UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

Discovering Medical Diagnosis Patterns in Electronic Medical Records using Association Rule Mining

By

Stephen Mungai Kang'ethe

P58/75902/2012

Supervisor

Prof. Peter Wagacha

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science in the School of Computing and Informatics, University of Nairobi

2014

DECLARATION

This project, as presented in this report, is my original work and has not been presented for a degree in any other university.

Stephen M. Kang'ethe
P58/75902/2012

Signed: ……………………

Date: ….………………

This thesis has been submitted for examination with my approval as university supervisor.

Prof. Peter W. Wagacha
School of Computing & Informatics

Signed: ………………………

Date:..………………………..

DEDICATION

To my dear wife Winnie Nyoro-Mungai and daughter Leona Njeri whose laughter has kept me going throughout this project and whose joyful spirits have given me a reason to wake up and look forward to another day.

To my parents Kang’ethe Mungai and Monicah Njeri who have supported and encouraged me to go on and pursue my dreams and be the best wherever my interests take me.

ACKNOWLEDGEMENT

I first wish to thank the School of Computing and Informatics of the University of Nairobi for according me the opportunity to pursue this course at this great Institution.

I would also like to thank my project supervisor, Prof. Peter Waiganjo for his indispensable guidance and dedication of time towards the achievement of this project's objectives. I wish to extend this gratitude to the panel members who also gave much needed advice throughout the project.

Many thanks also go to medical colleagues of the United Nations Joint Medical Service for their expert advice before and during the entire period of this thesis. In particular I would like to thank Dr. James Mbai for his extended dedication in handling my sometimes "over inquisitiveness" which I needed to successfully execute this project.

Lastly I wish to acknowledge my entire family's support throughout my studies, without which none of this would have been possible.

ABSTRACT

Data mining technologies have been used extensively in the commercial retail sectors to extract data from their “big data” warehouses. In healthcare, data mining has been used as well in various aspects which we explore. The voluminous amounts of data generated by medical systems form a good basis for discovery of interesting patterns that may aid decision making and saving of lives not to mention reduction of costs in research work and possibly reduced morbidity prevalence. It is from this that we set out to implement a concept using association rule mining technology to find out any possible diagnostic associations that may have arisen in patients’ medical records spanning across multiple contacts of care. The dataset was obtained from Practice Fusion’s open research data that contained over 98,000 patient clinic visits from all American states.

Using an implementation of the classical apriori algorithm, we were able to mine for patterns arising from medical diagnosis data. The diagnosis data was based on ICD-9 coding and this helped limit the set of possible diagnostic groups for the analysis. We then subjected the results to domain expert opinion. The panel of experts validated some of the most common associations that had a minimum confidence level of between 56-76% with a concurrence of 90% whereas others elicited debate amongst the medical practitioners. The results of our research showed that association rule mining can be used to confirm what is already known from health data in form of comorbidity patterns while generating some very interesting disease diagnosis associations that can provide a good starting point and room for further exploration through studies by medical researchers to explain the patterns that are seemingly unknown or peculiar in the concerned populations.

Table of Contents

Table of Figures

Acronyms

CCR	Continuity of Care Record
CDC	Center for Disease Control
CDO	Care Delivery Organization
EHR	Electronic Health Records
EMR	Electronic Medical Records
HIS	Health Information System
HITECH	Health Information Technology for Economic and Clinical Health Act (2009)
HL7	Health Level 7 International
ICD	International Statistical Classification of Diseases
ISO	International Organization for Standardization
WHO	World Health Organization

Definition of Key Terms

Multimorbidity - The co-occurrence of multiple chronic or acute diseases and medical conditions within one person without any reference to an index condition. (van den Akker et al., 1998). Also comorbidity.

Standardized EMR - In our context, this refers to an EMR that complies with the guidelines developed by the Kenya government contained in the document (*Standards and Guidelines for Electronic Medical Record Systems in Kenya*, 2009), as well as other guidelines particularly those of the world health governing body WHO.

Differential Diagnosis- is a systematic diagnostic method used to identify the presence of an entity where multiple alternatives are possible (and the process may be termed differential diagnostic procedure), and may also refer to any of the included candidate alternatives (which may also be termed candidate condition) ("Differential diagnosis," 2014).

CHAPTER ONE

INTRODUCTION

1.0 Introduction

The health sector worldwide has been involved in automation of medical records worldwide not to be left behind in the digital technology age. Medical practitioners have had to learn new ways of capturing their findings and treatment plans on their patients after having had years of the same on paper. Different health institutions who have adopted Electronic Medical Record systems (EMR) have done it in their own ways before owing to the lack of standardization of such implementations in the years past. In recent times however, world governing institutions like WHO and ISO have embraced the advent of Health Information Systems (HIS) and spearheaded the development of standards that were hitherto unavailable to implementers of health systems. These standards make it easy not only to capture and share data across multiple and seemingly disparate implementations, but to also query, analyze and extract useful statistics from data entered in the same systems.

Data mining technologies have been used extensively particularly in the commercial retail sectors to extract data from their "big data" warehouses. In healthcare, data mining has been used as well in various aspects which we will explore later. The voluminous amounts of data generated by these systems form a good basis for discovery of interesting patterns that may aid decision making and saving of lives not to mention reduction of costs in research work and possibly reduced morbidity prevalence. It is from this that we seek to implement a concept using association rule mining technology (ARM) to find out any possible diagnostic associations that may arise in patients' medical records spanning across multiple contacts of care (visits).

1.1 Background

The term EMR stands for Electronic Medical Records. In their work, "*Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference*", (Garets and Davis, 2006) define an EMR system as an application environment composed of the clinical data repository, clinical decision support, controlled medical vocabulary, order entry, computerized provider order entry, pharmacy, and clinical documentation applications. This environment supports the patient's electronic medical record across inpatient and outpatient environments, and is used by healthcare practitioners to document, monitor, and manage health care delivery within a care delivery organization (CDO). The data in the EMR is the legal record of what happened to the patient during their encounter at the CDO and is owned by the CDO. This is to be differentiated with Electronic Health Record systems (EHR) which they again define as a subset of each care delivery organization's EMR, presently assumed to be summaries like ASTM's Continuity of

Care Record (CCR) or HL7's Continuity of Care Document (CCD), is owned by the patient and has patient input and access that spans episodes of care across multiple CDOs within a community, region, or state (or in some countries, the entire country). The terms are often used interchangeably though the difference, subtle as it may seem, may be of particular significance in this research.

EMRs have been in use in several countries by different health facilities over the years but standardization of the different electronic medical records implementation has been left as an individual task for different governments to pursue. The US, research has shown, had been lagging behind other (particularly Scandinavian countries) in the adoption of EHRs (Schoen et al., 2009). This is changing as they have aggressively embarked with the implementation of the Health Information Technology for Economic and Clinical Health Act (HITECH) of 2009, which provides $27 billion over 10 years for federal incentive payments to hospitals and clinicians for adopting EHRs (Gray et al., 2011). The use of certified electronic health records (EHR) and pertinent objectives to be achieved over several stages are known as "Meaningful Use".

Closer home, the government of Kenya openly admits to challenges of obtaining health data, due to the weak health information infrastructure, a poor information culture that does not spur demand for information, multiple and parallel information systems, a thin and stretched human resource to support data collection, transformation, presentation and archiving among others. This is in its health information policy and strategic plan (*Standards and Guidelines for Electronic Medical Record Systems in Kenya*, 2009). The Division of Health Information Systems (HIS), in this policy document continues to say: "It is with this background that the ministries through the Division of Health Information system (HIS) undertook to develop a health information policy and strategic plan (2009- 2014) to guide the health information strengthening agenda in the country. In its Strategic Plan, the HIS has planned to improve data management and strengthen the use and application of information technology in data management. To effectively do this, there is need to develop standards that will ensure quality of software, compatibility of data sharing, ease of maintenance and common understanding among the workforce.

Data complexity, volumes of patients served and the desire to have efficient health information systems have defined the need for Electronic Medical Record (EMR) Systems in Kenya. EMR systems, when well developed and implemented, can improve the process of data collection resulting in better quality and more reliable health information. These systems can also greatly improve aggregation and reporting of data from facilities. EMR systems support provision of health care through the integrated clinical decision support functions and by ensuring that patient information is available across facilities for continuity of care."

The policy document goes ahead to lay a regulatory framework that is based on international standards from institutions such as WHO, ISO and CDC. The standards that will be of particular interest in our research are the International Classification of Diseases (ICD) standards, (both ICD-9 and ICD-10) and HL7 health information interchange standards.

This only demonstrates the significance of standardized EMR systems and the evolutionary role they are likely to play not only in the world in general but in Kenya in particular.

It is with this standardized data capture and storage, that there emanates useful data which can not only be analyzed but that can also have useful patterns discovered that could aid governments and medical practitioners alike in improving healthcare services to the public and their patients respectively.

1.2 Problem Statement

The availability of standardized medical data creates a large pool of data with a lot of hidden and potentially useful information. Where there are a few records from a medical institution, it is possible to apply simple, semi-automated methods and tools to do analysis, albeit with a sacrifice in accuracy speed, reliability and validity of findings. However, when "big data" is employed there is need to employ adequate methods to enable the realization of full benefits of such data. The data obtained from EMR systems is expensive to acquire and maintain and therefore holds potential for secondary use key amongst which is unraveling the interesting patterns that could lie therein.

1.3 Research Objectives

In order to achieve the aim of this research, we establish the following key objectives:

1. Identify and adapt for use, an association rule mining algorithm to patient diagnostic data obtained from one or more standardized EMR, using ICD coding as its basis.

2. Demonstrate that we can generate and discover strong rules (relationships) that indicate multimorbidity trends from the EMR data with varying confidence levels.

3. Match these newly discovered associations to patient demographics and extract new knowledge from them.

4. Validate the discovered associations against existing domain knowledge.

1.4 Research Questions

At the end of the research, we will seek to answer the following questions:

1. How has ARM been applied in other areas and how if at all, has it been used in health systems and can we apply association rule mining to EMR's ICD coded diagnosis data to discover new or possibly unknown patterns, or support and reinforce already known associations?

2. What levels of multimorbidity can be extracted from this data and with what confidence can we indicate them?

3. What are the relations between these comorbidity trends and patient demographics for example age and gender and can any linkages be derived from them?

4. Are there any similarities between our results and what is already known? If not, how different are the results?

1.5 Significance of the Research

The research is based on a model that can be replicated across multiple EMR implementations as long as they adhere to the stipulated standards and have the same output format. This means that we will have the ability to apply this technique to data that could span say the entire country or continent in a bid to discover new disease multimorbidity patterns. Based on the newly discovered strong associations, we are able to know which diseases tend to appear together from amongst the patients, which could help the government place intervention measures in advance. This could include putting in place health measures requiring pathological tests for a certain disease given that another closely associated one has been diagnosed.

Based on the findings, policy makers could also focus on conducting health campaigns for certain diseases with the knowledge that the success of such will essentially have a certain related effect on the other associated diseases.

It also helps the health industry to finally take advantage of years' worth of input, in the sense that it will be possible to utilize multiple sources of medical data to aid in decision making given certain morbidity patterns. This comes in a welcome secondary use for the data since there is value addition to primarily medical records collected for other purposes.

1.6 Assumptions

The research exists within a well governed health domain and as such makes several basic assumptions. Key to this is that the diagnosis codes used are from the ICD coding standards. This is to limit the number of diagnosis groups and to ensure that we have a consistent pool of data to draw our comparisons from.

We also make an assumption that the practitioners observed the guidelines in (*ICD-9-CM Official Guidelines for Coding and Reporting*, 2011). This, amongst others, provides for recording of the most accurate diagnosis describing the patient's condition and avoiding "provisional" or "working" diagnosis and where diagnosis is "probable" or "questionable", it is coded as if it existed (*ICD-9-CM Official Guidelines for Coding and Reporting*, 2011, p. 92).

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

In this chapter we will highlight the efforts that have been put in the realization of electronic health records not just by the global community but by Kenya as well. We also look at several standardization efforts like ICD and HL7 that are making it easier by the day to implement these systems and to share data across multiple system implementations. We highlight the importance that this plays in the realization and enabling role it plays in our research.

We will also go further to look at how data mining has been used by other researchers in the health industry and in particular the use of the Apriori algorithm in the same. We will look at how and why the algorithm best suits our research in comparison with other data mining techniques for association mining. We delve deeper into our specific realm, the electronic medical records use of the same and highlight a few works on what others have done in this field.

In their work, *Fast algorithms for mining association rules in large databases,* (Agrawal and Srikant, 1994), the authors presented an algorithm, known as Apriori, for discovering association rules within large, primarily transactional, sales databases. This algorithm was a development of previously known algorithms for itemset mining and association rules discovery. We have a brief look at how this algorithm works and its known uses in the commercial, particularly retail sales databases, for which the authors admit the algorithm was originally conceived. We will also explore the benefits accrued by using this algorithm over other known algorithms for association rules mining.

Lastly, we look to acquaint ourselves with the application of Apriori in previous EMR systems both standardized and non-standardized, and indicate the research gaps and weaknesses that these present, and that we intend to address in this research.

2.1 EMR Systems Adoption in the Kenyan and Global Context

2.1.0 The Kenyan context

Earlier in this report, we made a distinction between Electronic Medical Records (EMR) and Electronic Health Records (EHR) as defined by (Garets and Davis, 2006, p. 2). EHR systems play a much wider role as they span multiple EMR's whether integrated or not, several Care Delivery Organizations (CDOs) which could include hospitals, emergency or ambulatory care services, psychological or mental institutions and other health organizations. The government of

Kenya in this context has been keen to strategically spearhead the adoption of EMRs in the country. This is with the establishment of guidelines through the Division of Health Information System (HIS). Though the government's main aim, as it indicates in the policy, is to improve health data management, use of ICT in health and ease sharing of health data, there will be a lot of other benefits in the long run that will be as a result of this, part of which we wish to exploit in this research.

Currently, Kenyan health care givers implement multiple non-standardized EMR systems. Each organization adopts a system, either in-house developed or off the shelf, based on customized needs. Examples of such systems include OpenMRS, IQ-CARE, and C-PAD amongst others.

The government has chosen to adopt standards adopted by other "partner" countries the world over, mainly guided by WHO and ISO standards. Of particular interest to our case, the guideline requires any system being implemented to have the ability to maintain a coded list of problems/diagnoses (*Standards and Guidelines for Electronic Medical Record Systems in Kenya*, 2009, p. 23). It also goes ahead to indicate that the EMR ought to maintain a "Problem List" associated with the patient, its status, and the coded list of problems/diagnoses. These are some of the developments that have developed our interest in this research.

2.1.1 The Global Context

The federal government of United States of America has been on the headlines in recent times in the adoption of EMRs in the country. A survey of eleven countries carried out by (Schoen et al., 2009) found USA to be lagging behind other mainly Scandinavian countries. As (Gray et al., 2011) argue, the availing of $27 billion over 10 years to health providers by the US government for not just adopting EHRs but attaining "Meaningful Use" in improving patient care is set to change the outlook. This has an effect of ensuring that EHR implementers trample on each other to deliver systems to their clients seeking to attain Meaningful Use objectives in order to benefit from the financial incentives. According to the US government's official health IT website, the Meaningful Use objectives are divided into three stages, each with a target date as follows ("Meaningful Use Definition & Objectives," n.d.):

Stage 1: Data capture and sharing (2011-2012).

Stage 2: Advance clinical processes (2014).

Stage 3: Improved outcomes (2016).

Meaningful use is defined as using certified electronic health record (EHR) technology to:

- Improve quality, safety, efficiency, and reduce health disparities.
- Engage patients and family.

- Improve care coordination, and population and public health.
- Maintain privacy and security of patient health information.

They go on to outline that the meaningful use compliance will result in:

- Better clinical outcomes
- Improved population health outcomes
- Increased transparency and efficiency
- Empowered individuals
- More robust research data on health systems (which again we are going to benefit from in this research).

Soon more countries are expected to jump in and whereas they may not start out with the meaningful Use strategy as their strategy, EHR adoption alone is sure to set them on this path, and more data will be available for researchers in this area with all the benefit that their output brings.

2.2 Major Standards

There are several standards that are in use across EMR systems. Due to the scope of this research, we only focus on those that touch on medical diagnosis.

2.2.0 International Statistical Classification of Diseases (ICD)

International Statistical Classification of Diseases is the standard diagnostic tool for epidemiology, health management and clinical purposes. ("WHO | International Classification of Diseases (ICD)," n.d.). It contains standard diagnostic codes that attempt to cover all known morbidity and mortality causes statistics.

ICD-10 is the current standard and is a replacement of the widely used ICD-9. The latest version is the 2010 version. ICD 9 has also been in use for a while and is in the process of being replaced by ICD 10. WHO also state that the 11th revision of the classification (ICD-11) is in place and is set to go on until the year 2017 ("WHO | World Health Organization," 2014).

ICD-9, codes are three to five digits. The first digit is either numeric or alpha (the letters E or V only) and all other digits are numeric as shown in Figure 2.1 ("ICD-10 Conversion and Mapping - AAPC," 2014)

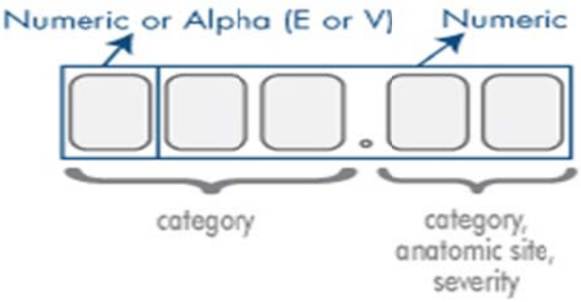


Figure 2.1 The structure of an ICD-9 code

In ICD-10-CM, however, codes can be up to seven digits. The first digit is always alpha (it can be any letter except U), the second digit is always numeric, and the remaining five digits can be any combination (see Figure 2.2 below).

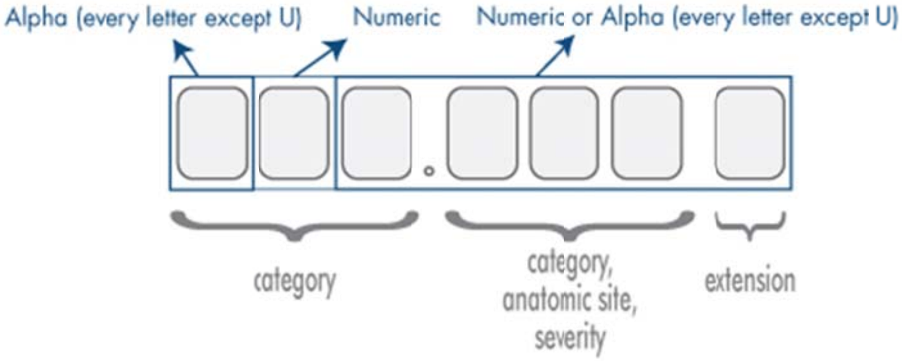

Alpha (every letter except U)
Numeric
Numeric or Alpha (every letter except U)
category
category, anatomic site, severity
extension

Figure 2.2 The structure of an ICD 10 code

For our research, any EMR that uses either of the two codes will suffice provided they are used consistently across the implementation.

2.2.1 Health Level 7 Standards (HL7)

Health Level Seven International (HL7) is a not-for-profit, ANSI-accredited standards developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services (“About Health Level Seven International,” 2007). According to HL7 International, the most widely used standard is the HL7 V2.7 updated in 2011. Chapter 12 (patient care) supports most of the standards basis of this research. The document states that a problem of a given individual can be described by formal diagnosis coding systems (such as DRGs, NANDA Nursing Diagnosis, ICD9, DSM, etc.) or by other professional descriptions of healthcare issues affecting an individual. Problems can be short- or long-term in nature, chronic or acute, and have a status. In a longitudinal record, all problems may be of importance in the overall long-term care of an individual, and may undergo changes in status repeatedly. Problems are identified during patient visits, and may span multiple visits, encounters, or episodes of care. The problem detail segment goes on to define specific message interchange standards and machine-human readable messages formats ("HL7 Standards V2.71," 2011).

These are important as they enable different EMRs to exchange and output health data in a common format where we can mine our patterns from.

2.3 Association Rule Mining and the Apriori Algorithm.

Association rule mining has been used extensively in the commercial industry particularly in the retail sector. It has mainly been used to do market basket analysis where the focus is on analyzing the contents of the customer's "basket". As (Berry and Linoff, 2004, p. 287) explain, Market basket analysis provides insight into the merchandise by telling us which products tend to be purchased together and which are most amenable to promotion. Association rules identify strong relations that exist in databases using several measures of interestingness (usually based on minimum support and minimum confidence) (Matheus et al., 1993).

The patterns discovered may have different uses in nature and they may be categorized as actionable rules (contain high-quality, actionable information), trivial rules (already known by anyone at all familiar with the business) or inexplicable rules (these seem to have no explanation and do not suggest a course of action), (Berry and Linoff, 2004, pp. 296–298).

When large databases are involved, an efficient algorithm to find frequently items that exist together (frequent itemsets) and find any patterns amongst these is needed. (Agrawal and Srikant, 1994) present an algorithm (Apriori) that aims at discovering association rules between items in a large database of sales transactions. The algorithm is simple in concept and is split into two main sub problems:

1. Find all sets of items (itemsets) that have transaction support above minimum support. The support for an itemset is the number of transactions that contain the itemset. Itemsets with minimum support are called large itemsets, and all others small itemsets.

2. Use the large itemsets to generate the desired rules.

The minimum support and confidence are given as follows (Bhargavi et al., 2013):

Support:

$$supp(X) = \frac{no.of\ transactions\ which\ contain\ the\ itemset\ X}{total\ no.of\ transactions}$$

Confidence:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

2.4 Association Rule Mining in Healthcare.

In health informatics, a lot of work has gone in the use of data mining to previously commercial-only applications. Key amongst the uses has been in matching patient diagnosis with symptoms which intertwines a lot with the use of knowledge based systems. It is difficult to induce reliable diagnostic rules from amongst a set of possibly infinite permutations of symptoms since the resulting hypotheses may have unsatisfactory prediction accuracy (Rajak and Gupta, 2008).

However, other researchers have come up with further refinements by using association rules to improve the prediction level claimed at 90% by (Serban et al., 2006) by combining it with supervised learning methods. The researchers applied their work to cancer but they claim that this can be extended to other disease diagnosis.

Association analysis as it is also called has been used to give probabilistic statements such as “If patients undergo treatment A, there is a 0.35 probability that they will exhibit symptom Z” (Koh and Tan, 2011). These can be useful when establishing relationships that affect effectiveness of particular patient treatment plans.

2.4.0 Related and Specific Applications in EMR Implementations.

In this research’s specific field, some work has been done to take advantage of association rule mining in general and the Apriori algorithm in particular. Most of it centers on mining patterns in relation to a specific disease or diagnostic factor.

(Tai and Chiu, 2009) used association rule mining to discover associations from data obtained from the National Health Insurance Database of Taiwan. Their work was intended not only to discover the comorbidity patterns of Attention Deficiency Hyperactive Disorder (ADHD) but to also examine the application of association rule mining in clinical databases.

The database used ICD-9 diagnosis coding and drew a sample of about 18,000 patients aged 18 and below with a diagnosis of ADHD. The researchers then made comparisons using Apriori algorithm to check the strength of associations amongst comorbidity rates and relative risk (RR) ratios of both groups of each diagnosis which were compared to one another. The results were published along with the resultant levels of interestingness.

More work was also done (Kim et al., 2012) to analyze comorbidity in patients with type 2 diabetes mellitus (T2DM). The data was obtained from a medical center in Korea with an EMR that uses ICD-10 coding for the clinical diagnosis. The researchers developed a tool that uses Apriori algorithm to generate the strongest rules (diagnosis) that are associated with the T2DM. They then published the results of their findings with the resultant support and confidence levels.

Another prototype namely Clinical State Correlation Prediction (CSCP) was developed in order to predict the correlation(s) amongst the primary disease (the disease for which the patient visits the doctor) and secondary disease/s (which is/are other associated disease/s carried by the same patient having the primary disease (Rashid et al., 2014). The system developed uses the Apriori algorithm as well and checks the correlation between the primary disease and other secondary diseases. The CSCP is built on top of the transaction based health system which they base on and refer to as the OLTP. The diagnoses are not based on any diagnosis group like ICD.

They also use data from this health OLTP, and pass the algorithm over data selected for different age groups and sex. The results of the top two-item itemsets are then analyzed for any meaningful information.

2.5 Conceptual Framework

The research and the arising system prototype to be implemented fall within a wider health organization consisting of different components such as:

- Care delivery Organizations (CDO) - hospitals, dispensaries, other clinics, Radiology departments.
- Electronic prescription from pharmacies.
- Pathology units.

We are mainly interested in the data arising from Care Delivery Organizations.

We will implement ARM analysis on the EMR data that adheres to the specified standards, and from this, get rules that give us associations based on user specified measures of interestingness.

The framework we conceptualize our research is the Performance of Routine Information Systems Management (PRISM) framework. It identifies how technical, behavioral, and organizational determinant(s) relate and interact to determine Routine health Information Systems (RHIS) performance. Figure 2.3 spells out these factors.

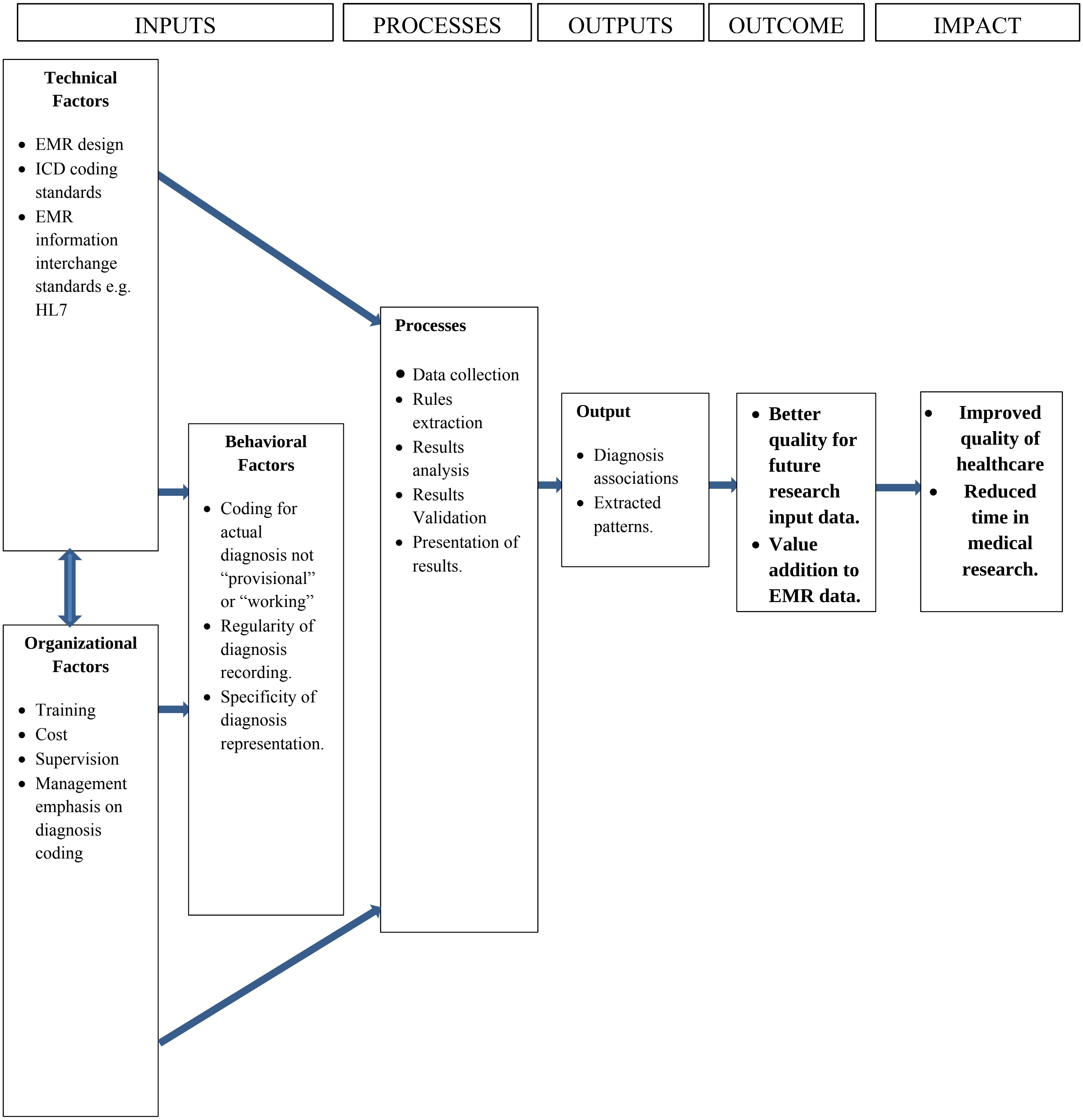

INPUTS
PROCESSES
OUTPUTS
OUTCOME
IMPACT
Technical Factors
EMR design
ICD coding standards
EMR information interchange standards e.g. HL7
Behavioral Factors
Coding for actual diagnosis not "provisional" or "working"
Regularity of diagnosis recording.
Specificity of diagnosis representation.
Organizational Factors
Training
Cost
Supervision
Management emphasis on diagnosis coding
Processes
Data collection
Rules extraction
Results analysis
Results Validation
Presentation of results.
Output
Diagnosis associations
Extracted patterns.
Better quality for future research input data.
Value addition to EMR data.
Improved quality of healthcare
Reduced time in medical research.

Figure 2.3 Conceptual framework

2.6 Limitations and Research gaps.

The research papers we have reviewed in this chapter thus far are some that show the extent to which data mining in general and association rule mining in particular has been used to discover interesting patterns in health systems.

Most advances in data mining and health systems research is based on diagnosis prediction systems that try to map symptoms with as small a differential diagnosis list as possible.

There are several other research works that follow similar approaches where either the same Apriori or another algorithm is used to mine associations between already known diseases with the intent of knowing which other conditions are most associated with those in question.

In the case of discovering previously unknown or thought disease patterns as in the work of (Rashid et al., 2014), they do not use a standardized EMR and the results are captured in the health system by free-text entered diagnosis which do not seem to adhere to any standard coding practice as presented in both their methodology and their results. This makes it difficult to statistically analyze or potentially group broader categories of diagnosis in order to get a wide variety of actionable rules/diagnosis.

This research aims at filling in these gaps by discovering associations from EMR systems that are built on this standard model and exploit the presence of data that is generated by multiple health providers that use EMRs governed by the same standards.

CHAPTER THREE

METHODOLOGY

Computer science applies several research methods over and above the traditional quantitative and qualitative methods used in other disciplines like social sciences. (Glass et al., 2004), define over nineteen methods applied in the related fields of Computer Science, information Systems and Software Engineering. They conduct for each discipline, an analysis of the most dominant research topics, research approaches, research methods, reference disciplines, and levels of analysis. In previous work by (Ramesh et al., 2004), formulative research was found to be the most widely used research approach in computer science disciplines (79.15%) as compared to descriptive and evaluative research approaches. Among the top three research methods were Conceptual analysis (15.13%), Conceptual analysis/mathematical (74.13%) and concept implementation (proof of concept) (2.87%).

In building on this work, (Holz et al., 2006) mention concept implementation (also proof of concept or proof of principle) as "a claim about the value of a system design (or the design of a part of a system) is validated by building a system based on that design. Typically, the system that is built is not fully featured, but has enough functionality to convince the readers that the design can be effective. The proof-of-concept system is usually measured for performance or usability, to show that the new design is not so bad as to be unworkable".

It is on this method that this research wants to align its work, by concept implementation and present a prototype that uses the association rule mining algorithm on a standardized EMR implementation.

3.0 Research Design

The research is structured as follows:

1. Problem Identification and Selection.
2. Literature Review and Concept Development.
3. Data Collection, Preparation and Processing.
4. Prototype Development.
5. Prototype Testing and Implementation.
6. Analysis, Validation and Presentation of Results.

3.0.0. Problem Identification and Selection.

In this stage, an attempt is made to select and explain the problem that the research intends to solve which in our case is taking advantage of standardized EMRs and advances in the data mining field. It is placed in perspective of the more general problem and there is an explanation of why it is a problem in the context of the research. There is also a brief description of how the researcher intends to approach the same and expected benefits that would accrue in tackling the issue at hand.

3.0.1 Literature Review and Concept Development.

Here we attempt to understand the advances of data mining technology, in particular the use of ARM. We explore the purposes that this has been traditionally used and originally conceived.

We also go ahead to look at the specific Apriori algorithm that is to be used for this research and we look at a few of the closest related work around the healthcare industry in general and in electronic health systems in particular. We develop the concept of ARM use in EMR systems and the several requirements in standards like ICD coding for diagnosis and HL7 for information exchange across the EMR systems that would be necessary to make the collection of data and analysis for this research possible.

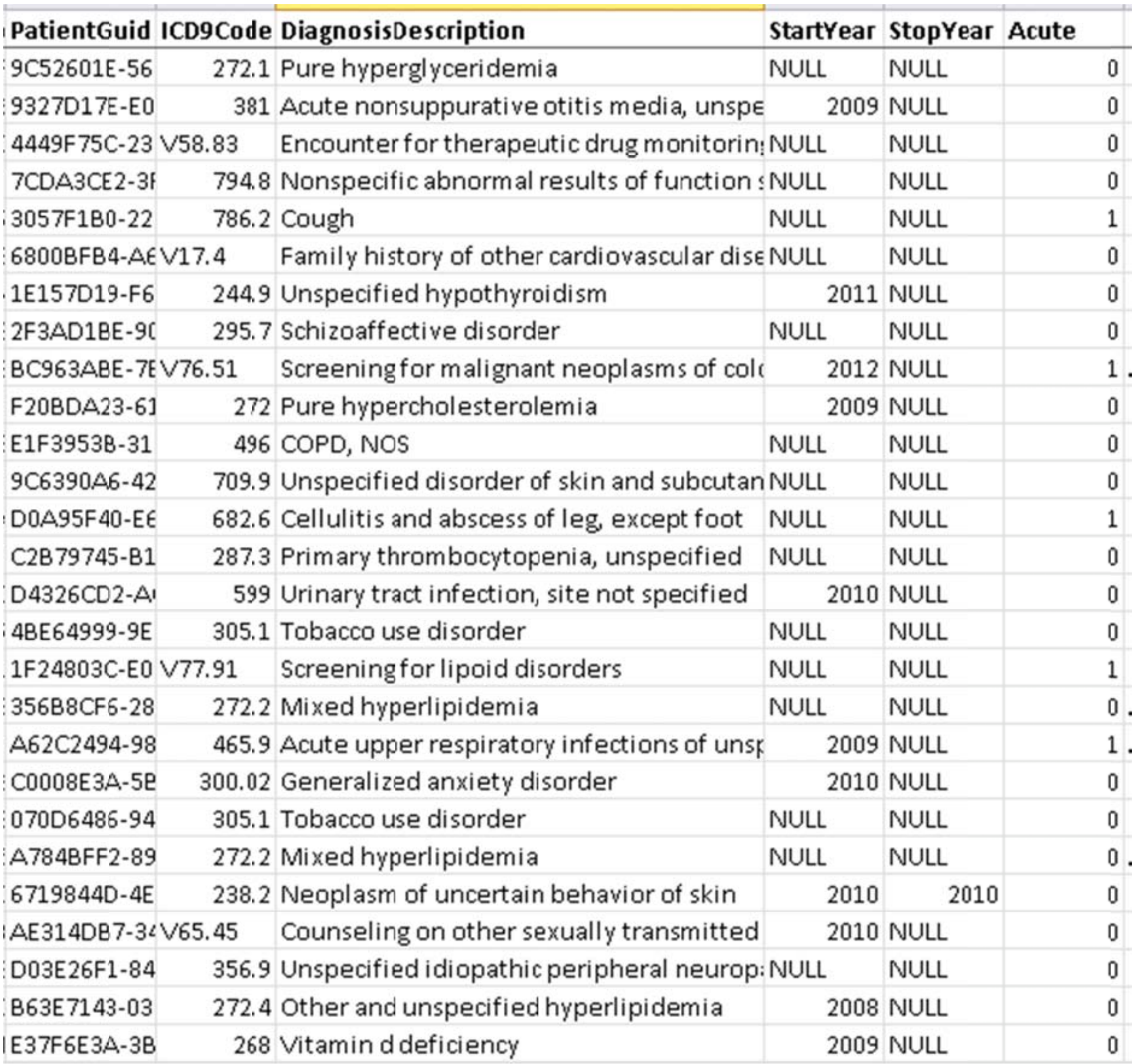
3.0.2 Data Collection, Preparation and Processing.

3.0.2.0 Data Collection

Here, we obtain data that is necessary for this research. The data had met the standards defined earlier in order for it to be usable. For this research, we obtained our data from Practice Fusion research data that is one of the leading EMR implementers in the United States of America. The dataset availed contains over 10,000 de identified patient records ("Analyze This! | Research Division," 2012) that contain over 98,000 individual contact points from over 150,000 medical practitioners across the country, in ICD-9 diagnosis codes. The data obtained was in the format shown in Figure 3.1.

The larger dataset from which they extracted this contains over 30 million records and has already been used to track the spread of H1N1 to help physicians obtain vaccines, and Practice Fusion's Research Division is partnering with leading academic institutions and public health agencies to pursue ambitious new health studies ("Big Data Gets Put to Work for Public Health," 2012).

We were unable to obtain Kenyan data from Kenyan EMR's as attempted through the International Training and Education Center for Health (I-TECH). This was so as one of the factors that have been key to generating the associations has been the consistency and adherence to the set out standards of medical diagnosis, which was difficult to find in the Kenyan context, whose practitioners tend to use the traditional free-text for their disease diagnosis. This opens up the possibilities of spelling inaccuracies, almost infinite permutations of how conditions can be described by each practitioner, amongst others.



PatientGuid	ICD9Code	DiagnosisDescription	StartYear	StopYear	Acute
9C52601E-56	272.1	Pure hyperglyceridemia	NULL	NULL	0
9327D17E-E0	381	Acute nonsuppurative otitis media, unspe	2009	NULL	0
4449F75C-23	V58.83	Encounter for therapeutic drug monitorin	NULL	NULL	0
7CDA3CE2-3F	794.8	Nonspecific abnormal results of function s	NULL	NULL	0
3057F1B0-22	786.2	Cough	NULL	NULL	1
6800BFB4-A6	V17.4	Family history of other cardiovascular dise	NULL	NULL	0
1E157D19-F6	244.9	Unspecified hypothyroidism	2011	NULL	0
2F3AD1BE-90	295.7	Schizoaffective disorder	NULL	NULL	0
BC963ABE-7E	V76.51	Screening for malignant neoplasms of colc	2012	NULL	1
F20BDA23-61	272	Pure hypercholesterolemia	2009	NULL	0
E1F3953B-31	496	COPD, NOS	NULL	NULL	0
9C6390A6-42	709.9	Unspecified disorder of skin and subcutan	NULL	NULL	0
D0A95F40-E6	682.6	Cellulitis and abscess of leg, except foot	NULL	NULL	1
C2B79745-B1	287.3	Primary thrombocytopenia, unspecified	NULL	NULL	0
D4326CD2-A	599	Urinary tract infection, site not specified	2010	NULL	0
4BE64999-9E	305.1	Tobacco use disorder	NULL	NULL	0
1F24803C-E0	V77.91	Screening for lipoid disorders	NULL	NULL	1
356B8CF6-28	272.2	Mixed hyperlipidemia	NULL	NULL	0
A62C2494-98	465.9	Acute upper respiratory infections of unsp	2009	NULL	1
C0008E3A-5B	300.02	Generalized anxiety disorder	2010	NULL	0
070D6486-94	305.1	Tobacco use disorder	NULL	NULL	0
A784BFF2-89	272.2	Mixed hyperlipidemia	NULL	NULL	0
6719844D-4E	238.2	Neoplasm of uncertain behavior of skin	2010	2010	0
AE314DB7-34	V65.45	Counseling on other sexually transmitted	2010	NULL	0
D03E26F1-84	356.9	Unspecified idiopathic peripheral neurop	NULL	NULL	0
B63E7143-03	272.4	Other and unspecified hyperlipidemia	2008	NULL	0
E37F6E3A-3B	268	Vitamin d deficiency	2009	NULL	0

Figure 3.1 Section of the raw dataset

3.0.2.1 Data Preparation & Processing

The data we obtained was then prepared and processed by the following steps:

i. Extracting the major diagnostic groups for each ICD-9 diagnosis for every patient record. This is due to the fact that every sub-diagnosis group after the period (.) still represents the anatomic site or severity of the specific disease category. For example the ICD-9 code 473 represents “Chronic Sinusitis”. Others under this would be 473.1 – “Chronic Sinusitis- Maxillary Antritis (chronic)”, 473.1 “Chronic Sinusitis Frontal”, 473.2 “Chronic Sinusitis – Ethmoidal” all the way to 473.9 - “Unspecified Chronic Sinusitis”.
ii. We are therefore still able to obtain interesting associations on a higher level without losing meaning from where we can dig further.
iii. We then filtered out the codes that begin with E & V since these represent External causes of injury (e.g. accidents) and Supplementary classification of factors influencing health status and contact with health services respectively (CMS, 2014)
iv. After this, we transform the data into itemsets where each patient has all the associated diagnosis combined into a single comma separated record. We have done this by use of an SQL script (Appendix A). A sample record would look like: (PID0001 | 420,421,618).
v. We are able to filter from the beginning if we want to observe chronic diagnosis only, acute only or all combined to give us the potential patterns we want.

3.0.3 Prototype Development

We have developed a prototype that implements the Apriori algorithm. The prototype borrows some implementations that have been used in market basket analysis. The prototype is capable of taking the data and finding associations based on the user defined values for the minimum support and confidence level. It was implemented in C# and relies on a backend database of MS-SQL server. The IDE used for this was Visual Studio 2012.

The system prototype developed takes an execution path as outlined in Appendix B (code map) with the main nodes.

This follows the classical Apriori algorithm with the steps as explained above in section 2.3.

3.0.4 Prototype Testing and Implementation

The prototype was developed and tested using the de-identified patient records. We were able to run the data and observe the association rules generated. Using varying minimum support and confidence values generated a number of rules. The top rules were those with the strongest confidence level above a support threshold.

Since there is no globally accepted minimum support (as this is a custom user generated variable that depends on what they want to achieve, and how far deep they want to dig into the associations), we varied these values to observe the results and recorded each observation. Just as in other works using the values of support and confidence in this mining for strong associations like in (Kim et al., 2012) and that of (Tai and Chiu, 2009), we vary the same measures and indicate the values of support and confidence for each rule.

3.0.5 Analysis, Validation and Presentation of Results.

Based on the rules observed, we compare this with the demographic data and select the demographic distribution of the top associations. These are mainly age groups and gender.

We then use measures of central tendency (as appropriate for the nominal and ordinal variables) and classify the data into the different categories that they fall in.

We also used a panel of experts drawn from the medical field who gave their opinion over the results. The survey was done using the questionnaire attached as Appendix C. The confidence levels and support for each question was left out deliberately so as to avoid user bias while answering the questions.

We used the Likert scale to gather expert opinion and listened to their overall advice while noting explanations to some of their responses.

CHAPTER IV

SYSTEM ANALYSIS AND DESIGN

4.0 Introduction

In this chapter, we describe the system user interface design, architecture and backend database design that have been used to structure the entire system prototype.

Since this is a prototype and not a fully functioning system, areas such as system security (logging in and out) and complex data entry validation have not been implemented.

We hope to illustrate the implementation of ARM and particularly using the Apriori algorithm in achieving our objectives. For an illustration of the research context see Figure 4.0.

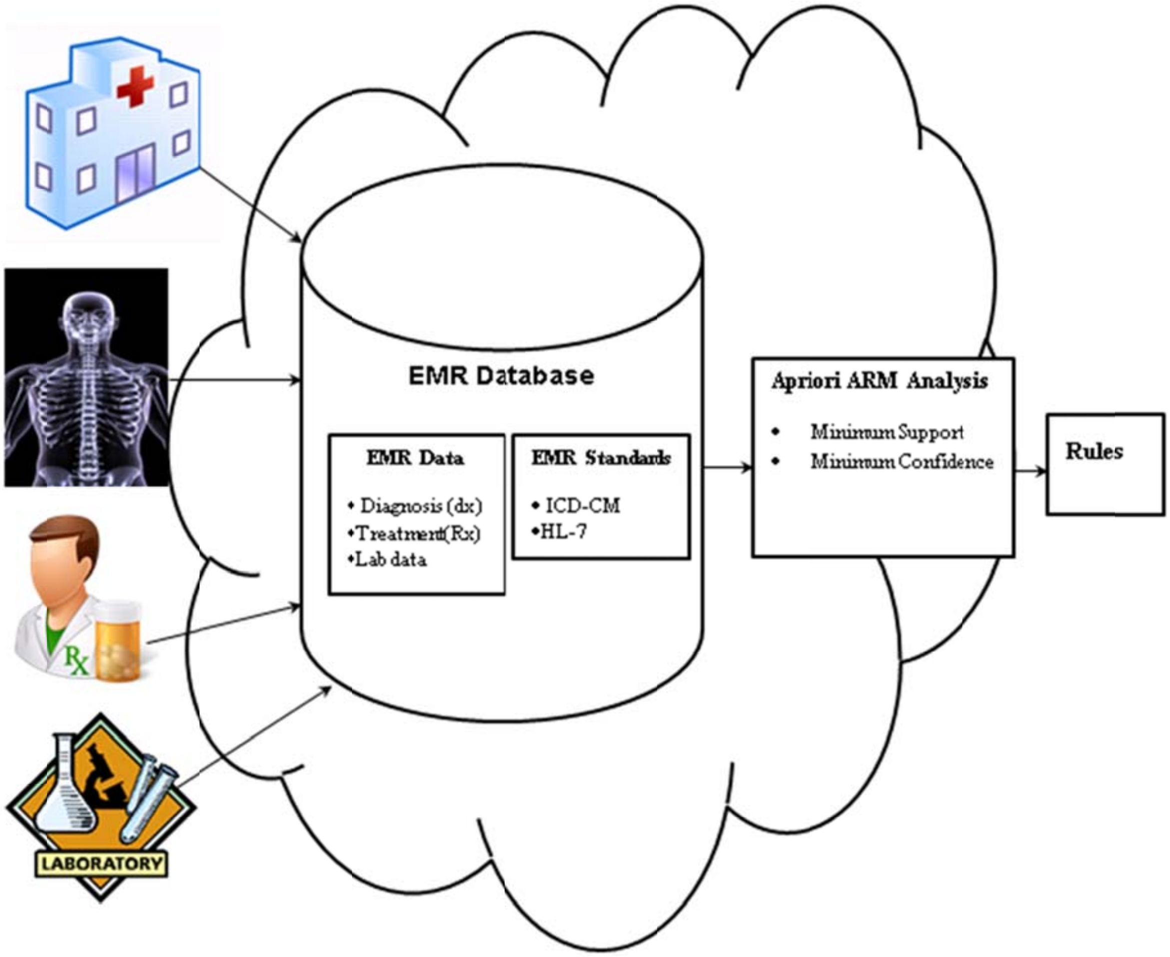


Figure 4.0 Illustration of research context

4.1.0 Key System Prototype Elements

The system consists of these three main constituent elements in order to accomplish its functions.

i) Input

This is the starting point of the system and is provided by the user. In our case, this consists of the input variables of minimum support and the minimum confidence threshold.

The minimum support determines the number of candidate itemsets from which the rules will be generated and ordered by the confidence variable.

ii) Transaction data

This is the actual data that the system will process. The data processing stage earlier on will have transformed the multiple patient clinic visits into individual records for each patient and their entire visits' diagnosis code in one delimited column.

iii) Output

The output of our system will be the strongest rules that are based on the most frequent itemsets and that satisfy the minimum confidence level.

4.1.1 System Design & Architecture

The prototype consists of three major components.

i. Database Design
ii. Logic
iii. User Interface.

i.) Database Design

The database consists of three tables.

- Patient demographics table
- Clinic visits table
- Transaction Table

The design of each and their relationships are as show below.

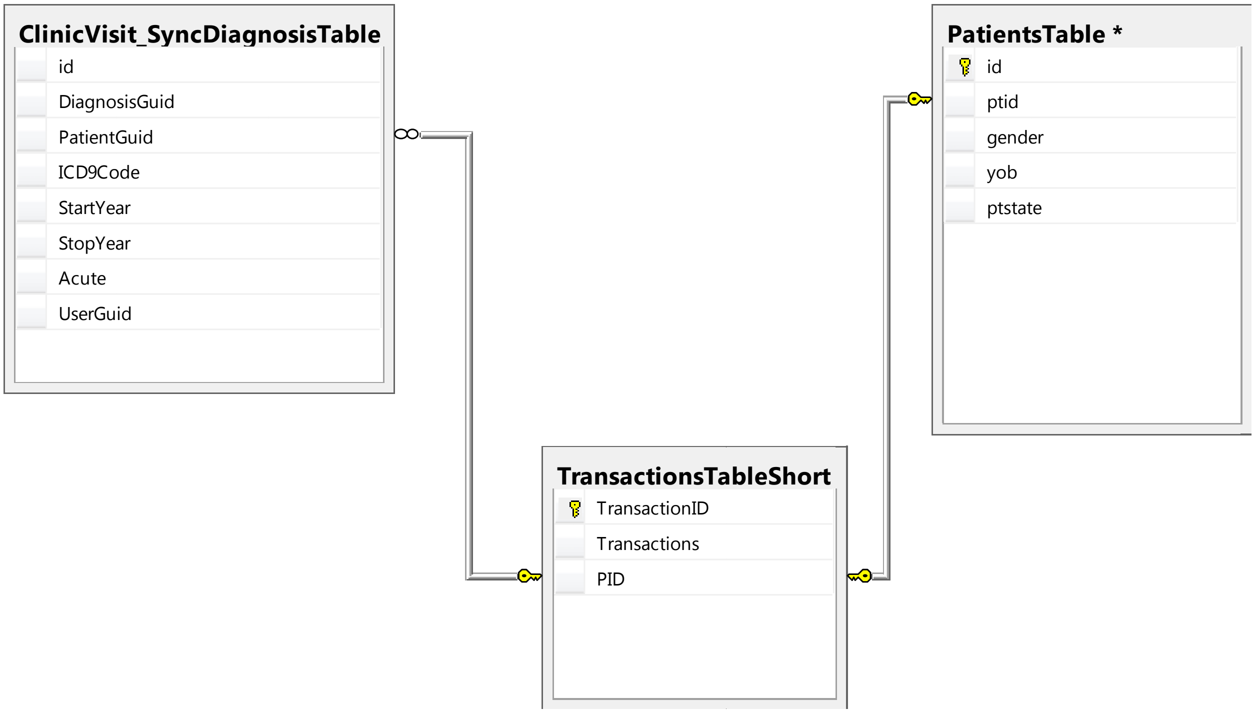
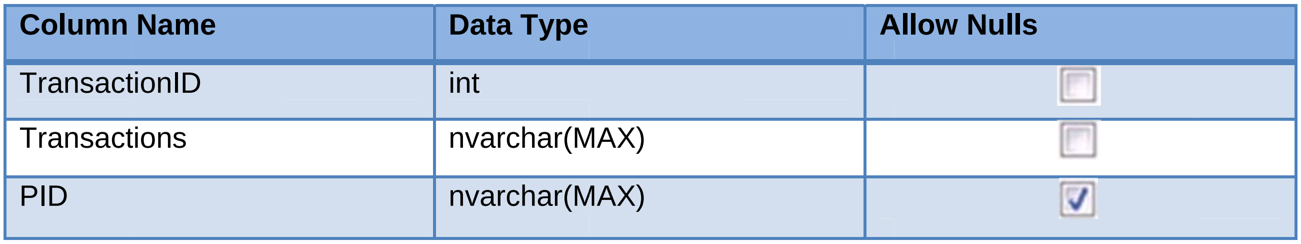


Figure 4.1 Database Schema



Column Name	Data Type	Allow Nulls
TransactionID	int	□
Transactions	nvarchar(MAX)	□
PID	nvarchar(MAX)	☑

Figure 4.2 Transactions Table

Column Name	Data Type	Allow Nulls
id	int	☐
ptid	nvarchar(100)	☑
gender	nchar(1)	☑
yob	nvarchar(4)	☑

Figure 4.3 Patient Bio data Table

Column Name	Data Type	Allow Nulls
id	int	☐
DiagnosisGUID	varchar(50)	☑
PatientGUID	varchar(50)	☑
ICD9code	varchar(50)	☑
StartYear	varchar(50)	☑
StopYear	varchar(50)	☑
Acute	varchar(50)	☑
UserGUID	varchar(50)	☑

Figure 4.4 Clinic Visits Table

ii) Logic

This is the most important component as it implements the algorithm. Using the input entered by the system we follow these steps to obtain the associations:

1. Find all sets of items (itemsets) that have transaction support above minimum support. The support for an itemset is the number of transactions that contain the itemset. Itemsets with minimum support are called large itemsets, and all others small itemsets.

If for example we define our support as 40% we select those diagnoses that have met this threshold e.g.

(240) (272) (530)

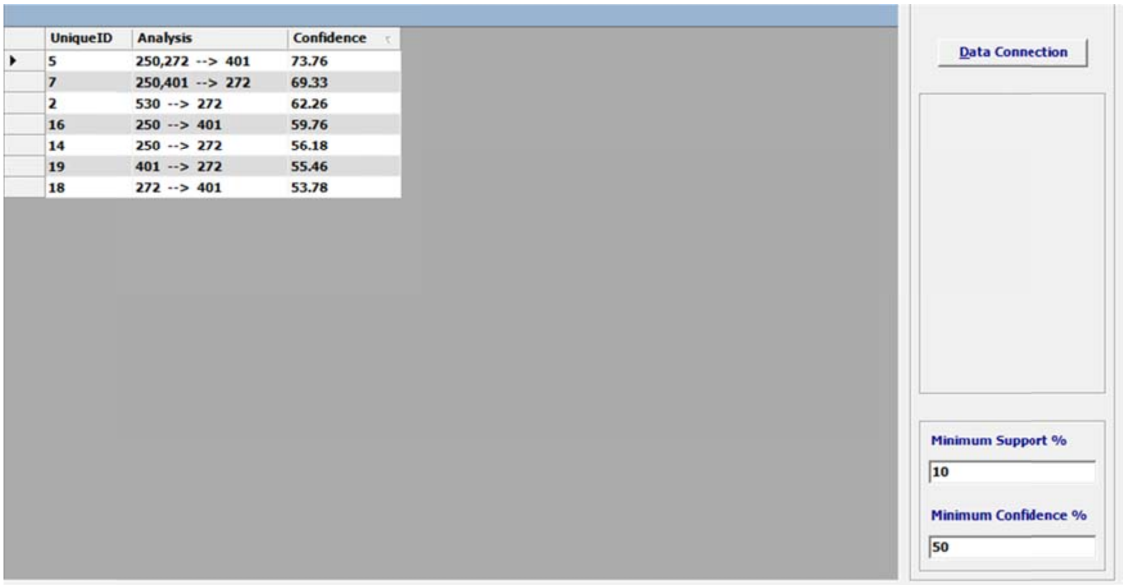
This would mean that these (level 1) itemsets meet the defined minimum support.

In the first iteration of the algorithm, each item is a member of the set of candidate. The set of frequent 1-itemsets, L1 , consists of the candidate 1-itemsets satisfying minimum support.

We keep joining the newly formed itemsets with themselves and use the Apriori Property or the downward-closure property according to (Agrawal et al., 1993), that all subsets of a frequent itemset must also be frequent. Thus obtaining itemsets such as (272,240) (401,272)

We iterate until there is no itemset meeting our minimum support threshold.

2. These large itemsets are then used to calculate the strongest rules now based on the user provided value for minimum confidence, e.g. (272--> 401) see fig 4.5 below.



UniqueID	Analysis	Confidence
5	250,272 --> 401	73.76
7	250,401 --> 272	69.33
2	530 --> 272	62.26
16	250 --> 401	59.76
14	250 --> 272	56.18
19	401 --> 272	55.46
18	272 --> 401	53.78

Minimum Support %: 10

Minimum Confidence %: 50

Figure 4.5 Resultant Strong Rules Screen

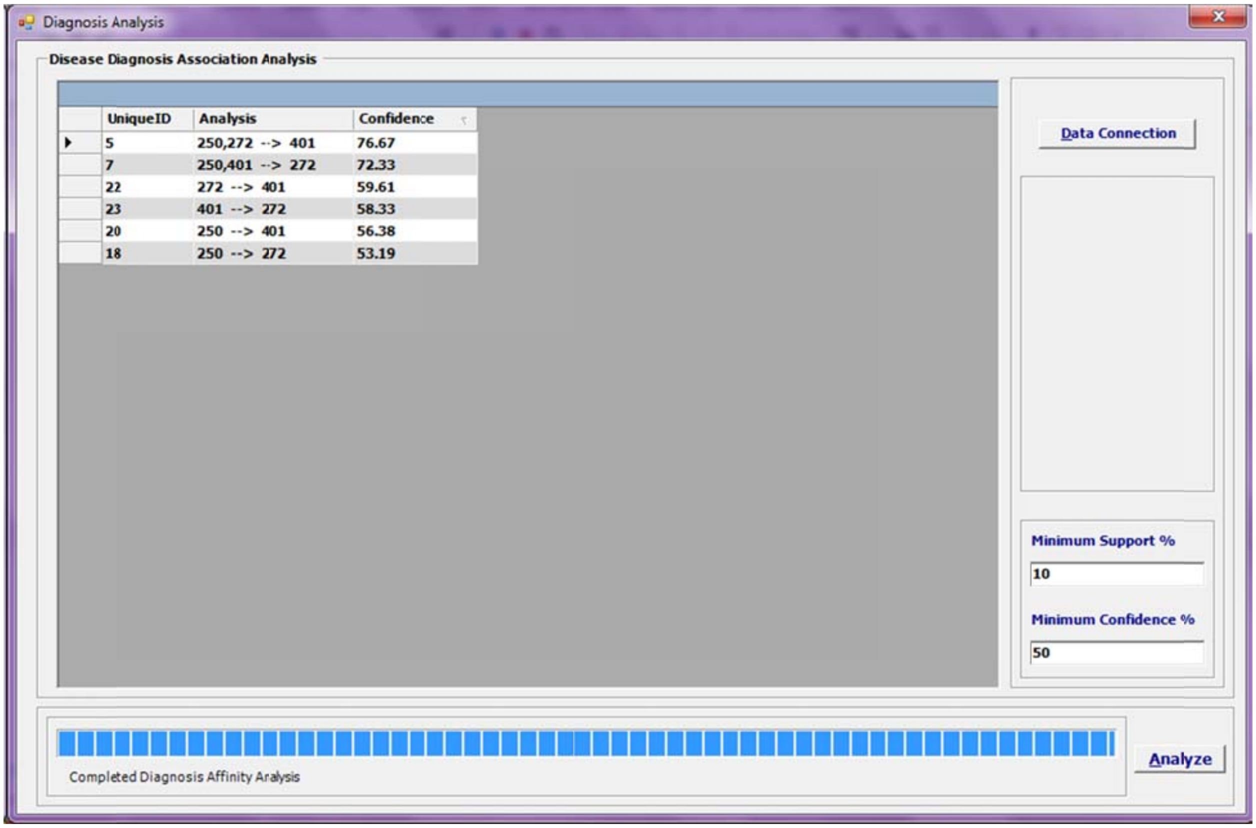
The execution map for this code is attached in Appendix B.

iii) User Interface

This is where the user interacts with the system. They are able to provide the values of minimum support and minimum confidence and with this several associations can be obtained.

The user is then presented with a data grid with the results along with their confidence values.

A sample screen is shown below:



UniqueID	Analysis	Confidence
5	250,272 --> 401	76.67
7	250,401 --> 272	72.33
22	272 --> 401	59.61
23	401 --> 272	58.33
20	250 --> 401	56.38
18	250 --> 272	53.19

Minimum Support %: 10

Minimum Confidence %: 50

Completed Diagnosis Affinity Analysis

Figure 4.6 Sample User Interface Screen

4.1.2 Execution of Runs

The system could be ran in batch for all the transactions. However for the 9,971 records, it takes around 1740 seconds to complete, and increasing the chances of failure.

However, with the reduction in the size of dataset size and with the use of execution runs, to 1/10 for each dataset, we are able to take approximately 180 seconds for each run.

We are then able to aggregate these results and come up with the figures that are discussed in the next chapter.

CHAPTER V

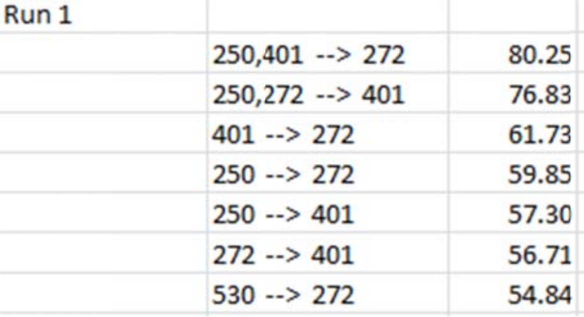
RESULTS AND DISCUSSION

Here we are now able to present the results of the research within the context of the study.

After execution of the runs and aggregation of the same, we were able to come up with a number of rules based on a support factor of 10% and a minimum confidence of 50% for the first instance in order to see if we would obtain less than 10 rules, from which we expected to see rules already known by anyone in the medical field, also known as trivial rules (Berry and Linoff, 2004, pp. 296–298).

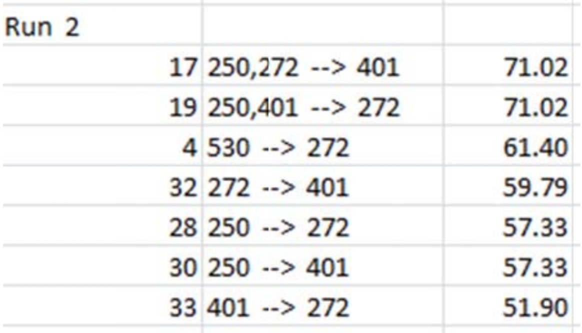
5.0 Potentially Trivial Associations

The results of each run are shown below. In each are the top associations and the associated confidence levels.

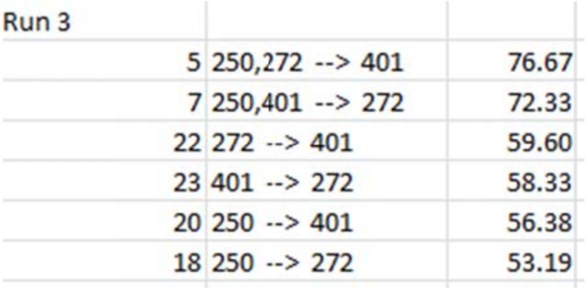


Run 1		
	250,401 --> 272	80.25
	250,272 --> 401	76.83
	401 --> 272	61.73
	250 --> 272	59.85
	250 --> 401	57.30
	272 --> 401	56.71
	530 --> 272	54.84

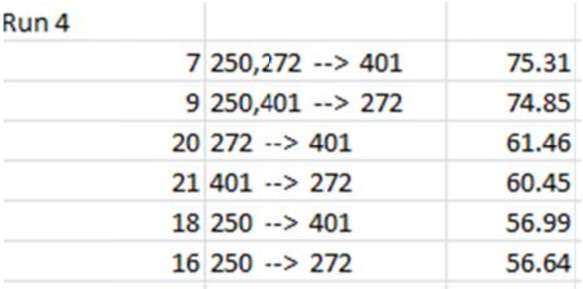
Run 1 (Code)	Description
250,401 --> 272	Diabetes Mellitus, Essential Hypertension-> Disorders of Lipoid metabolism
250,272 --> 401	Diabetes Mellitus, Disorders of Lipoid metabolism-> Essential Hypertension
401 --> 272	Essential Hypertension->Disorders of Lipoid metabolism
250 --> 272	Diabetes Mellitus->Disorders of Lipoid metabolism
250 --> 401	Diabetes Mellitus->Essential Hypertension
272 --> 401	Disorders of Lipoid metabolism->Essential Hypertension
530 --> 272	Esophageal disease->Disorders of Lipoid metabolism
530 --> 401	Esophageal disease->Essential Hypertension



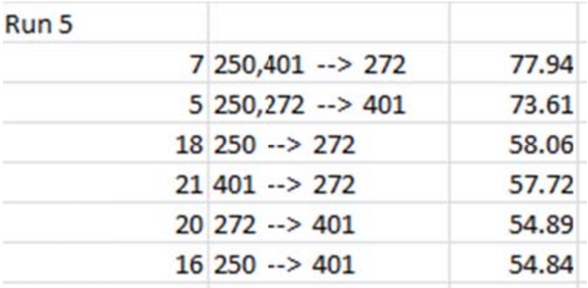
Run 2		
17	250,272 --> 401	71.02
19	250,401 --> 272	71.02
4	530 --> 272	61.40
32	272 --> 401	59.79
28	250 --> 272	57.33
30	250 --> 401	57.33
33	401 --> 272	51.90



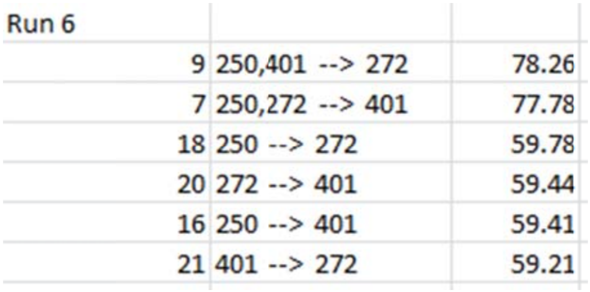
Run 3		
5	250,272 --> 401	76.67
7	250,401 --> 272	72.33
22	272 --> 401	59.60
23	401 --> 272	58.33
20	250 --> 401	56.38
18	250 --> 272	53.19



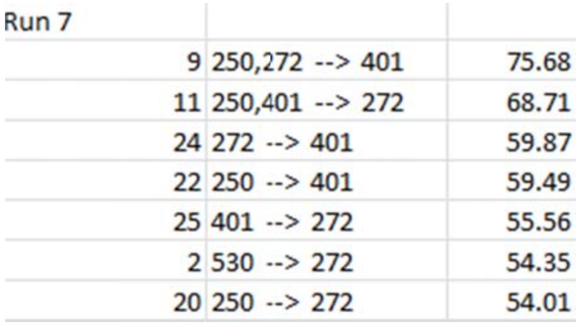
Run 4		
7	250,272 --> 401	75.31
9	250,401 --> 272	74.85
20	272 --> 401	61.46
21	401 --> 272	60.45
18	250 --> 401	56.99
16	250 --> 272	56.64



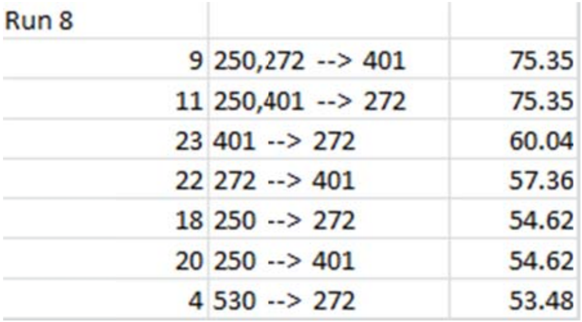
Run 5		
7	250,401 --> 272	77.94
5	250,272 --> 401	73.61
18	250 --> 272	58.06
21	401 --> 272	57.72
20	272 --> 401	54.89
16	250 --> 401	54.84



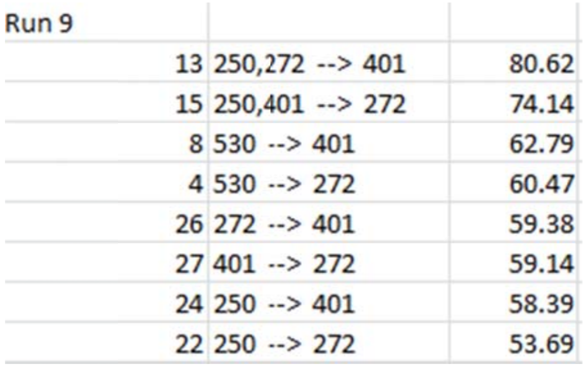
Run 6		
9	250,401 --> 272	78.26
7	250,272 --> 401	77.78
18	250 --> 272	59.78
20	272 --> 401	59.44
16	250 --> 401	59.41
21	401 --> 272	59.21



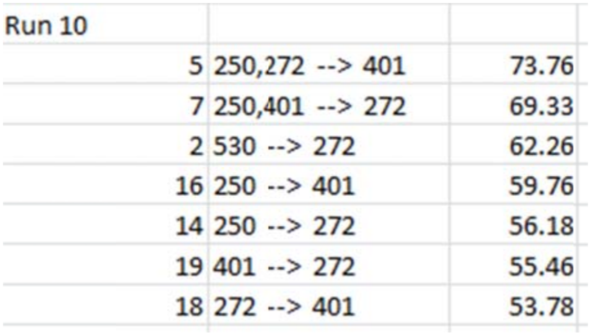
Run 7		
9	250,272 --> 401	75.68
11	250,401 --> 272	68.71
24	272 --> 401	59.87
22	250 --> 401	59.49
25	401 --> 272	55.56
2	530 --> 272	54.35
20	250 --> 272	54.01



Run 8		
9	250,272 --> 401	75.35
11	250,401 --> 272	75.35
23	401 --> 272	60.04
22	272 --> 401	57.36
18	250 --> 272	54.62
20	250 --> 401	54.62
4	530 --> 272	53.48



Run 9		
13	250,272 --> 401	80.62
15	250,401 --> 272	74.14
8	530 --> 401	62.79
4	530 --> 272	60.47
26	272 --> 401	59.38
27	401 --> 272	59.14
24	250 --> 401	58.39
22	250 --> 272	53.69



Run 10		
5	250,272 --> 401	73.76
7	250,401 --> 272	69.33
2	530 --> 272	62.26
16	250 --> 401	59.76
14	250 --> 272	56.18
19	401 --> 272	55.46
18	272 --> 401	53.78

Figure 5.1 All Runs

These aggregate for each was obtained and are listed in fig 5.2 below:

ICD9 Disease Description	Average confidence
Diabetes Mellitus, Essential Hypertension-> Disorders of Lipoid metabolism	74.22
Diabetes Mellitus, Disorders of Lipoid metabolism-> Essential Hypertension	75.66
Essential Hypertension->Disorders of Lipoid metabolism	57.95
Diabetes Mellitus->Disorders of Lipoid metabolism	56.34
Diabetes Mellitus->Essential Hypertension	57.45
Disorders of Lipoid metabolism->Essential Hypertension	58.23
Esophageal disease->Disorders of Lipoid metabolism	34.68
Esophageal disease->Essential Hypertension	6.28

Figure 5.2 Top Associations (note the bottom two as well)

There were two that met the minimum support but not the confidence levels (the bottom two in Fig 5.1) and in order to test if they were inexplicable rules, we allowed them to be in the survey and they were proved to be so as shown in the survey interpretation later in this chapter.

A graphical representation is also shown in Fig 5.3:

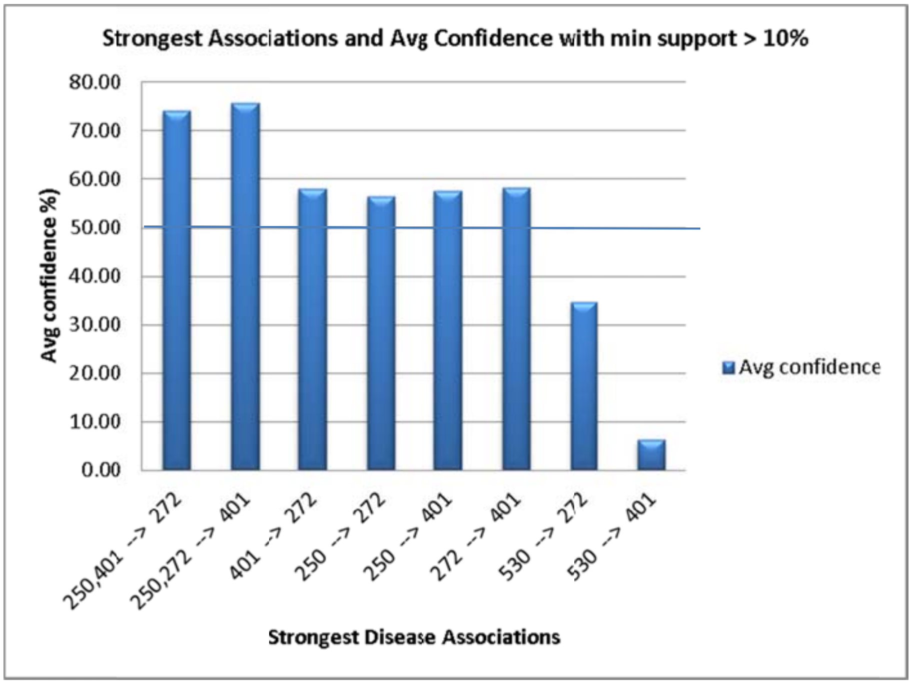
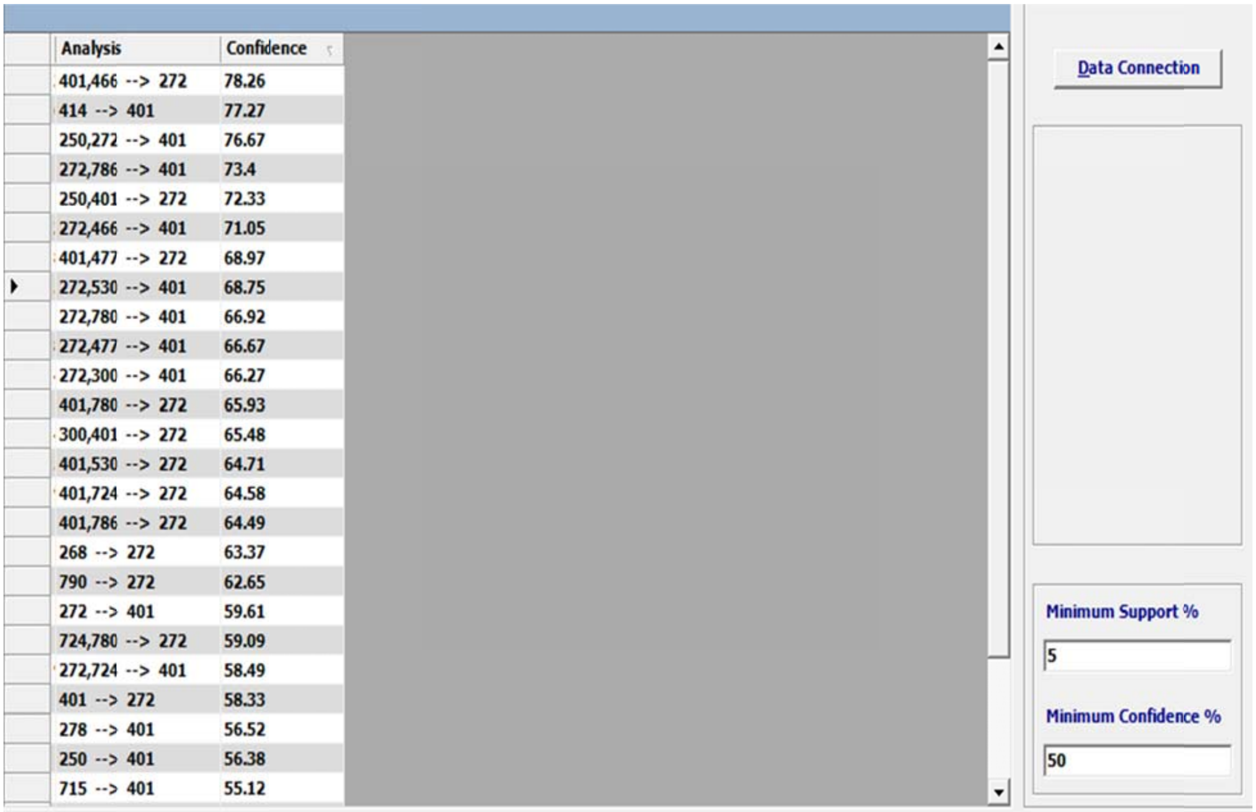


Figure 5.3 Strongest Associations with support above 10%

5.1 Potentially Actionable Rules

According to (Berry and Linoff, 2004, pp. 296–298) actionable rules as those that contain high-quality, actionable information. We intended to obtain these and us them in our comparisons.

Since the support of 10% only generated eight associations with a confidence level above 50%, reducing the minimum support to 5% but still maintaining a confidence level above 50% in order to avoid compromising on the quality of associations gave a number of interesting rules. They are listed in Figure 5.4:



Analysis	Confidence
401,466 --> 272	78.26
414 --> 401	77.27
250,272 --> 401	76.67
272,786 --> 401	73.4
250,401 --> 272	72.33
272,466 --> 401	71.05
401,477 --> 272	68.97
272,530 --> 401	68.75
272,780 --> 401	66.92
272,477 --> 401	66.67
272,300 --> 401	66.27
401,780 --> 272	65.93
300,401 --> 272	65.48
401,530 --> 272	64.71
401,724 --> 272	64.58
401,786 --> 272	64.49
268 --> 272	63.37
790 --> 272	62.65
272 --> 401	59.61
724,780 --> 272	59.09
272,724 --> 401	58.49
401 --> 272	58.33
278 --> 401	56.52
250 --> 401	56.38
715 --> 401	55.12

Minimum Support %

5

Minimum Confidence %

50

Figure 5.4 Other Association Codes

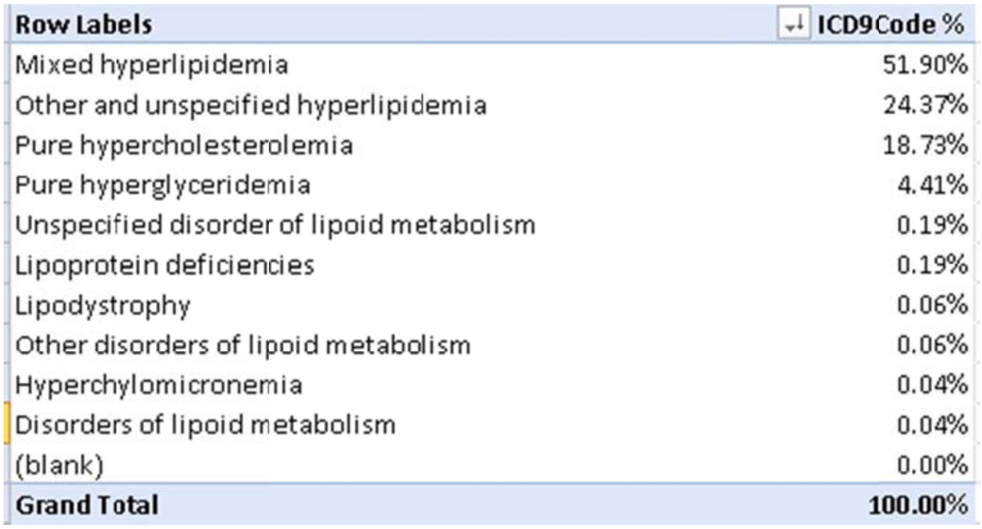
These translate to ICD descriptions as (Figure 5.5):

Essential hypertension,Allergic rhinitis --> Disorders of lipoid metabolism
Disorders of lipoid metabolism,Diseases of esophagus (excludes esophageal varices) --> Essential hypertension
Disorders of lipoid metabolism,Allergic rhinitis --> Essential hypertension
Disorders of lipoid metabolism,Anxiety, dissociative and somatoform disorders --> Essential hypertension
Anxiety, dissociative and somatoform disorders,Essential hypertension --> Disorders of lipoid metabolism
Essential hypertension,Diseases of esophagus (excludes esophageal varices) --> Disorders of lipoid metabolism
Essential hypertension,Other and unspecified disorders of back (see excludes) --> Disorders of lipoid metabolism
Essential hypertension,Symptoms involving respiratory system and other chest symptoms --> Disorders of lipoid metabolism
Vitamin D deficiency --> Disorders of lipoid metabolism
Nonspecific abnormal findings on examination of blood (see excludes 2) --> Disorders of lipoid metabolism
Disorders of lipoid metabolism --> Essential hypertension
Disorders of lipoid metabolism,Other and unspecified disorders of back (see excludes 1) --> Essential hypertension

Essential hypertension --> Disorders of lipoid metabolism
Overweight, obesity and other hyperalimentation --> Essential hypertension
Diabetes mellitus --> Essential hypertension
Osteoarthrosis and allied disorders --> Essential hypertension
Edema--> Essential hypertension
Excludes: 1. Excludes collapsed vertebra (code to cause, e.g., osteoporosis) conditions due to intervertebral disc disorders, spondylosis 2. Excludes abnormality of: platelets , thrombocytes, white blood cells

Figure 5.5 ICD 9 Descriptions

The diagnosis distribution amongst the group of “Disorders of lipoid metabolism” coded as 272 is shown below in Figure 5.6. It is important to note that “Mixed Hyperlipidemia” and “Other Unspecified hyperlipidemia” formed the bulk (76.27%) of the “Disorders of lipoid metabolism”



Row Labels	ICD9Code %
Mixed hyperlipidemia	51.90%
Other and unspecified hyperlipidemia	24.37%
Pure hypercholesterolemia	18.73%
Pure hyperglyceridemia	4.41%
Unspecified disorder of lipoid metabolism	0.19%
Lipoprotein deficiencies	0.19%
Lipodystrophy	0.06%
Other disorders of lipoid metabolism	0.06%
Hyperchylomicronemia	0.04%
Disorders of lipoid metabolism	0.04%
(blank)	0.00%
Grand Total	**100.00%**

Figure 5.6 Distribution for "Disorders of lipoid metabolism"

5.3 Demographic Comparison

We compared the diagnoses with the demographic prevalence where they were comorbid and these were indicated as shown in Appendix D. It is worth noting that the average age of the population was age 52.

5.4 Validation of Results

In this stage we are able to compare the results of our prototype and the opinion of experts regarding whether the associations obtained here are known to them or not, and if not whether they agree that they could be linked (probably indirectly) and by how much (strongly or otherwise). This is done through a questionnaire survey (see appendix C).

Each of the questions can be scored as follows:

Question	LK	HC	JM	DO	JA	BA	Median
1	5	4	5	4	5	3	4.5
2	5	3	5	4	5	2	4.5
3	5	4	5	4	5	4	4.5
4	5	4	5	4	5	4	4.5
5	5	4	5	5	5	3	5
6	5	4	5	4	5	4	4.5
7	3	1	1	2	1	2	1.5
8	3	1	1	2	1	2	1.5
9	3	1	3	-	-	2	2.5
10	5	4	5	1	5	4	4.5
11	5	3	5	5	5	3	5
12	3	4	4	1	4	4	4
13	5	4	5	4	5	3	4.5
14	3	3	3	1	-	2	3
15	3	1	3	2	-	3	3
16	3	2	3	1	-	2	2
17	3	2	3	2	-	2	2
18	4	2	3	2	4	3	3
19	4	2	3	1	4	3	3
20	3	2	3	1	-	3	3
21	2	2	3	1	4	3	2.5
22	2	2	4	1	4	4	3
23	1	1	4	1	2	5	1.5
24	1	2	2	1	-	2	2
25	5	3	5	4	5	4	4.5
26	3	3	3	2	4	4	3
27	5	4	5	4	5	2	4.5
28	4	5	5	4	5	5	5
29	5	3	5	4	5	2	4.5
30	3	3	-	3	2	2	3
31	4	2	4	2	5	2	3

Figure 5.7 Likert Scale Scores

Each of the questions has a score associated that is calculated from the median of responses from all experts (since the scale consists of ordinal values).

5.5 Discussion of Results

After running through the dataset, we were able to generate several associations that differed based on what we set as the minimum support and confidence level. We did not find a universally applicable or acceptable threshold for minimum support and confidence, as this seems to be applicable in different ways to different domains, depending on what patterns the end user intends to accept or reject. As earlier discussed, as in other works using the values of support and confidence in this mining for strong associations like in (Kim et al., 2012) and that of (Tai and Chiu, 2009), we vary the same measures and indicate the values of support and confidence for each rule.

It is possible to obtain a very large number of rules since these increase as the values of minimum support and minimum confidence are decreased and are approaching zero. The outliers in the data in this case will be the rules that may not necessarily meet the selected user-specific threshold for minimum support and confidence. It is therefore up to the user to decide what the most acceptable values for minimum support and confidence are, and what criteria to use to discard or accept the generated associations.

We observed that some rules were generated which happened to be consistent with common knowledge amongst the members of the medical fraternity, for example the link between Essential Hypertension and Disorders of the Lipoid Metabolism, or Diabetes Mellitus (as shown in the first six rules of Figure 5.4 and subsequent description in Figure 5.5). The panelists accepted this with a concurrence of 4.5/5 translating to a 90% nod. These known associations also had all high measures of confidence (between 56.34-75.66% from our system) as shown in Figure 5.2. Some diagnosis were also consistent with some of previous specific research like that of (Kim et al., 2012) that indicate the strongest link between Type 2 Diabetes mellitus and Essential Hypertension with a confidence of 34.86%. This is captured as rule 5 in our results with a confidence of 57.45%.

There are other rules which most of the panelists chose to neither agree nor disagree. They attributed this to the fact that some of the associations may be incidental to some specific patients and it may be observed in a few cases but not necessarily a majority of the cases. The presence of one qualifying diagnosis from amongst the set on the left being linked to that on the right also caused a mixed reaction in most of the practitioners, an example being that of:

Essential hypertension, Allergic rhinitis --> Disorders of lipoid metabolism.

In such a case, the panelists argued that it is the link of Essential hypertension to Disorders of lipoid metabolism and not the Allergic rhinitis that would trigger the association.

We also observe that some associations were out rightly rejected by the same panel of experts as expected (e.g. the association between *Esophageal disease->Disorders of Lipoid metabolism*). These were listed despite having very low confidence levels (as low as 6.28%) in order to compare and therefore validate the responses with the others that had higher values for confidence.

Some experts indicated that some of the associations could be comorbid but not necessarily linked, that is without a cause-effect relation and that some conditions coexist but are not very frequent. This, they said, also determined how they scaled the associations.

However, particularly with the second run where the minimum support was lowered to 5%, but the confidence level maintained, it is interesting to note that there was mixed opinion, or outright rejection of some interesting associations. An association that seemed interesting to the researcher that got a "Strongly Disagree" despite having a high confidence level (63.37%) is that of (*Vitamin D deficiency -> Disorders of lipoid metabolism*). As shown in Figure 5.6, more than 75% of this was hyperlipidemia (mixed and unspecified).

This result presents an interesting dimension as the experts indicated little or no known association between the two. In ensuing discussions over the results, one panelist noted that this association could have different indirect associations that could potentially explain it. Hyperlipidemia (explained to the researcher as a condition resulting from elevated levels of lipids in the blood) could have been as an indirect result of vitamin D deficiency since people who lack in vitamin D may be those that tend to stay indoors most of the time (one of the major sources of Vitamin D being skin exposure to sunlight). This could arguably be in line with the average age of 52 for the patients in the sampled dataset. The hyperlipidemia therefore in his reasoning, comes not as a result of the Vitamin D deficiency, but as a result of the lifestyle likely to be found amongst patients with Vitamin D deficiency. That relation alone as a real possibility could be subject to investigation outside the scope of this research.

Another panelist was also keen to indicate that the associations that we seek to investigate can only be investigated as comorbidity patterns and causal relations may not necessarily be possible to state comprehensively at this level. This is what the research emphasizes as the output of its findings.

Findings to mining medical datasets requires a lot of domain expertise to interpret the rules as was reiterated by (Roddick et al., 2003). Most of them will be known but others may be less known while those that seem unusual may be discarded at a first look. However, output to this research may prove to be of utmost importance to curious specialists since some of the rules generated, however few, could be used as a starting point for future research by the domain experts. Of great interest would be to attempt to establish the reasons for comorbidity amongst our associations that seem unusual or unknown, a good example being the *Vitamin D-> Hyperlipidemia* association. These reasons could be causal links or outright co-existence due to the condition of the patient. As one panelist explained, a patient diagnosed separately with allergic rhinitis, bronchitis and eczema (dermatitis) will have allergic tendencies that make such conditions, whereas unrelated, to be present in the same patient over time. When this happens frequently in the sampled population, some associations like these will certainly emerge from our system, and only further investigation by domain experts will show that.

Comparisons with demographics showed some expected patterns like some disease prevalence being higher in older patients e.g. the combination of Hypertension and Diabetes Mellitus being found in patients with an average age of 63.5, presenting a distance of 11.5 years above our average age. This is true for the most common diagnosis associations from our results. Further demographic analysis could be done on individual sets of associations as far as one would desire to find more relevant demographic patterns and compare them with the expected patterns.

CHAPTER VI

SUMMARY, CONCLUSION AND RECOMMENDATIONS

6.0 Summary

This research set out to identify any hidden diagnosis patterns within the big EMR data. We intended to find out the current use of association rule mining in both the commercial world in general and the health sector in particular. Of great interest, was the adaption to the apriori algorithm to mine the associations by prototype implementation.

We also intended to investigate the applicability of association rule mining in the context of electronic medical record systems that adhere to certain standards as well as show that we can generate and discover strong rules (relationships) that indicate multimorbidity trends from the EMR data with varying confidence levels.

It was also of interest to establish whether there would be any interesting or new trends between the age and gender factors to what is known currently.

6.1 Conclusion

Using this prototype, we are now able to mine data from EMR systems that implement any standardized diagnosis coding guideline. In our case, it is the WHO recommended standard of ICD-CM coding. Multiple systems can exchange their data and we are therefore able to take advantage of big data and generate patterns from it based on user defined measures of interestingness on what suits one as the minimum support and confidence.

It is also key to note that the data used for mining the associations was primarily intended for other clinical purposes. In this research, we were able to take advantage and build our system to find interesting patterns that could arise from this kind of well-organized big data. This goes to demonstrate the power of having standardized clinical data across multiple implementations of electronic medical records systems.

We were able to see that although the medical practitioners agreed on some already known associations, it would not be prudent to expect them to agree on all previously unknown associations. This research would therefore prove to be key as input to another research on causation, and would be a good starting point for any medical researcher seeking to look for multi-morbidity trends amongst patients in any given patient population.

We are particularly encouraged by previous studies that seemed to suggest that Vitamin D deficiency is associated with Hypertension but the causal relationship is not known (Vimaleswaran et al., 2014).

This is the same way in which there could be a (perhaps less prevalent but nonetheless unknown and important) relationship between Vitamin D deficiency and disorders of lipoid metabolism mostly hyperlipidemia (mixed and unspecified type).

This would ideally then be used as input to another study that seeks to dwell on the specific association and finding if there is any causal association.

The demographic prevalence of our associations showed no much difference with the expected outcome as discussed in the previous chapter.

6.2 Limitations and Challenges

There were several limitations of the research and challenges faced while carrying this research.

From the outset, obtaining the data used for this research was quite a challenge as getting EMR data is an issue that carries quite a lot of confidentiality and medical-legal challenges with it. Most hospitals would therefore be unwilling to release such data.

We also required data that adhered to our specified data in order to make data mining possible and this is currently a challenge in Kenya since the standards and guidelines for this have only been recently established. Not all major CDO’s have been using EMR systems which reduces even further the pool from which to obtain this data. We were thus unable to obtain Kenyan data within the duration of this research but should this become available in future, the same system could be applied to this data and hopefully interesting patterns will be obtained.

Interpreting results obtained from mining medical data also requires familiarity with diagnosis coding and expertise in the medical field. The researcher was fortunate enough to be working in a health organization where domain expertise was available. Although the practitioners tended to be very busy, we were able to schedule discussions and their feedback and guidance in some of the areas during the research proved to be quite useful.

6.3 Future Work

As stated in the previous chapter, we recommend that the output of this research particularly with results from the rules that had high confidence levels but lower support levels be investigated by another domain specific study to explain the comorbidity trends to those that are unknown to the medical fraternity.

In order to better the functioning of the system performance while handling large data, we recommend that another improvement of ARM algorithm be adopted to see if will perform better than the current system on the same data. If this works then it would speed up and even encourage the use of the same prototype implementation with even larger datasets that are more likely to yield further and more usable associations.

Work on Kenyan data with the same implementation would also be encouraged as soon as there is enough data also adhering to the specified standards in this research from which to attempt to generate these associations.

REFERENCES

About Health Level Seven International [WWW Document], 2007. URL https://www.hl7.org/about/index.cfm?ref=quicklinks (accessed 2.4.14).

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining Association Rules Between Sets of Items in Large Databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93. ACM, New York, NY, USA, pp. 207–216. doi:10.1145/170035.170072

Agrawal, R., Srikant, R., 1994. Fast Algorithms For Mining Association Rules In Datamining, in: 20th International Conference on Very Large Data Bases. Presented at the VLDB, Santiago, Chile, pp. 487–499.

Analyze This! | Research Division [WWW Document], 2012. . Res. Pract. Fusion. URL http://www.practicefusion.com/research/analyze-this/ (accessed 2.6.14).

Berry, M.J.A., Linoff, G., 2004. Data mining techniques for marketing, sales, and customer relationship management, 2nd ed. Wiley, Indianapolis.

Bhargavi, B., Venkanna, B., Prasad, V.H., 2013. Mining Frequent Items Using Directed Graphs. Int. J. Sci. Res. Comput. Sci. 1, 21–24.

Big Data Gets Put to Work for Public Health [WWW Document], 2012. . Pract. Fusion. URL http://www.practicefusion.com/pages/pr/big-data-public-health.html (accessed 2.6.14).

CMS, 2014. ICD-9 Code Lookup [WWW Document]. Cent. Medicare Medicaid Serv. URL http://www.cms.gov/medicare-coverage-database/staticpages/icd-9-code-lookup.aspx (accessed 8.6.14).

Differential diagnosis, 2014. . Wikipedia Free Encycl.

Garets, D., Davis, M., 2006. Electronic medical records vs. electronic health records: yes, there is a difference. Policy White Pap. Chic. HIMSS Anal.

Glass, R.L., Ramesh, V., Vessey, I., 2004. An analysis of research in computing disciplines. Commun. ACM 47, 89–94.

Gray, B., Bowden, T., Ib, J., Koch, S., 2011. Electronic Health Records: An International Perspective on "Meaningful Use." Commonw. Fund Pub 1565 28.

HL7 Standards V2.71, 2011.

Holz, H.J., Applin, A., Haberman, B., Joyce, D., Purchase, H., Reed, C., 2006. Research Methods in Computing: What are they, and how should we teach them?, in: ACM SIGCSE Bulletin. ACM, pp. 96–114.

ICD-9-CM Official Guidelines for Coding and Reporting, 2011. . Department of Health and Human Services (DHHS).

ICD-10 Conversion and Mapping - AAPC [WWW Document], 2014. URL http://www.aapc.com/icd-10/conversion-mapping.aspx (accessed 2.6.14).

Kim, H.S., Shin, A.M., Kim, M.K., Kim, Y.N., 2012. Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining. Korean J. Intern. Med. 27, 197–202. doi:10.3904/kjim.2012.27.2.197

Koh, H.C., Tan, G., 2011. Data mining applications in healthcare. J. Healthc. Inf. Manag. 19, 65.

Matheus, C.J., Chan, P.K., Piatetsky-Shapiro, G., 1993. Systems for Knowledge Discovery in Databases. IEEE Trans Knowl Data Eng 5, 903–913. doi:10.1109/69.250073

Meaningful Use Definition & Objectives [WWW Document], n.d. URL http://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives (accessed 2.3.14).

Rajak, A., Gupta, M.K., 2008. Association rule mining-applications in various areas, in: Proceedings of International Conference on Data Management, Ghaziabad, India. pp. 3–7.

Ramesh, V., Glass, R.L., Vessey, I., 2004. Research in computer science: an empirical study. J. Syst. Softw. 70, 165–176. doi:10.1016/S0164-1212(03)00015-3

Rashid, M.A., Hoque, M.T., Sattar, A., 2014. Association Rules Mining Based Clinical Observations. ArXiv Prepr. ArXiv14012571.

Roddick, J.F., Fule, P., Graco, W.J., 2003. Exploratory medical knowledge discovery: Experiences and issues. ACM SIGKDD Explor. Newsl. 5, 94–99.

Schoen, C., Osborn, R., Doty, M.M., Squires, D., Peugh, J., Applebaum, S., 2009. A Survey Of Primary Care Physicians In Eleven Countries, 2009: Perspectives On Care, Costs, And Experiences. Health Aff. (Millwood) 28, w1171–w1183. doi:10.1377/hlthaff.28.6.w1171

Serban, G., Istvan-Gergely, C., Alina, C., 2006. A Programming Interface For Medical diagnosis Prediction. Stud. Univ. Babes - Bolyai Inform., 1 LI, 21–30.

Standards and Guidelines for Electronic Medical Record Systems in Kenya (Health Information Policy), 2009. . Ministries of Health, Government of Kenya.

Tai, Y.-M., Chiu, H.-W., 2009. Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan. Int. J. Med. Inf. 78, e75–83. doi:10.1016/j.ijmedinf.2009.09.005

Van den Akker, M., Buntinx, F., Metsemakers, J.F., Roos, S., Knottnerus, J.A., 1998. Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. J. Clin. Epidemiol. 51, 367–375.

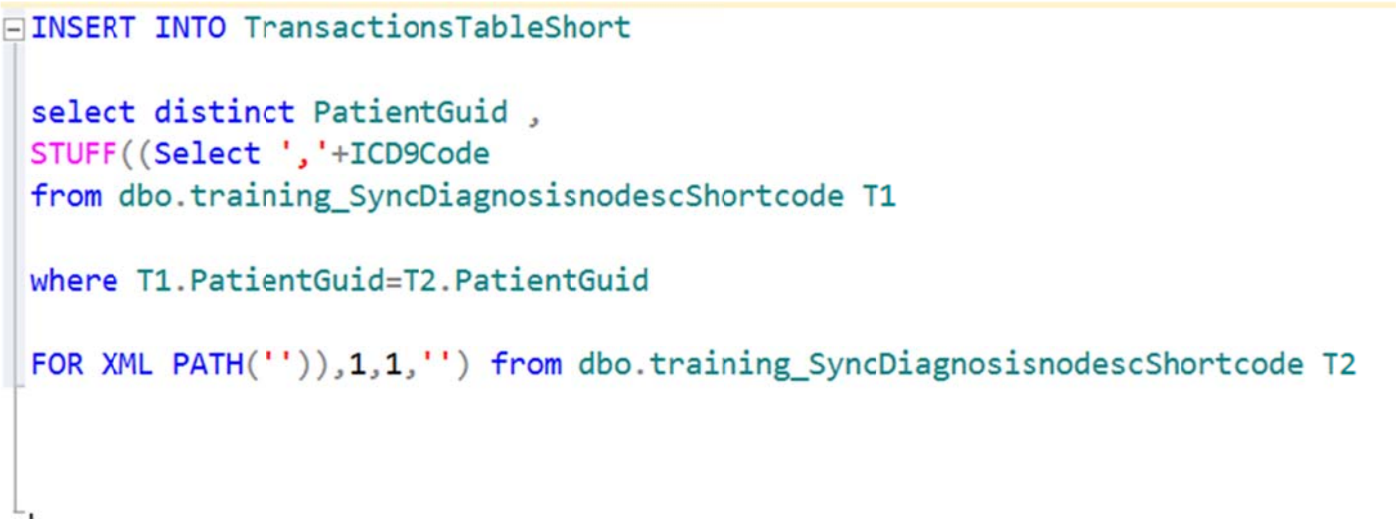
Vimaleswaran, K.S., Cavadino, A., Berry, D.J., Jorde, R., Dieffenbach, A.K., Lu, C., Alves, A.C., Heerspink, H.J.L., Tikkanen, E., Eriksson, J., Wong, A., Mangino, M., Jablonski, K.A., Nolte, I.M., Houston, D.K., Ahluwalia, T.S., van der Most, P.J., Pasko, D., Zgaga, L., Thiering, E., Vitart, V., Fraser, R.M., Huffman, J.E., de Boer, R.A., Schöttker, B., Saum, K.-U., McCarthy, M.I., Dupuis, J., Herzig, K.-H., Sebert, S., Pouta, A., Laitinen, J., Kleber, M.E., Navis, G., Lorentzon, M., Jameson, K., Arden, N., Cooper, J.A., Acharya, J., Hardy, R., Raitakari, O., Ripatti, S., Billings, L.K., Lahti, J., Osmond, C., Penninx, B.W., Rejnmark, L., Lohman, K.K., Paternoster, L., Stolk, R.P., Hernandez, D.G., Byberg, L., Hagström, E., Melhus, H., Ingelsson, E., Mellström, D., Ljunggren, Ö., Tzoulaki, I., McLachlan, S., Theodoratou, E., Tiesler, C.M.T., Jula, A., Navarro, P., Wright, A.F., Polasek, O., Wilson, J.F., Rudan, I., Salomaa, V., Heinrich, J., Campbell, H., Price, J.F., Karlsson, M., Lind, L., Michaëlsson, K., Bandinelli, S., Frayling, T.M., Hartman, C.A., Sørensen, T.I.A., Kritchevsky, S.B., Langdahl, B.L., Eriksson, J.G., Florez, J.C., Spector, T.D., Lehtimäki, T., Kuh, D., Humphries, S.E., Cooper, C., Ohlsson, C., März, W., de Borst, M.H., Kumari, M., Kivimaki, M., Wang, T.J., Power, C., Brenner, H., Grimnes, G., van der Harst, P., Snieder, H., Hingorani, A.D., Pilz, S., Whittaker, J.C., Järvelin, M.-R., Hyppönen, E., 2014. Association of vitamin D status with arterial blood pressure and hypertension risk: a mendelian randomisation study. Lancet Diabetes Endocrinol. doi:10.1016/S2213-8587(14)70113-5

WHO | International Classification of Diseases (ICD) [WWW Document], n.d. . WHO. URL http://www.who.int/classifications/icd/en/ (accessed 2.3.14).

WHO | World Health Organization [WWW Document], 2014. . WHO. URL http://www.who.int/classifications/icd/revision/en/ (accessed 2.5.14).

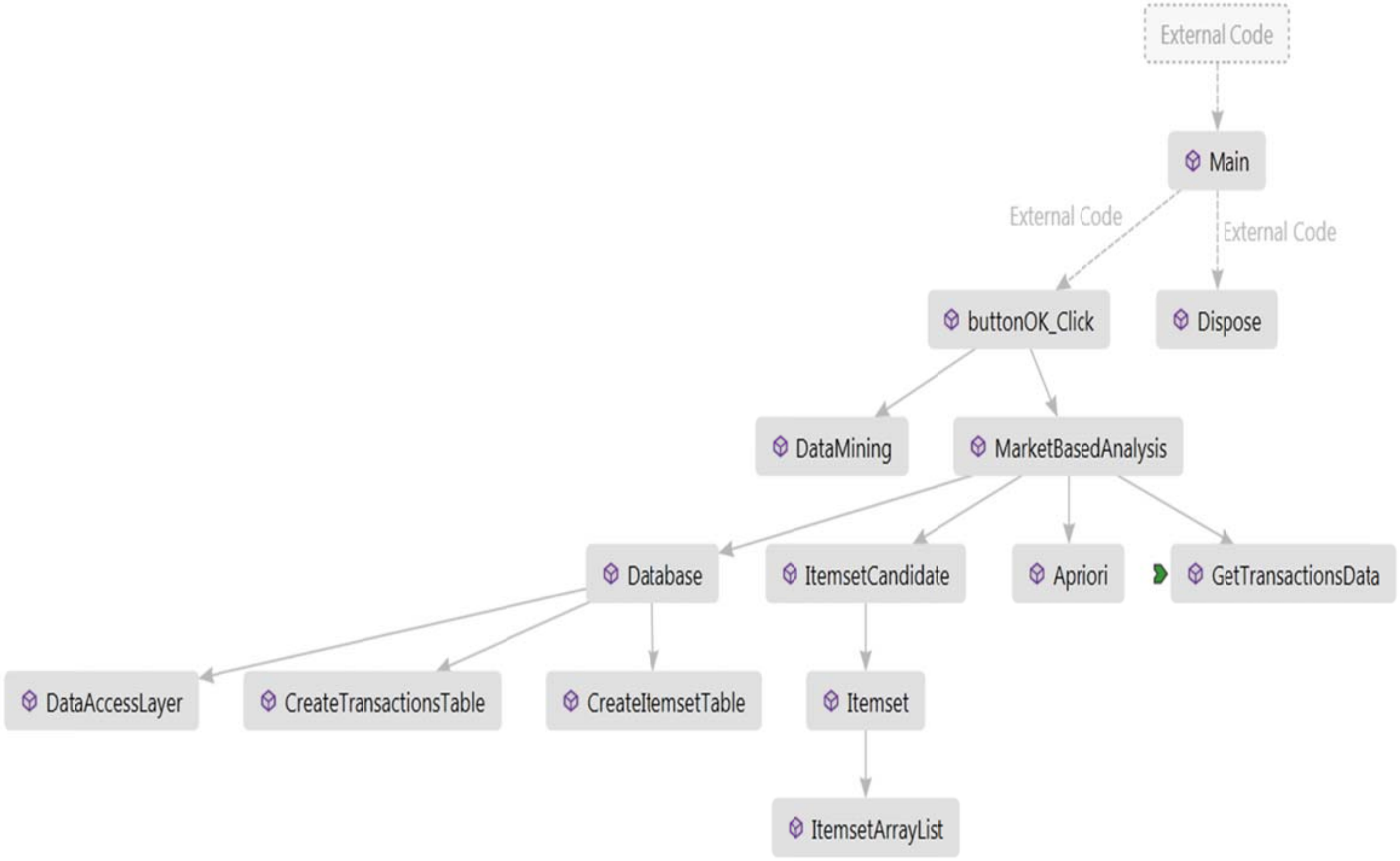
APPENDICES

Appendix A: Transformation of Patient Data to Distinct Comma Separated Values

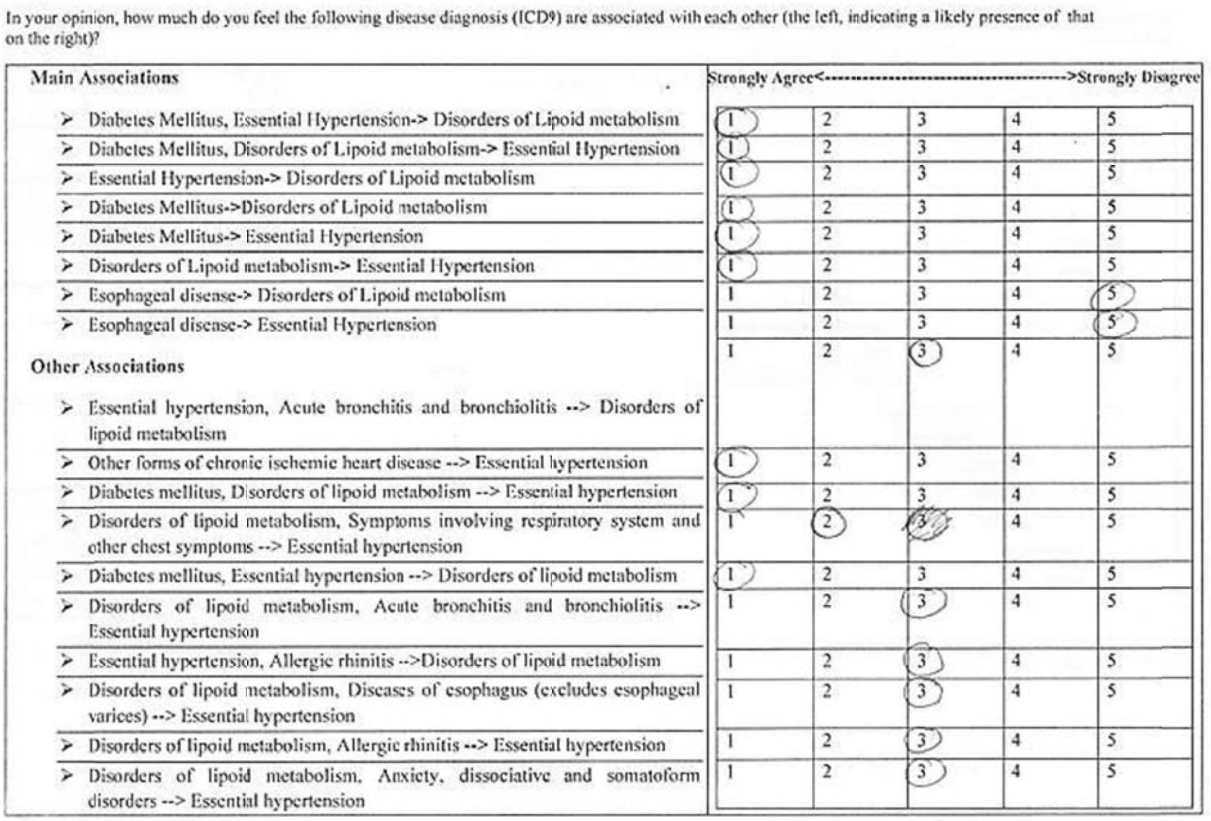


```
INSERT INTO TransactionsTableShort

select distinct PatientGuid ,
STUFF((Select ','+ICD9Code
from dbo.training_SyncDiagnosisnodescShortcode T1

where T1.PatientGuid=T2.PatientGuid

FOR XML PATH('')),1,1,'') from dbo.training_SyncDiagnosisnodescShortcode T2
```

Appendix B: Code Map (Execution Path)

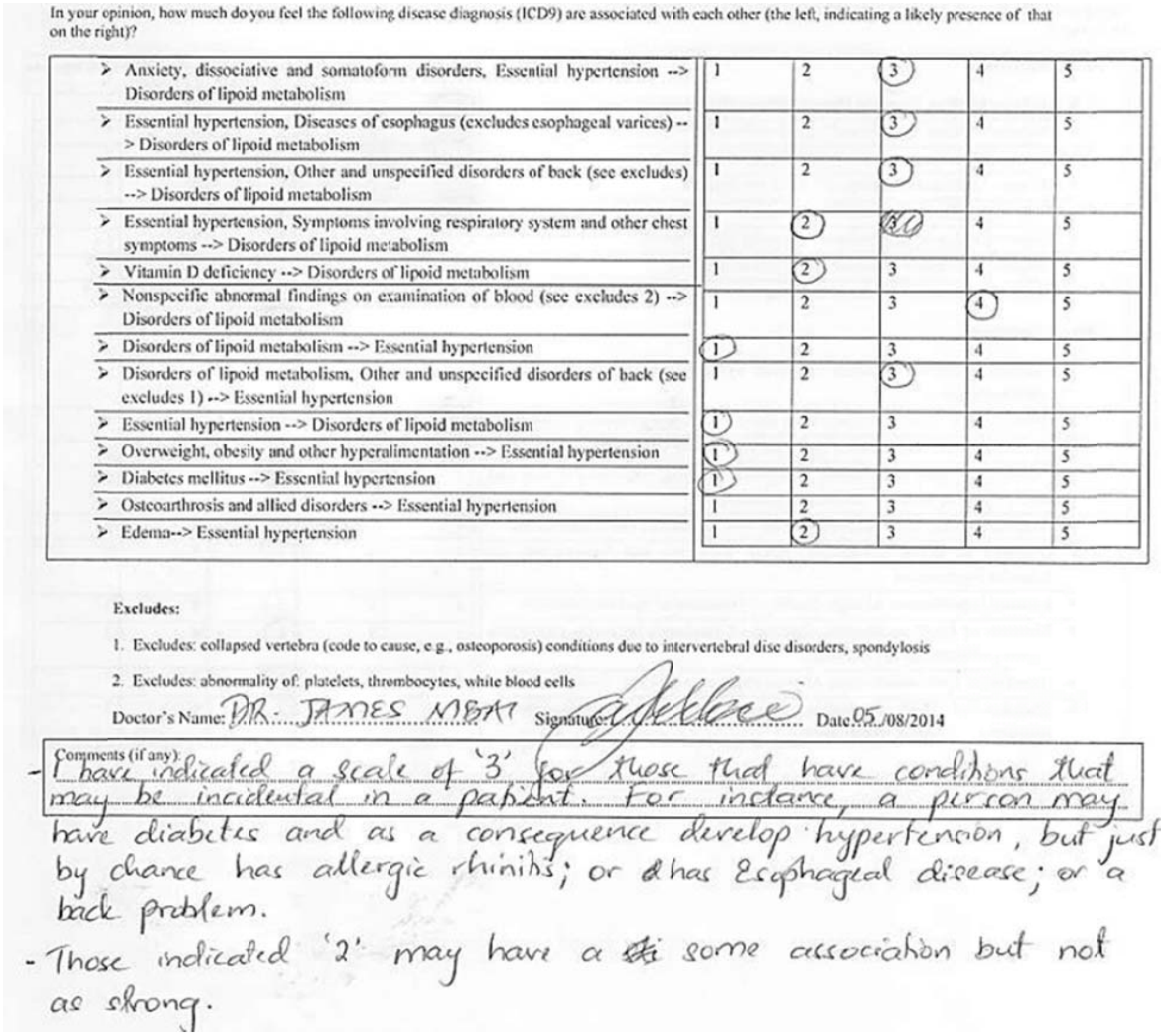


Appendix C: Questionnaire for Survey



In your opinion, how much do you feel the following disease diagnosis (ICD9) are associated with each other (the left, indicating a likely presence of that on the right)?

Main Associations	Strongly Agree<------------------------------------->Strongly Disagree				
➢ Diabetes Mellitus, Essential Hypertension-> Disorders of Lipoid metabolism	[✓ 1]	2	3	4	5
➢ Diabetes Mellitus, Disorders of Lipoid metabolism-> Essential Hypertension	[✓ 1]	2	3	4	5
➢ Essential Hypertension-> Disorders of Lipoid metabolism	[✓ 1]	2	3	4	5
➢ Diabetes Mellitus->Disorders of Lipoid metabolism	[✓ 1]	2	3	4	5
➢ Diabetes Mellitus-> Essential Hypertension	[✓ 1]	2	3	4	5
➢ Disorders of Lipoid metabolism-> Essential Hypertension	[✓ 1]	2	3	4	5
➢ Esophageal disease-> Disorders of Lipoid metabolism	1	2	3	4	[✓ 5]
➢ Esophageal disease-> Essential Hypertension	1	2	3	4	[✓ 5]
Other Associations ➢ Essential hypertension, Acute bronchitis and bronchiolitis --> Disorders of lipoid metabolism	1	2	[✓ 3]	4	5
➢ Other forms of chronic ischemic heart disease --> Essential hypertension	[✓ 1]	2	3	4	5
➢ Diabetes mellitus, Disorders of lipoid metabolism --> Essential hypertension	[✓ 1]	2	3	4	5
➢ Disorders of lipoid metabolism, Symptoms involving respiratory system and other chest symptoms --> Essential hypertension	1	[✓ 2]	~~3~~	4	5
➢ Diabetes mellitus, Essential hypertension --> Disorders of lipoid metabolism	[✓ 1]	2	3	4	5
➢ Disorders of lipoid metabolism, Acute bronchitis and bronchiolitis --> Essential hypertension	1	2	[✓ 3]	4	5
➢ Essential hypertension, Allergic rhinitis -->Disorders of lipoid metabolism	1	2	[✓ 3]	4	5
➢ Disorders of lipoid metabolism, Diseases of esophagus (excludes esophageal varices) --> Essential hypertension	1	2	[✓ 3]	4	5
➢ Disorders of lipoid metabolism, Allergic rhinitis --> Essential hypertension	1	2	[✓ 3]	4	5
➢ Disorders of lipoid metabolism, Anxiety, dissociative and somatoform disorders --> Essential hypertension	1	2	[✓ 3]	4	5



In your opinion, how much do you feel the following disease diagnosis (ICD9) are associated with each other (the left, indicating a likely presence of that on the right)?

➢ Anxiety, dissociative and somatoform disorders, Essential hypertension --> Disorders of lipoid metabolism	1	2	[✓ 3]	4	5
➢ Essential hypertension, Diseases of esophagus (excludes esophageal varices) --> Disorders of lipoid metabolism	1	2	[✓ 3]	4	5
➢ Essential hypertension, Other and unspecified disorders of back (see excludes) --> Disorders of lipoid metabolism	1	2	[✓ 3]	4	5
➢ Essential hypertension, Symptoms involving respiratory system and other chest symptoms --> Disorders of lipoid metabolism	1	[✓ 2]	~~3~~	4	5
➢ Vitamin D deficiency --> Disorders of lipoid metabolism	1	[✓ 2]	3	4	5
➢ Nonspecific abnormal findings on examination of blood (see excludes 2) --> Disorders of lipoid metabolism	1	2	3	[✓ 4]	5
➢ Disorders of lipoid metabolism --> Essential hypertension	[✓ 1]	2	3	4	5
➢ Disorders of lipoid metabolism, Other and unspecified disorders of back (see excludes 1) --> Essential hypertension	1	2	[✓ 3]	4	5
➢ Essential hypertension --> Disorders of lipoid metabolism	[✓ 1]	2	3	4	5
➢ Overweight, obesity and other hyperalimentation --> Essential hypertension	[✓ 1]	2	3	4	5
➢ Diabetes mellitus --> Essential hypertension	[✓ 1]	2	3	4	5
➢ Osteoarthrosis and allied disorders --> Essential hypertension	1	2	3	4	5
➢ Edema--> Essential hypertension	1	[✓ 2]	3	4	5

Excludes:

1. Excludes: collapsed vertebra (code to cause, e.g., osteoporosis) conditions due to intervertebral disc disorders, spondylosis

2. Excludes: abnormality of: platelets, thrombocytes, white blood cells

Doctor's Name: DR: JAMES MBAT Signature: Date 05/08/2014

Comments (if any):

- I have indicated a scale of '3' for those that have conditions that may be incidental in a patient. For instance, a person may have diabetes and as a consequence develop hypertension, but just by chance has allergic rhinitis; or ~~&~~ has Esophageal disease; or a back problem.
- Those indicated '2' may have a ~~sti~~ some association but not as strong.

Appendix D: Demographic Prevalence

ICD Code	Dx Description	Male	Female	% Male	% Female	Avg Age
250,401 --> 272	Diabetes Mellitus, Essential Hypertension-> Disorders of Lipoid metabolism	578	555	51	49	64
401 --> 272	Essential Hypertension->Disorders of Lipoid metabolism	1321	1359	49	51	63
250 --> 272	Diabetes Mellitus->Disorders of Lipoid metabolism	746	708	51	49	62
250 --> 401	Diabetes Mellitus->Essential Hypertension	750	725	51	49	63.5
530 --> 272	Esophageal disease->Disorders of Lipoid metabolism	425	509	46	54	60.5
530 --> 401	Esophageal disease->Essential Hypertension	380	495	43	57	61.5
401,466 --> 272	Essential Hypertension, Acute Bronchitis->Disorders of Lipoid metabolism	204	201	50	50	61.5
414 --> 401	Other forms of chronic ischemic heart disease(Coronary atherosclerosis (of native coronary artery))->Essentia	268	163	62	38	69
272,466 --> 401	Disorders of lipoid metabolism,Acute bronchitis and bronchiolitis --> Essential hypertension	204	201	50	50	61.5
401,477 --> 272	Essential hypertension,Allergic rhinitis --> Disorders of lipoid metabolism	272	327	45	55	63
272,530 --> 401	Disorders of lipoid metabolism,Diseases of esophagus (excludes esophageal varices) --> Essential hyperter	267	340	44	56	63
272,300 --> 401	Disorders of lipoid metabolism,Anxiety, dissociative and somatoform disorders --> Essential hypertension	183	270	40	60	59.5
401,724 --> 272	Essential hypertension,Other and unspecified disorders of back (see excludes) --> Disorders of lipoid metab	291	334	47	53	61.5
401,786 --> 272	Essential hypertension,Other and unspecified disorders of back (see excludes) --> Disorders of lipoid metab	249	301	45	55	62
268 --> 272	Vitamin D deficiency --> Disorders of lipoid metabolism	225	286	44	56	59
790 --> 272	Nonspecific abnormal findings on examination of blood (see excludes 2) --> Disorders of lipoid metabolism	331	185	64	36	59
278 --> 401	Overweight, obesity and other hyperalimentation --> Essential hypertension	312	413	43	57	55.5
715 --> 401	Osteoarthrosis and allied disorders --> Essential hypertension	271	424	39	61	67
782 --> 401	Edema--> Essential hypertension	177	332	35	65	65

Appendix E: Sample Code

Datamining.cs

```
using System;
using System.Data;
using VISUAL_BASIC_DATA_MINING_NET.APriori;
using VISUAL_BASIC_DATA_MINING_NET.CustomEvents;


       /// <summary>
       /// The VISUAL_BASIC_DATA_MINING_NET namespace contains namespaces and classes 
used by this assembly.
       /// </summary>
namespace VISUAL_BASIC_DATA_MINING_NET
{

       /// <summary>
       /// A class that provides data mining services using C#.NET, ADO.NET, XML.NET and 
a Diagnosis Analysis 
       /// Data Mining Algorithm.
       /// </summary>
       public class DataMining
       {

              //test
              /// <summary>
              /// Initializes a new instance of the DataMining class using a 
parameterless default constructor.
              /// </summary>
              public DataMining()
              {
              }


              /// <summary>
              /// Initializes a new instance of the Diagnosis Analysis Data Mining class 
and sets it's 
              /// properties.
              /// See <see 
cref="VISUAL_BASIC_DATA_MINING_NET.DataMining.MarketBasedAnalysis"/> .
              /// </summary>
              /// <returns>
              /// A System.Data.DataSet object containing the tables of the dx  Based 
Data Mining Analysis.
              /// </returns>
              public DataMining(double supportCount, double minimumConfidence, string 
connectionString, string 
                     dataSource, CommandType dataSourceCommand)
              {
                     this.minimumSupportCount = supportCount;

                     this.minimumConfidence = minimumConfidence;

                     this.connectionString = connectionString;

                     this.dataSource = dataSource;
                     this.dataSourceCommand = dataSourceCommand;
              }


              /// <param name="e">
              /// A CustomEvents.ProgressMonitorEventArgs object.
              /// </param>
              /// <remarks>
              /// This method is used to invoke a dalegate that notifies users about the
progress of an executing code.
              /// </remarks>
              public void OnProgressMonitorEvent(ProgressMonitorEventArgs e)
              {
                     if (ProgressMonitorEvent != null)
                     {
                            // Invokes the delegates.
                            ProgressMonitorEvent(this, e);
                     }
              }

              /// <summary>
              /// The public OnProgressMonitoringCompletedEvent raises the
ProgressMonitorEvent event by invoking
              /// the public OnProgressMonitorEvent. This method is invoked by an event
source when an event monitoring
              /// is completed.
              /// </summary>
              /// <param name="e">
              /// A CustomEvents.ProgressMonitorEventArgs object.
              /// </param>
              /// <remarks>
              /// This method is used to pass messages from event sources to users.
              /// </remarks>
              public void OnProgressMonitoringCompletedEvent(object sender,
ProgressMonitorEventArgs e)
              {
                     if (ProgressMonitorEvent != null)
                     {
                            // Invokes the delegates.
                            ProgressMonitorEvent(this, e);
                     }

                     else
                     {
                            ProgressMonitorEventArgs newE = new
ProgressMonitorEventArgs(e);

                            ProgressMonitorEvent(this, e);
                     }
              }

              /// <summary>
              /// A custom event that notifies user  about the progress of the executing
code.
              /// </summary>
              public event ProgressMonitorEventHandler ProgressMonitorEvent;
              /// <summary>
              /// APriori.Apriori implements the C#.NET diagnosis data mining algorithm
that we will use to mine the diagnosis affinity data.
              /// </summary>
              protected Apriori AP;

              /// <summary>
              /// A strongly typed DataSet containing an in-memory cache of the results
of the Diagnosis Analysis Data Mining.
              /// </summary>

              protected Data dataBase;

              /// <summary>
              /// A System.Data.DataView object for viewing the tables of the Diagnosis
Analysis tables.
              /// </summary>
              protected DataView viewDataMiningAnalysis;

              /// <summary>
              /// Stores the minimum support count required for every frequent set of
items.
              /// </summary>
              protected double minimumSupportCount;

              /// <summary>
              /// The minimum confidence required for the diagnosis data mining rules
created.
              /// See <see
cref="VISUAL_BASIC_DATA_MINING_NET.DataMining.MarketBasedAnalysis"/>
              /// </summary>
              protected double minimumConfidence;

              /// <summary>
              /// A string used to connect to a relational database like SQL Server, Ms
Access or Oracle.
              /// </summary>
              protected string connectionString;

              /// <summary>
              /// Stores the path to an XML file that contains Transactions data.
              /// </summary>
              protected string xmlFilePath;

              /// <summary>
              /// A stored procedure, table or SQL SELECT statement that will provide the
diagnosis transactions data.
              /// </summary>
              protected string dataSource;

              /// <summary>
              /// A CommandType enumeration of CommandType.StoredProcedure or
CommandType.Text or CommandType.TableDirect.
              /// </summary>
              protected CommandType dataSourceCommand;

              /// <summary>
              /// A System.Int32 variable that contains the number of transactions in the
transactions table.
              /// </summary>
              protected int transactionsCount;


              /// <summary>
              ///The support count is the number of transactions in a database containing
a set of items.
              /// </summary>
              /// <value>
              /// A public read only System.Int32 variable.
              /// </value>
              public double MinimumSupportCount
              {
                     get
                     {
                            return minimumSupportCount;
                     }
              }


              /// <summary>
              /// The minimum confidence required for the diagnosis data mining rules
created.
              /// </summary>
              /// <value>
              /// A public read only System.Int32 variable.
              /// </value>
              public double MinimumConfidence
              {
                     get
                     {
                            return minimumConfidence;
                     }
              }



              /// <example>
              ///  "SELECT TransactionID, Transactions FROM TransactionsTable"
              /// </example>
              /// </para>
              /// </summary>
              /// <value>
              /// A public read only System.String variable.
              /// </value>
              public string ConnectionString
              {
                     get
                     {
                            return connectionString;
                     }
              }


              /// <summary>
              /// The path to an XML file that contains Transactions data.
              /// </summary>
              /// <value>
              /// A public read only System.String variable.
              /// </value>
              /// <include file='APrioriExamples.xml'
path='Documentation/SourceCode[@name="StartingSampleC"]/*' />
              public string XMLFilePath
              {
                     get
                     {
                            return xmlFilePath;
                     }
              }


              ///<summary>
              /// A string containing a SQL statement, a table name or the name of a
stored procedure.
              /// <para>
              /// To use a table it must have a TransactionID field and a Transactions
field.
              /// </para>
              /// <para>
              /// To use a stored procedure named usp_GetTransactions, create the stored
procedure in SQL Server using :
              /// <code>
              /// <example>
              /// CREATE  PROCEDURE usp_GetTransactions AS
              ///
              ///    SELECT TransactionID, Transactions FROM TransactionsTable
              ///    </example>
              /// </code>
              /// </para>
              /// </summary>
              /// <value>
              /// A public System.String variable.
              /// </value>
              public string DataSource
              {
                     get
                     {
                            return dataSource;
                     }
              }


              /// <summary>
              /// A CommandType enumeration of CommandType.StoredProcedure or
CommandType.Text or CommandType.TableDirect.
              /// </summary>
              /// <value>
              /// A public System.Data.CommandType enumeration.
              /// </value>
              public CommandType DataSourceCommand
              {
                     get
                     {
                            return dataSourceCommand;
```

```
            }
        }


        /// <summary>
        /// A strongly typed DataSet containing an in-memory cache of the results 
of the Diagnosis Analysis Data Mining.
        /// </summary>
        /// <value>
        /// A VISUAL_BASIC_DATA_MINING_NET.Data strongly typed System.Data.DataSet 
object.
        /// </value>
        public Data DataBase
        {
            get
            {
                return dataBase;
            }
        }


        /// <summary>
        /// Retrieves the results of a Diagnosis Analysis as an in-memory cache of 
data.
        /// </summary>
        /// <param name="supportCount">
        /// The support count is the number of transactions containing a set of 
items.
        /// </param>
        /// <param name="minimumConfidence">
        /// The confidence of two sets of items A and B is the number of 
transactions supported by A and B
        /// divided by the number of transactions divided by A and vice versa.
        /// <example>
        /// confidence[A->B) = (number of transactions containing both A and B) / 
(number of transactions
        ///
                                        containing only A)
        /// </example>
        /// </param>
        /// <param name="connectionString">
        /// The connection string used to establish connection to a relational 
database using ADO.NET.
        /// <example>
        /// An example of a connection string using Windows Integrated Security :
        /// <para>
        /// string connect = "Provider=SQLOLEDB;Data Source=localhost;Initial 
Catalog=Sales;" + "Integrated Security=SSPI;";
        /// </para>
        /// An example of a string not using Windows Integrated Security :
        /// <para>
        /// "Provider=SQLOLEDB;Data Source=localhost;User 
ID=Analyst;Password=DataMining;Initial Catalog=Shopping Cart";
        /// </para>
        /// </example>
        /// </param>
        /// <param name="dataSource">
```

```
              /// A string containing a SQL statement, a table name or the name of a
stored procedure.
              /// <para>
              /// The table must have a TransactionID field and a Transactions field.
              /// </para>
              /// <para>
              /// To use a stored procedure named usp_GetTransactions, create the stored
procedure in SQL Server using :
              /// <code>
              /// <example>
              /// CREATE  PROCEDURE usp_GetTransactions AS
              ///
              ///    SELECT TransactionID, Transactions FROM TransactionsTable
              ///    </example>
              /// </code>
              /// </para>
              /// <para>
              ///  The SQL statement used to select transactions data.
              /// <example>
              ///  "SELECT TransactionID, Transactions FROM TransactionsTable"
              /// </example>
              /// </para>
              /// </param>
              /// <param name="dataSourceCommand">
              ///  A CommandType enumeration of CommandType.StoredProcedure or
CommandType.Text or CommandType.TableDirect.
              /// </param>
              /// <returns>
              /// A System.Data.DataSet in-memory database containing the Diagnosis
Analysis results in
              /// the TransactionsTable, ItemsetTable, SubsetTable, Rulestable.
              /// </returns>
              /// <remarks>
              /// See  <see
cref="VISUAL_BASIC_DATA_MINING_NET.DataAccessLayer.GetTransactionsData"/>
              /// </remarks>
              public Data MarketBasedAnalysis(double supportCount, double
minimumConfidence, string connectionString, string

       dataSource, CommandType dataSourceCommand)
              {

                     Database database = new Database();

                     ItemsetCandidate Item = new ItemsetCandidate();

                     this.AP = new APriori.Apriori();

                     this.AP.ProgressMonitorEvent += new
ProgressMonitorEventHandler(this.OnProgressMonitoringCompletedEvent);


                     this.dataBase = database.GetTransactionsData(connectionString,
dataSource, dataSourceCommand);

                     database.Transactions = this.dataBase;
                     this.transactionsCount = this.dataBase.TransactionTable.Count;

                     supportCount = ((supportCount / 100) * this.transactionsCount);

                     minimumConfidence = (minimumConfidence / 100);

                     string support = "SupportCount >= " + supportCount + " AND Level > 
1";

                     string sort = "SupportCount, Level";

                     ItemsetCandidate uniqueItems = AP.CreateOneItemsets(database);


       AP.AprioriGenerator(uniqueItems,database,Convert.ToInt32(supportCount));

                     ItemsetArrayList [] keys = database.GetItemset(support, sort);


                     string msg = "Creating Frequent Subsets for Items";

                     ProgressMonitorEventArgs e = new 
ProgressMonitorEventArgs(1,100,95,"DataMining.MarketBasedAnalysis(3)",msg );

                     this.OnProgressMonitorEvent(e);

                     for(int counter = 0; counter < keys.Length; counter++)
                     {
                            AP.CreateItemsetSubsets(0,keys[counter], null, database);
                     }


                     msg = "Completed Diagnosis Affinity Analysis";

                     e = new 
ProgressMonitorEventArgs(1,100,100,"DataMining.MarketBasedAnalysis(3)",msg );

                     this.OnProgressMonitorEvent(e);


                     //Set the public properties of the class
                     this.minimumSupportCount = supportCount;

                     this.minimumConfidence = minimumConfidence;

                     this.connectionString = connectionString;

                     this.dataSource = dataSource;
                     this.dataSourceCommand = dataSourceCommand;

                     //return the database of transactions
                     return this.dataBase;

              }


              /// <summary>
              /// Retrieves the results of a Diagnosis Analysis as an in-memory cache of
data from an XML file.
              /// </summary>
              /// <param name="supportCount">
              /// The support count is the number of transactions containing a set of
items.
              /// </param>
              /// <param name="minimumConfidence">
              /// The confidence of two sets of items A and B is the number of
transactions supported by A and B
              /// divided by the number of transactions divided by A and vice versa.
              /// <example>
              /// confidence[A->B) = (number of transactions containing both A and B) /
(number of transactions
              ///
                                                        containing only A)
              /// </example>
              /// </param>
              /// <param name="xmlFilePath">
              /// The path to an XML file containing transaction data.
              /// </param>
              /// <returns>
              /// A System.Data.DataSet in-memory database containing the Diagnosis
Analysis results in
              /// the TransactionsTable, ItemsetTable, SubsetTable, Rulestable.
              /// </returns>

              public Data MarketBasedAnalysis(double supportCount, double
minimumConfidence, string xmlFilePath)
              {

                     Database database = new Database();

                     ItemsetCandidate Item = new ItemsetCandidate();


                     this.AP = new APriori.Apriori();

                     this.AP.ProgressMonitorEvent += new
ProgressMonitorEventHandler(this.OnProgressMonitoringCompletedEvent);


                     this.dataBase = database.GetXMLData(xmlFilePath);


                     database.Transactions = this.dataBase;


                     this.transactionsCount = this.dataBase.TransactionTable.Count;
                     supportCount = ((supportCount / 100) * this.transactionsCount);

                     minimumConfidence = (minimumConfidence / 100);


                     string support = "SupportCount >= " + supportCount + " AND Level >
1";

                     string sort = "SupportCount, Level";


                     ItemsetCandidate uniqueItems = AP.CreateOneItemsets(database);



       AP.AprioriGenerator(uniqueItems,database,Convert.ToInt32(supportCount));


                     ItemsetArrayList [] keys = database.GetItemset(support, sort);



                     string msg = "Creating Frequent Subsets for Items";

                     ProgressMonitorEventArgs e = new
ProgressMonitorEventArgs(1,100,95,"DataMining.MarketBasedAnalysis(3)",msg );

                     this.OnProgressMonitorEvent(e);


                     for(int counter = 0; counter < keys.Length; counter++)
                     {
                            AP.CreateItemsetSubsets(0,keys[counter], null, database);
                     }




                     msg = "Completed Diagnosis Affinity Analysis";

                     e = new
ProgressMonitorEventArgs(1,100,100,"DataMining.MarketBasedAnalysis(3)",msg );

                     this.OnProgressMonitorEvent(e);


                     //Set the public properties of the class
                     this.minimumSupportCount = supportCount;

                     this.minimumConfidence = minimumConfidence;

                     this.xmlFilePath = xmlFilePath;

                     //return the database of transactions
                     return this.dataBase;

              }
              public Data MarketBasedAnalysis(double supportCount, double 
minimumConfidence, Data transactionsData)
              {

                     Database database = new Database();

                     ItemsetCandidate Item = new ItemsetCandidate();


                     this.AP = new APriori.Apriori();

                     this.AP.ProgressMonitorEvent += new 
ProgressMonitorEventHandler(this.OnProgressMonitoringCompletedEvent);


                     this.dataBase = transactionsData;


                     database.Transactions = this.dataBase;


                     this.transactionsCount = this.dataBase.TransactionTable.Count;


                     supportCount = ((supportCount / 100) * this.transactionsCount);

                     minimumConfidence = (minimumConfidence / 100);


                     string support = "SupportCount >= " + supportCount + " AND Level > 
1";

                     string sort = "SupportCount, Level";


                     ItemsetCandidate uniqueItems = AP.CreateOneItemsets(database);



       AP.AprioriGenerator(uniqueItems,database,Convert.ToInt32(supportCount));


                     ItemsetArrayList [] keys = database.GetItemset(support, sort);



                     string msg = "Creating Frequent Subsets for Items";

                     ProgressMonitorEventArgs e = new 
ProgressMonitorEventArgs(1,100,95,"DataMining.MarketBasedAnalysis(3)",msg );

                     this.OnProgressMonitorEvent(e);


                     for(int counter = 0; counter < keys.Length; counter++)
                     {
                            AP.CreateItemsetSubsets(0,keys[counter], null, database);
                     }


            msg = "Completed Diagnosis Affinity Analysis";

                     e = new
ProgressMonitorEventArgs(1,100,100,"DataMining.MarketBasedAnalysis(3)",msg );

                     this.OnProgressMonitorEvent(e);



                     //Set the public properties of the class
                     this.minimumSupportCount = supportCount;

                     this.minimumConfidence = minimumConfidence;

                     //return the database of transactions
                     return this.dataBase;

              }


              /// <summary>
              /// A DataView for viewing the results of the Diagnosis Analysis Data
Mining results.
              /// </summary>
              /// <returns>
              ///A System.Data.DataView object for viewing the tables of the Diagnosis
Analysis tables.
              /// </returns>
              public DataView ViewDataMiningAnalysis()
              {
                     double minimumconfidence = ((this.minimumConfidence) * 100);

                     string confidence = "Confidence >= " + minimumconfidence + "%";

                     viewDataMiningAnalysis = new
DataView(this.dataBase.Tables["ViewRulesTable"], confidence, "Confidence",

       DataViewRowState.CurrentRows);

                     return viewDataMiningAnalysis;
              }


              /// <summary>
              /// A DataView for viewing the contents of any of the Diagnosis Analysis
tables.
              /// </summary>
              /// <param name="tableName">
              /// The name of the table containing the results to be viewed with the
DataView object.
              /// </param>
              /// <returns>
              /// A System.Data.DataView object for viewing the tables of the Diagnosis
Analysis tables.
```

```
              /// </returns>
              public DataView ViewDataMiningAnalysis(string tableName)
              {
                     viewDataMiningAnalysis = new
DataView(this.dataBase.Tables[tableName]);

                     return viewDataMiningAnalysis;
              }


              /// <summary>
              /// A DataView for viewing the contents of any of the Diagnosis Analysis
tables.
              /// </summary>
              /// <param name="tableName">
              /// The name of the table containing the results to be viewed with the
DataView object.
              /// </param>
              /// <param name="sortColumn">
              /// The name of the table column to use in sorting the data to be viewed.
              /// </param>
              /// <returns>
              /// A System.Data.DataView object for viewing the tables of the Diagnosis
Analysis tables.
              /// </returns>
              public DataView ViewDataMiningAnalysis(string tableName, string sortColumn)
              {
                     viewDataMiningAnalysis = new
DataView(this.dataBase.Tables[tableName],"",sortColumn,DataViewRowState.CurrentRows);

                     return viewDataMiningAnalysis;
              }


              /// <summary>
              /// A DataView for viewing the contents of any of the Diagnosis Analysis
tables.
              /// </summary>
              /// <param name="tableName">
              /// The name of the table containing the results to be viewed with the
DataView object.
              /// </param>
              /// <returns>
              /// A System.Data.DataView object for viewing the tables of the Diagnosis
Analysis tables.
              /// </returns>
              public DataView ViewDataMiningAnalysis(string tableName, double
minimumConfidence)
              {
                     string confidence = "Confidence >= " + minimumConfidence;

                     viewDataMiningAnalysis = new
DataView(this.dataBase.Tables[tableName], confidence, "Confidence",

       DataViewRowState.CurrentRows);

                     return viewDataMiningAnalysis;
              }
```

```
    }
}
```

Viewdata.cs

```
/// <summary>
        /// Fetches and transforms data in the RulesTable into the ViewRulesTable.
        /// </summary>
        /// <param name="dataset">
        /// A DataSet containing RulesTable and SubsetsTable.
        /// </param>
        /// <param name="minimumConfidence">
        /// The minimum confidence for each item in the ViewRulesTable.
        /// </param>
        /// <returns>
        /// A System.Data.DataTable object.
        /// </returns>
        /// <remarks>
        /// Creates a ViewRulesTable containing only items that satisfy a minimum
confidence.
        /// </remarks>
        public DataTable CreateViewRulesTable(double minimumConfidence, Data
dataset)
        {

            int leftRuleID = 0;

            int rightRuleID = 0;

            double confidence = 0;

            DataRow newRow;

            Data.ItemsetTableRow itemLeftRow;

            Data.ItemsetTableRow itemRightRow;


            minimumConfidence = (minimumConfidence/100);


            foreach(Data.RulesTableRow ruleRow in dataset.RulesTable.Rows)
            {

                confidence = ruleRow.Confidence;


                if( confidence >= minimumConfidence )
                {

                    newRow = this.viewRulesTable.NewRow();


                    newRow["UniqueID"] = ruleRow.FirstKeyID;
```

```
                                   leftRuleID = ruleRow.LeftRule;

                                   itemLeftRow =
dataset.ItemsetTable.FindByItemID(leftRuleID);

                                   rightRuleID = ruleRow.RightRule;

                                   itemRightRow =
dataset.ItemsetTable.FindByItemID(rightRuleID);

                                   newRow["Analysis"] = itemLeftRow.Itemset + "  -->  " +
itemRightRow.Itemset;


                                   //newRow["Confidence"] = ((ruleRow.Confidence) * (100)
+ "%");
                    newRow["Confidence"] = ((Math.Round(ruleRow.Confidence,4)) * (100));

                                   this.viewRulesTable.Rows.Add(newRow);

                            }

                     }

                     return this.viewRulesTable;
              }
```

DataAccesLayer.cs

```
  /// <summary>
        /// Retrieves diagnosis transaction data from a transactions table.
        ///  See <see
cref="VISUAL_BASIC_DATA_MINING_NET.DataAccessLayer.GetTransactionsData"/>
        /// </summary>
        /// <param name="rdbmsConnectionString">
        /// The connection string used to establish connection to a relational database
using ADO.NET.
        ///  See <see
cref="VISUAL_BASIC_DATA_MINING_NET.DataAccessLayer.GetTransactionsData"/>
        /// </param>
        /// <param name="dataSource">
        /// A string containing a SQL statement, a table name or the name of a stored
procedure.
        /// <para>
        /// The table must have a TransactionID field and a Transactions field.
        /// </para>
        /// <para>
        /// To use a stored procedure named usp_GetTransactions, create the stored
procedure in SQL Server using :
        /// <code>
        /// <example>
        /// CREATE  PROCEDURE usp_GetTransactions AS
        ///
        ///   SELECT TransactionID, Transactions FROM TransactionsTable
        ///   </example>
        /// </code>
        /// </para>
        /// <para>
        ///  The SQL statement used to select transactions data.
        /// <example>
        ///  "SELECT TransactionID, Transactions FROM TransactionsTable"
        /// </example>
        /// </para>
        /// </param>
        /// <param name="commandType">
        /// A CommandType enumeration of CommandType.StoredProcedure or CommandType.Text
or CommandType.TableDirect.
        /// </param>
        /// <returns>
        /// See
        ///  <see
cref="VISUAL_BASIC_DATA_MINING_NET.DataAccessLayer.GetTransactionsData"/>
        /// </returns>
        public Data GetTransactionsData(string rdbmsConnectionString, string dataSource,
CommandType commandType)
        {

            myDatabase = new Data();

            myConnection = new OleDbConnection(rdbmsConnectionString);

            myCommand = new OleDbCommand(dataSource, myConnection);


            myCommand.CommandType = CommandType.Text;
            myCommand.CommandText = "SELECT top 1000 TransactionID, Transactions FROM
TransactionsTableShort where TransactionID>2000";


            myAdapter = new OleDbDataAdapter();

            myAdapter.SelectCommand = myCommand;

            myAdapter.Fill(myDatabase, "TransactionTable");


            return myDatabase;

        }
```