 to 5 dependent family members. Moreover people whose income was in the range of 16000 to 20000.

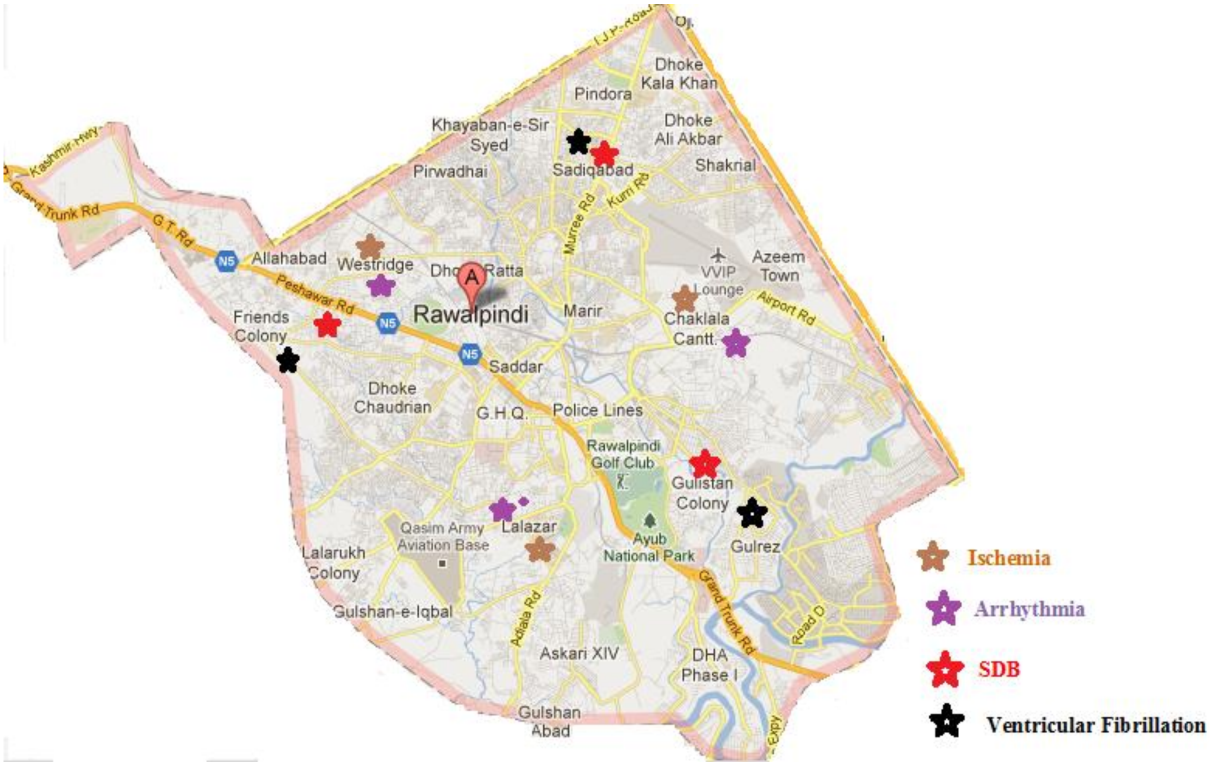


Figure 12 : Map of Rawalpindi showing CVD affected areas

So by analyzing the above factors many hidden factors are concluded. Same disease in different areas had different cause. The causes of *ischemia* and *arrhythmia* having good financial conditions (region A) were due to obesity , high blood pressure and high cholesterol level. This indicates the behavioral patterns of people who caused the disease. The suggestion is that these people should reduce the disease through primary prevention from risk. Healthy diet rich in nutrient and low in fat should be used and should indulge themselves in physical activity and exercise etc.

By analyzing region B people suffered from Ischemia and *SDB* due to hypertension. The major problem was low income and more dependent family members due to people suffered from poor sleep and hypertension causing SDB disease and also people suffered from Ischemia and main reason of ischemia in this region was hypertension due to low income. Moreover through survey and questionnaire from the people living in these area it also come to know that people living in this area also had water issue. Due to lack of availability of pure drinking water people suffered from other diseases such as liver and lung and medication [32]. So medication side effects triggered such diseases into CVD diseases. So decisions should be taken by the government to improve the health conditions by improving the important underlying cause such as water and such issue should be solved.

By analyzing region C important information of CVD disease was collected according to morbidity and mortality data as death of most of the people due to CVD occurred in this region. The dominant disease in this area was *Atherosclerosis*. Many hidden factors are revealed as main reason in this area was smoking and drug abuse and low income that was not sufficient to meet daily needs of life . People were hypertensive and became easy victims of narcotics and drug abuse. A personal survey through questionnaire also revealed that in this area there were some criminals who were selling drugs and narcotics. So this region needs special attention of authorities to save people from such abuse. Also provision of free medicines and medical campaigns can save people of this area. Financial assistance to unemployed class and effective use of the manpower by provision of work can play a significant role in minimizing this issue.

4.4.3 Analysis of CVD patterns in Lahore Region

Lahore is the second largest city of Pakistan and densely populated. The main city Lahore is developed and have hospitals and medical facilities while rural and sub-urban areas are not developed and even lacks in hospitals and basic medical facilities [33]. People from rural areas of Lahore have to come to the main city even for basic medical facilities and examination. By studying ECG data and patients profile various patterns of CVDs were found.

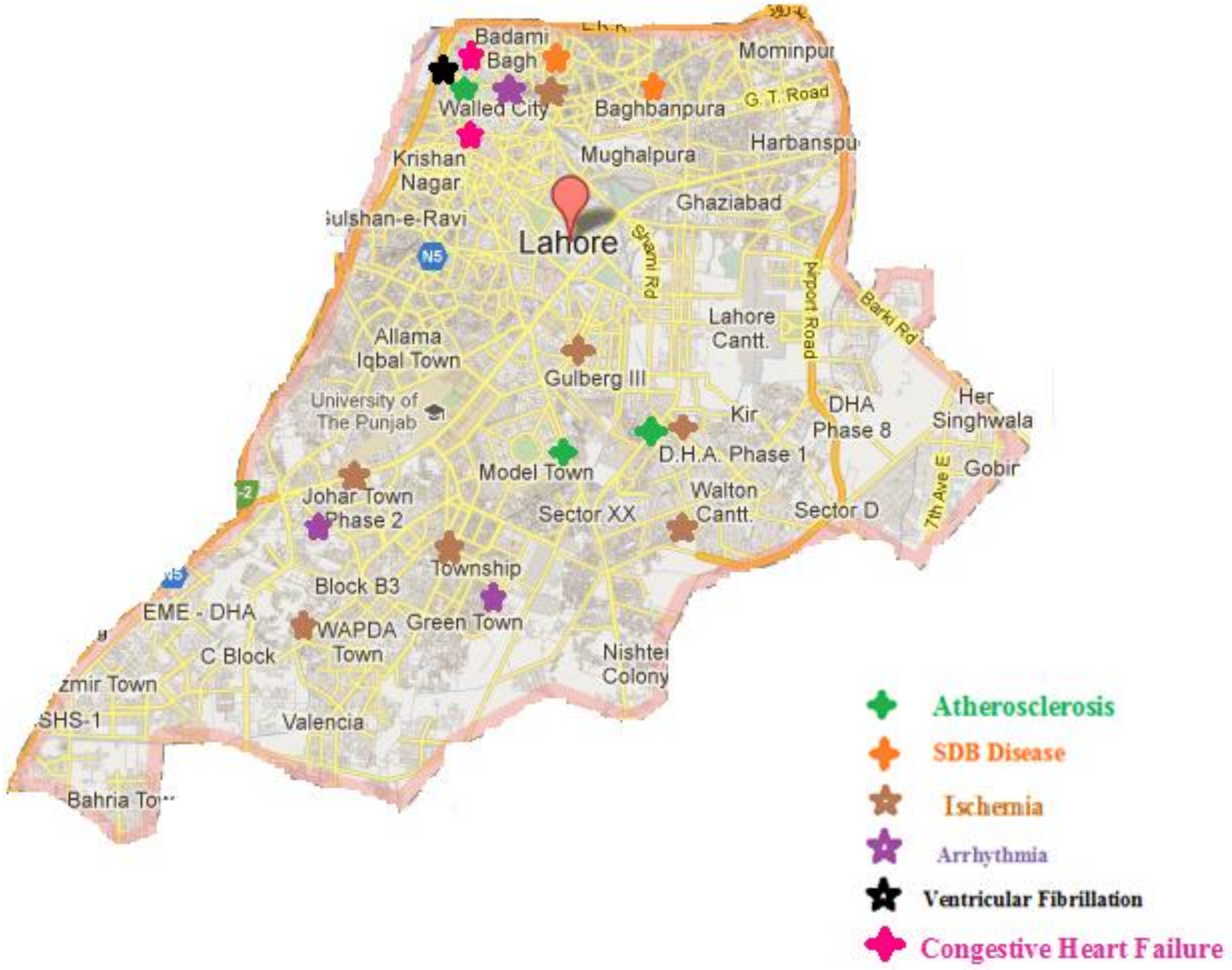


Fig 13: Map of Lahore showing CVDs in different areas

1. Ischemia was prominent disease in region A. The possible cause of the disease was hereditary, high blood pressure, and obesity. Moreover people whose age was from 45 to 75 years.

2. *Atherosclerosis* was prominent disease in region A. The possible cause of the disease was obesity. high blood pressure, smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 8000 to 100000 .

3. Ischemia was prominent disease in region B. The possible cause of the disease was , males , high blood pressure, hypertension , smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 25000 to 30000 .

4. Ischemia was prominent disease in region C. The possible cause of the disease was previous disease history , common in males , high blood pressure, hypertension , smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000 .

5. SDB was prominent disease in region C. The possible cause of the disease was Hypertension, gender , high blood pressure . Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000 .

6. *Atherosclerosis* was prominent disease in region C. The possible cause of the disease was high blood pressure, smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000 .

7. Arrhythmia was prominent disease in region C. The possible cause of the disease was previous history of CVD, high blood pressure, smoking , alcohol use and Hypertension. Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000.

8. Congestive Heart Failure prominent disease in region C. The possible cause of the disease was previous history of CVD, high blood pressure, smoking , alcohol use and Hypertension. Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000.

9. Ventricular fibrillation was prominent in region C. The possible cause of the disease was triggered activity of some other disease, medication side effects, history of other CVD and hypertension. Moreover, people whose income range was between 15000 to 20000 and number of dependent family members were 4 to 5.

These patterns reveal that CVDs in affluent class was due to obesity, physical inactivity, and these behaviors raise the cholesterol level and blood pressure. So the main risk factor is obesity and this can be controlled by the use of low fat and becoming physically active and also through regular exercise.

The reason of the CVDs in middle class was hereditary and hypertension and income problems and this situation becomes more adverse if person has large number of dependent family members. Moreover increased pollution in Lahore is the main cause of lung diseases and this also enhances stress which is already present in the form of hypertension due to socio-economic factors[34]. Moreover people living in rural areas have difficulty in accessing the medical facilities.
Also there are health hazards in these areas due to contamination of water due to industrial wastes as many minor industries are pumping wastes into the main canal. One important thing is also noticed that people living in these areas have small home industries which throw their wastes in the street canals and worsening the situation. So this thing should be controlled by the administration.

As far as lower middle class is concerned CVDs are the major cause of morbidity and mortality factor in those areas. People in these areas don’t have sufficient resources for meeting their basic necessities of life so making people hypertensive, moreover these areas are neglected by administration for providence of basic facilities, and cleanliness perspective so many infectious disease emerge from here and people don’t have access to specialist doctors and have to go to free dispensaries where no specialist or trained staff is present and wrong medicines not only worsen the condition of the patient but also triggers the other diseases and increases the mortality rate[35]. Use of contaminated water also worsens the condition. So the administration should be given the proper consideration to these areas by taking proper attention to these areas.

4.4.4 Analysis of CVD patterns in Gujranwala Region

Gujranwala is not a developed region . Even main city is not much developed and contains the basic medical facilities and facilities of cardiologists are not available. People have to move to move to other developed cities like Lahore and Islamabad for such treatment. While the people living in rural areas lack the basic facilities of medical health care [36]. Even if in those areas medical health units are present, they are not in working condition and doctors are not present there. People of those areas have to move to the main city for basic medical treatment. Population living in these areas is divided into middle, lower middle class and poor. ECG’s data of patients suffered from CVDs collected from those city hospitals by viewing the addresses from their profile and addresses and also personally by questionnaire and survey. The following patterns were found.

1. *Atherosclerosis* was prominent disease in region A. The possible cause of the disease was hereditary, obesity. high blood pressure, smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 50000 to 600000.

2. Ischemia was prominent disease in region B. The possible cause of the disease was males, high blood pressure, hypertension , smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 25000 to 30000 .

3. Ischemia was prominent disease in region C. The possible cause of the disease was previous disease history , common in males , high blood pressure, hypertension , smoking and diabetes. Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000 .

4. SDB was prominent disease in region C. The possible cause of the disease was Hypertension, gender , high blood pressure . Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000 .

5. *Atherosclerosis* was prominent disease in region C. The possible cause of the disease was high blood pressure, smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 8000 to 15000 .

6. Arrhythmia was prominent disease in region C. The possible cause of the disease was previous history of CVD, high blood pressure, smoking , alcohol use and Hypertension. Moreover people whose age was from 45 to 75 years and income range was between 8000 to 15000.

7. Congestive Heart Failure prominent disease in region C. The possible cause of the disease was previous history of CVD, high blood pressure, smoking ,drug and alcohol use and Hypertension. Moreover people whose age was from 45 to 75 years and income range was between 8000 to 15000.

8. Ventricular fibrillation was prominent in region C. The possible cause of the disease was triggered activity of some other disease, medication side effects, history of other CVD and hypertension. Moreover, people whose income range was between 8000 to 15000 and number of dependent family members were 4 to 5.

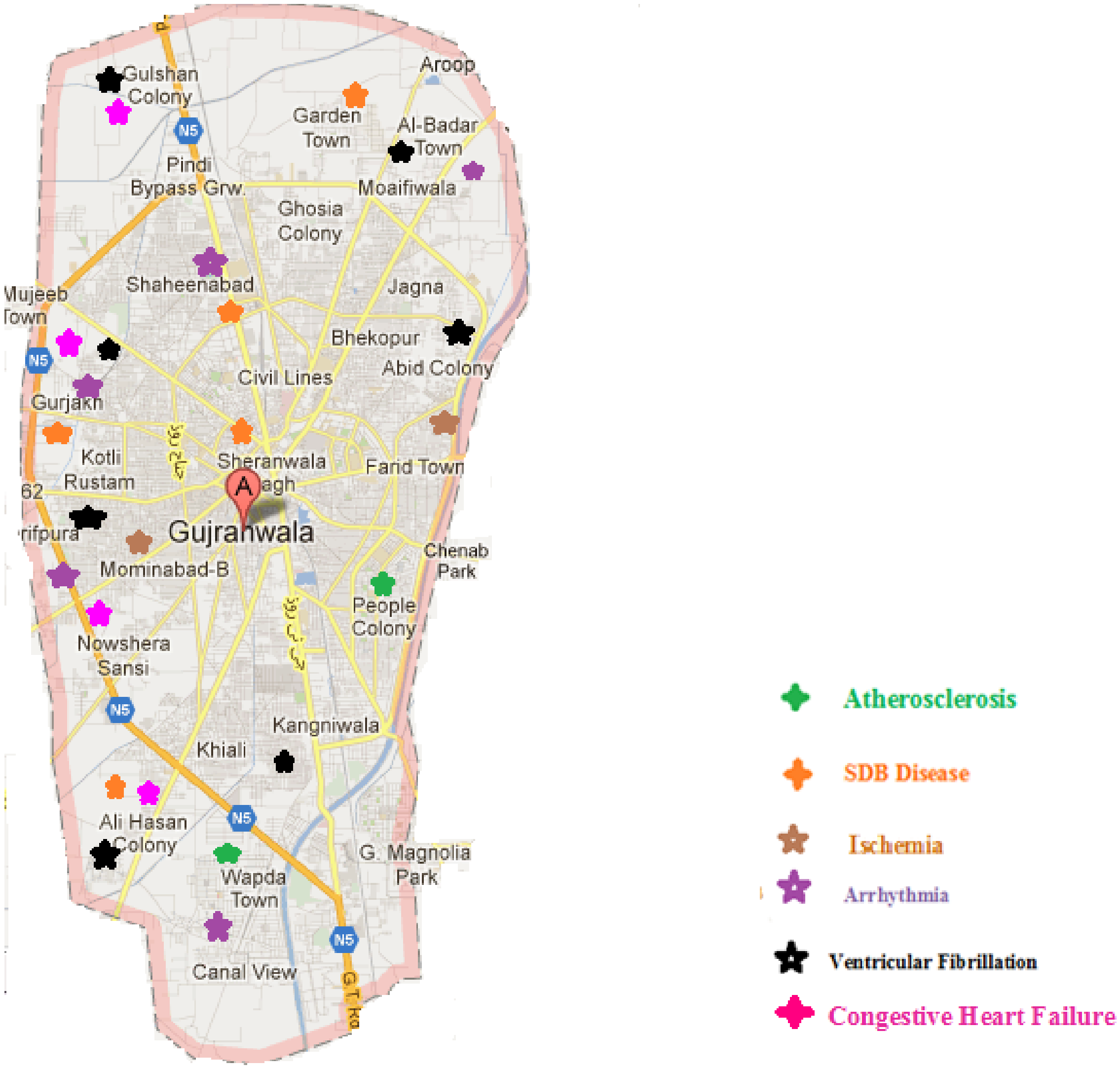


Figure 14: Map of Gujranwala showing distribution of different CVDs

The socio-economic status plays a vital role in the lives of human being . The lack of financial resources cause the hypertension and other CVDs in this area. The reason of the CVDs in middle class living in the highlighted areas in map such as peoples colony and satellite town was hereditary and obesity and becoming physically inactive so in such cases risk factors are controlled through exercise and becoming physically active. In some cases disease is hereditary and some have diabetes cause. This situation is avoided by medicine for controlling diabetes. In poor people who have low income and living in far off places specially in rural areas they even can't access basic facilities of health. They are more hypertensive and disease becomes more severe due to negligence towards their health moreover they don't have money to travel far off big cities and spending on their treatment and this situation leads to their death. Even in some areas of such as mohalla "Bakhtey wala" the tendency of smoking and drug addiction is seen through the patterns indicating that in such areas some criminals are selling narcotics due to which every third person out of ten is addict and this situation needs to be controlled by the administration for the betterment of society [37]. Moreover financial help and utilization of resources for the employment of unemployed people of such areas can be a solution, also free medical treatment and providence of medicines in such areas should be provided. The above suggested solution of treatment by the ECG machines and data retrieved trough internet will be a blessing for such areas.

4.4.5 Analysis of CVD patterns in Faisalabad Region

Faisalabad is the third largest city of Pakistan and is divided into six sub-divisions and have good economy as being industrial city. The main city is developed and have hospital and medical facilities while rural and sub-urban areas are not developed and even lacks in hospitals and basic medical facilities [38]. People from rural areas of Faisalabad have to come to the main city even for basic medical facilities and examination. By studying ECG data and patients profile various patterns of CVDs were found.

1. Ischemia was prominent disease in region A. The possible cause of the disease was hereditary, high blood pressure, and obesity. Moreover people whose age was from 45 to 75 years and income range was between 8000 to 100000 . .

2. *Atherosclerosis* was prominent disease in region A. The possible cause of the disease was obesity. high blood pressure, smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 8000 to 100000 .

3. Ischemia was prominent disease in region B. The possible cause of the disease was , males , high blood pressure, hypertension , smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 25000 to 30000 .

4. Ischemia was prominent disease in region B. The possible cause of the disease was previous disease history , common in males , high blood pressure, hypertension , smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 25000 to 30000 .

5. SDB was prominent disease in region C. The possible cause of the disease was Hypertension, gender , high blood pressure . Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000 .

6. *Atherosclerosis* was prominent disease in region C. The possible cause of the disease was high blood pressure, smoking and diabetes . Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000 .

7. Arrhythmia was prominent disease in region C. The possible cause of the disease was previous history of CVD, high blood pressure, smoking , alcohol use and Hypertension. Moreover people whose age was from 45 to 75 years and income range was between 15000 to 20000.

8. Congestive Heart Failure prominent disease in region C. The possible cause of the disease was previous history of CVD, high blood pressure, smoking , alcohol use and Hypertension. Moreover people whose age was from 45 to 75 years and income range was between 8000 to 14000.

9. Ventricular fibrillation was prominent in region C. The possible cause of the disease was triggered activity of some other disease, medication side effects, history of other CVD and hypertension. Moreover, people whose income range was between 8000 to 14000 and number of dependent family members were 4 to 5.

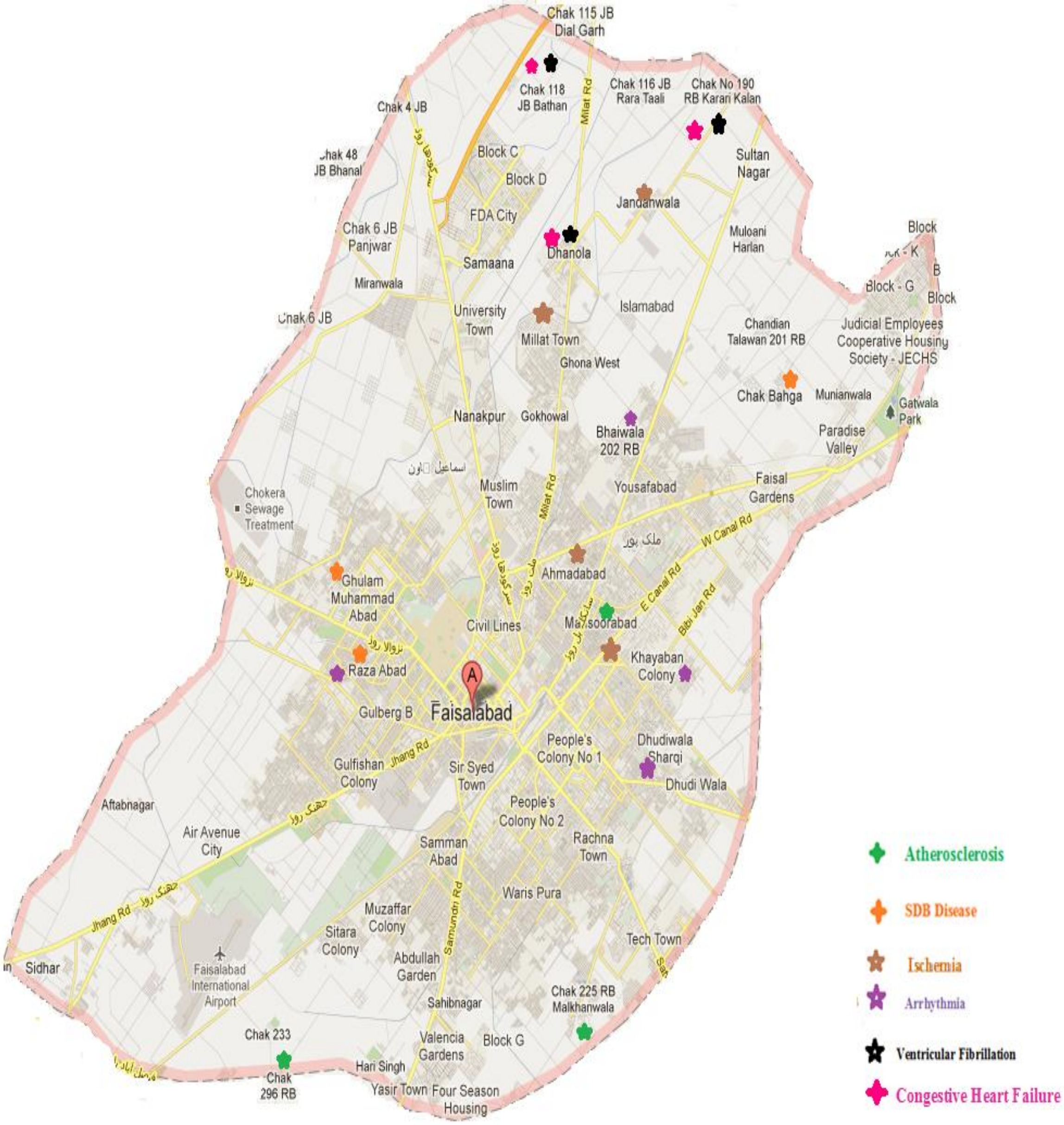


Figure 15 : Map showing the CVDs in Faisalabad

These patterns reveal that CVDs in affluent class as shown in the map in medina town was due to obesity, physical inactivity, and these behaviors raise the cholesterol level and blood pressure. So the main risk factor is obesity and this can be controlled by the use of low fat and becoming physically active and also through regular exercise.The reason of the CVDs in middle class livening in “Millat Town” was hereditary and hypertension and income problems and this situation becomes more adverse if person has large number of dependent family members. Moreover increased pollution in Faisalabad is the main cause of lung and liver diseases and this also enhances stress which is already present in the form of hypertension due to socio-economic factors[40]. Moreover people living in rural areas have difficulty in accessing the medical facilities.

Faisalabad being industrial city have numerous industries which don’t have proper disposal of wastes and dispose off their wastes directly into the city canals which is very dangerous to human beings. So this thing should be controlled by the administration by applying the strict rules regarding disposal of wastes. As far as lower middle class is concerned CVDs are the major cause of morbidity and mortality factor in those areas. People in these areas don’t have sufficient resources for meeting their basic necessities of life so making people hypertensive, moreover these areas are neglected by administration for providence of basic facilities, and cleanliness perspective so many infectious disease emerge from here and people don’t have access to specialist doctors and have to go to free dispensaries where no specialist or trained staff is present and wrong medicines not only worsen the condition of the patient but also triggers the other diseases and increases the mortality rate. Use of contaminated water also worsens the condition. So the administration should be given the proper consideration to these areas by taking proper attention to these areas.

4.5 Routing Medicines

Based on the trends analyzed in comparison with every location around the nation, it would become more clearer that which area has a tendency towards what arrhythmia or disease. Hence based on this information the Medicine required to treat that specific disease would be routed there as required and in the quantity required. This would improve the overall efficiency of the health care system as the medicine would be easily found where needed and there would be no problem of shortage as the shortfalls would always be calculated and remedied beforehand.

4.6 Future Recommendations

The research conducted in identifying arrhythmia patterns is complete however improvements and evolutions which should be implemented later and incorporated into this system are as follows:

Synchronization with NADRA

In the long run the government should aim for the centralization of all its records which would mean that the medical history collected through our system should be mapped onto the National

Database Records of every Pakistani as managed by NADRA using the National Identification Number as issued through our National Identity Card. When this synchronization has been achieved, medical records of any person should be easily accessible to all hospitals Nationwide via the National Health Repository System by simply entering the person's demographics or his National Identification Number. This would also mean that the government has a firmer control over its health care system and in a time such as this, where monitoring records of every person is very crucial to security, these records would prove to be very useful indeed.

Expansion to Cover Other Diseases

This system should be expanded to cover diseases other than arrhythmia. This should be central repository containing the complete medical records of every person rather than only arrhythmia. The benefits derived from this system would then be increased many fold if this recommendation is followed. Trend analysis, routing medicine, future trend prediction, remote data access, central record storage and faster detection of potential diseases could all be done for the wider range of all the possible diseases. Expansion for other diseases would simply mean expanding the database to cover for other medical records as well. No system from scratch would have to be built to cover this requirement.

4.7 Conclusion

The system proposed through this research is a dire need of our nation which requires a quick solution to the increasing medical atrocities she has to face on a day to day basis. With the present situation within the country spiraling out of control, the government also needs to implement checks as to who received its intended medical care and the effective health situation in any one area. The government needs to benefit from the functionalities which can be availed through this system and the health system of Pakistan is asking digitalization. This is one of the important research that Pakistan needs to implement to improve it's overall infrastructure especially in the field as important as health.

The proposed research not only concern the infrastructure of our nation. It also deals with the vast divisions in the medical field as well and how to overcome them as well. Providing health care to the rural areas and bridging the gap between the various types of hospitals is an essential task that we must tackle. Although for now the design is only for arrhythmias, it can be easily expanded to cater for other medical shortcomings as well. Digitalization of systems such as this is the only way to jump ahead and join the ranks of the most technologically advanced nations, benefitting from technology to provide solutions for its own people. This research is one such technologically advanced research which can help propel Pakistan's infrastructure decades ahead of its competitive countries into the ranks of the more advanced nations. Centralization of the health resource is necessary for the government. It will have to implement this, if not now, then maybe five or ten years in the future but the hard fact is that survival without is no question. Hence our research is aimed to provide an efficient and workable solution designed especially as per our own needs to provide all the mentioned functionality and more importantly, a degree of control over its resources through this system.

References

[1]. http://www.emro.who.int/pakistan/programmes_ncd.htm

[2]. http://www.who.int/cardiovascular_diseases/en/. accessed 19 September 2011

[3]. http:// tele-healthcare.org/2009/01/existing-infrastructure-of-health-care.html

[4]. Nishtar S. Health Indicators of Pakistan – Gateway Paper II. Islamabad, Pakistan: Heartfile; 2007.

[5]. Dawber, Thomas R., Meadors, Gilcin F. and Moore, Felix E. (1951), "Epidemiological Approaches to Heart Disease: The Framingham Study," American Journal of Public Health, 41, 279-286.

[6]. http://www.who.int/cardiovascular_diseases/en/

[7]. http://www.who.int/chp/chronic_disease_report/media/pakistan.pdf

[8]. Afsar Raza, Ashur Khan, Mujeeb-ul-Haq, Muhammad Imran Majeed, Ehsan Ahmed Alvi, Muhammad Zulfiqar Khan. Intravenous streptokinase for coronary thrombolysis. Aprospective study on more than 100 patients. Pakistan Armed Forces Medical Journal 41: 36-41, 1991.

[9]. Shah SM, Luby S, Rahbar M, Khan AW, McCormick JB. Hypertension and its Determinants among Adults in High Mountain Villages of the Northern Areas of Pakistan. J Hum Hypertens 2001;15(2):107-12

[10] Framingham Heart Study. Coronary Heart Disease. History of the Framingham Heart study [online]. Available at: www. framinghamheartstudy.org/about/history.html. Accessed May 25 , 2012.

[11]. Gordon, Tavia and Kannel, William B. (1972), "The Prospectve Study of Cardiovascular Disease," in Trends in Epidemiology: Applications to Health Service Research and Train-ing, (G.T. Steward, Ed.). Springfiled, Ill.: Charles C. Thomas, 198-211.

[12] Dawber, Thomas R. (1980), The Framingham Study: The Epidemiology of Atherosclerotic Disease, Cambridge, Mass.: Harvard University Press.

[13]. Higgins, Millicent W. (1984), "The Framingham Heart Study: Review of Epidemiological Design and Data, Limitations and Prospects," in Genetic Epidemiology of Coronary Heart Disease: Past, Present and Future, New York: Alan R. Liss, Inc., 51-64.P

[14] Dawber, Thomas R., and Moore, Felix E. (1952), "Longitudinal Study of Heart Disease in Framingham, Massachusetts: An Interim Report," in Research in Public Health, Papers presented at the 1951 Annual Conference of the Milbank Memorial Fund, 241-247.

[15]. Dawber, Thomas R., Kannel, W.B. and Lyell, L. (1963), "An Approach to Longitudinal Studies in a Community: The Framingham Study," Annals of the New York Academy of Sciences, 107, 539-556.

[16] Dawber, Thomas R., Meadors, Gilcin F. and Moore, Felix E. (1951), "EpidemiologicalApproaches to Heart Disease: The Framingham Study," American Journal of Public Health, 41,279-286.

[17] Dawber, Thomas R.,“ The Framingham Study: The Epidemiology of Atherosclerotic Disease” Cambridge, MA : Harvard University Press 1980. ISBN : 0674317300

[18] http://www.sph.umn.edu/epi/ecg/history/ Accessed 28 May 2012.

[19]. Syed Muhammad Imran Majeed, Muhammad Khalid Raja, Masud-ul-Hasan Nuri, Syed Afzal Ahmed. Brugada Syndrome: Intermittent manifestation of the electrocardiographic abnormality.
Journal of College of Physicians and Surgeons Pakistan

[20] Han, j. and M. Kamber, *Data Mining Concepts and Techniques*. 2006: Morgan Kaufmann Publishers.

[21] Ahmad Jawad Asghar, Haaris Bin Ghafoor , Irfan Ijaz and M. Zain Mustafa , *“Thesis on Storage Area Network For National Medical Records And Data Mining For Trend Analysis*”, at CEME NUST . Degree 29 Session 2007-2011 .

[22]. http://en.wikipedia.org/wiki/K-means_clustering Accessed 28 May 2012.

[23] Wa’Eljuma’Ahal_Zyadat, Rodzia Bintiatan , Hamidah Ibrahim , “ The Directs Impact to Pre-Filtering Process to Weather Datasheet“ Masrah Azrifa and Azmi Murad , in *Journal of Theoretical And Applied Information Technology*. Vol. 26 N0.1 ISSN 1817-3195

[24]. http://en.wikipedia.org/wiki/Mean Accessed 28 May 2012.

[25]. http://en.wikipedia.org/wiki/Standard_deviation Accessed 28 May 2012.

[26]. Late potentials in hypertensive patients with left ventricular hypertrophy, Pakistan Cardiac Society 12th International Congress of Cardiology, Islamabad, Pakistan, April 1996

[27]. Gordon, Tavia and Kannel, William B. (1968b), "The Framingham Study: Follow-up to the Eight Exam," in The Framingham Study: An Epidemiological Investigation of Cardiovas-cular Disease, Section 2, (W.B. Kannel and T. Gordon, Eds.). Bethesda: National Heart In-stitute.

[28]. http://en.wikipedia.org/wiki/Islamabad Accessed 28 May 2012.

[29]. S.A. Hussain Shah, S.A. Shami and Samina Jalali, “Consanguinity and Family History: Risk Factors of Cardiovascular Diseases”, Pakistan Journal of Zoology., vol. 36(4), pp. 301-305, 2004.

[30]. http://en.wikipedia.org/wiki/Rawalpindi Accessed 28 May 2012.

[31]. Shami, S.A., Schmit, L.H. and Bittles, A.H., 1989. Consanguinity related prenatal and postnatal mortality in the populations of seven Pakistani Punjab cities. *J. Med. Genet*., 26: 267-271.

[32]. Ch. Naseer Ahmad, Azizullah Khan, Ghalib Hasnain and Shahid Durez, “ Water Supply Problems in Rawalpidi City”, World Water Day- April 2011.

[33]. http://en.wikipedia.org/wiki/Lahore Accessed 28 May 2012.

[34]. M. Perwaiz Iqbal, “Lead pollution – A risk factor for cardiovascular disease in Asian developing countries”, Pak. J. Pharm. Sci., Vol.25, No.1, January 2012, pp.289-294

[35]. Tauqeer Hussain Shah, Huma Butt "Sleep Comes All The Way: A Study Of Homeless People In Lahore, Pakistan", SAVAP International Volume 1, Issue 3, November 2011 ISSN: 2223-9553

[36]. http://en.wikipedia.org/wiki/Gujranwala Accessed 28 May 2012

[37]. Sara Hitchman, Lorraine Craig, Pete Driezen, Michelle Bishop, and Geoffrey T. Fong, "Cardiovascular harms from tobacco use and secondhand smoke", global gaps in awareness and implications for action WHO report April 2012.

[38]. http://en.wikipedia.org/wiki/Faisalabad Accessed 20 June 2012.

[39]. M. I. Zafar, S. R. S. Abbasi*, Z. Batool and I. Shahid, " A Study Of Medical Facilities Provided By Punjab Employees Social Security Institutions To The Labourers In The Faisalabad City", J. Anim. Pl. Sci. 18(2-3): 2008.

[40]. Khan NI, Naz L, Mushtaq S, Rukhl, Ali S And Hussain Z, "Ischemic Stroke: Prevalence Of Modifiable Risk Factors In Male And Female Patients In Pakistan", Pak. J. Pharm. Sci., Vol.22, No.1, January 2009, pp.62-67

Appendix A

Serial No	Variable	Description	Definition	
Demographic				
1				
2				
3				
4	Height	Height in **meters** of each participant for BMI calculation	Integer	
5	Age	Age of each participant/patient in **years**	Year	Year
			A- 00	B-1-4
			C-5-9	D-10-14
			F- 15-19	G-20-24
			H-25-29	I-30-34
			J-35-39	K-40-44
			L-45-49	M-50-54
			N-55-59	O-60-64
			P-65-69	Q-70-74
			R-75-79	S-80-84
			T-85+	
6	Education	Education of each participant/patient	A-None B-Fifth grade or less C-Sixth to Ninth Grade D-Matric/Equivalent E- Fsc/Equivalent Under way F-Fsc/Equivalent Cleared G-Under Graduate Not Cleared H-Under Graduate I-Post Graduate U-unknown	
7	Marital Status	Marital status of each participant/patient	A-Single B-Married C-Widowed D-Divorced E-Separated F-Married more than once H-Not Apllicable(age<15)	
8	Employment			

		Employment Status of Patient/participant	A- Unemployed(> 3 months) B- Recently Unemployed (< 3 months) C- Employed D- self Employed including Employer E-Retired F- Not working due to family Constraints G- Disabled cannot work H- Student > 16yrs I- child < 16yrs
9	Employment category based or pay scale	Per month income from job/business	A < 6000Rs/month 6000Rs/month < B < 10000Rs/m 10000Rs/month < C < 15000Rs/month 15000Rs/month < D < 25000Rs/month 25000Rs/month < E < 40000Rs/month 40000Rs/month < F < 60000Rs/month 60000Rs/month < G < 80000Rs/month 80000Rs/month < H
10	Dependent Famil members	Members of family directly dependent on patient/participant	A- None B- Only one C- 2-4 D- 5-7 E- 8-10 F- >10
11	Number of bread earners in the family	Earning Hands in family whose income is used to run household	A- Only one B- Two C- > 2
12	Any other source income	Any other source of family income example property	A- Yes B- No
13	Total family income	Sum of income of all the sources in family	Integer
14	Total family members	Total no. of family members living in the house for more than 3 months	Integer

15	Per capita income	$\frac{Total\ family\ income(Rs)}{Total\ no.of\ family\ members}$	?
16	Permanent Addre	Permanent address of participant/patient (estimating population density)	Street: Colony: City: Province :
17	Present Address	Present address of patient/participant	Street: Colony: City: Province :

18	Displaced	Living in the area patient belongs to:	A:Yes(Permanent address = Present Address)
			B:No(Permanent address!=Present Address)

ECG			
1	P-P interval Variation		A < 0.16 B > 0.16
2	R-R interval Variation		A < 0.16 B > 0.16
3	P-P interval duration		0.6 < A < 1 0.24 < B < 0.4 0.33 < C < 0.6 D > 1
4	R-R interval duration		0.6 < A < 1 0.33 < B < 0.6 C > 1 1 > D < 1.5 E > 1.5

5	Atrial rate		60 < A < 100 B < 40 C < 60 100 < D < 180 150 < E < 250 250 < F < 400 G > 400
6	Ventricular rate		60 < A < 100 B < 40 C < 60 100 < D < 180 E > 100
7	P-wave		
8	P-R interval duration		0.12 < A < 0.2 B > 0.2
9	QRS interval duration		0.06 < A < 0.1 B > 0.1
10	T wave		0.36 < A < 0.44
11	Q-T interval duration		
12	ST segment shift		-0.05 < A < 1 B < -0.05 C > 1
Hospitals/Cardiologist			
1	Hospitals in area	Hospitals in range of participants/patients residence	A < 3km radius 3km radius < B < 5km radius 5km radius < C < 10km radius 10km radius < D < 20km radius 20 km radius < E < 30 km radius

			F > 30km radius G-Unknown	
2	Availability of cardiologist in hospitals	Whether or not cardiologist is available in hospitals in range	A- Not available B- Unknown C- Occasionally available D- Available	
3	Hospital condition	Condition of hospital including presence of ambulance, medicine, necessary medical equipment.	A- Unknown B- not Satisfactory C- Average D- Above average E-Good	
General Medical Investigations				
1	BM I $\frac{weight/kg}{(height/m)^2}$	Body Mass Index ranges	A < 16kg/ m^2 16 kg/m^2 < B < 20 kg/m^2 20 kg/ m^2< C < 25 kg/ m^2 25 kg/m^2 < D < 30 kg/m^2 30 kg/m^2< E < 35 kg/m^2 35 kg/m^2 < F < 40 kg/m^2 40 kg/m^2 < G	
2	Blood pressure	Blood Pressure of participant/patient	Systolic(mm/hg)	Diastolic(mm/hg)
			140<A<210	90<A<120
			110<B<130	75<B<85
			50<C<90	33<C<60
3	Breathing rate	Breathing rate according to age group	Age Group	Breathing Rate (Breaths/min)
			00-Years	30<B<60
			1-3 Years	A<24 24<B<40 C>40
			3-6 Years Per School	A<20 22<B<34

				C>35
			6-12 Years	A<16 16<B<30 C>30
			12 to above	A<12 12<B<20 C>20
Medical History				
1	Acute Infections		A- Negative/None B- Diphtheria C- Scarlet Fever D- Frequent Sore Throat E- Diphtheria And Scarlet Fever F- Frequent Sore Throat G- Scarlet Fever And Frequent Sore Throat H- Diphtheria, Scarlet Fever, And frequent sore throat I- Unknown	
2	Asthmas/Allergies		A- Negative B- Allergy, Alone C- Bronchial Asthma, Alone D- Allergy And Asthma, Together E- Unknown	
3	History of Hypertension		A- Negative B- Transient C- Permanent D- Type Unknown E- Doubtful F- Unknown	
4	History of non-Cardiovascular Disease, Exam 1	(Peptic Ulcer, Chronic Colitis, Kidney Disease)		

			A- Negative B- Peptic Ulcer C- Chronic Colitis D- Kidney Disease E- Ulcer And Colitis F- Ulcer And Kidney Disease G- Colitis And Kidney Disease H- Ulcer, Kidney Disease, And Colitis I- Doubtful J- Unknown	
5	History of Other Cardiovascular Disease		A- No B- Yes C- Unknown	
6	Number Of Pregnancies		A- Man, Single Woman, or no pregnancies for non-single woman B- One Pregnancy reported C- Two Pregnancies reported D- Three pregnancies reported E- Four pregnancies reported F- Five pregnancies reported G- Six pregnancies reported H- Seven pregnancies reported I- Eight or more pregnancies reported J- Unknown	

7	History of Drug Usage		Drug Type
			Cigarettes
				Intensity (Smoked Per day)
				X=1 still smoking
				X=2 Quitted
				A= Never Used