DATA ANALYTICS FOR HOSPITAL OPERATIONS

BY

ABDULRAHMAN EHAB EL SAFTY

BACHELOR OF ENGINEERING (HONS) ELECTRICAL & ELECTRONICS

UNIVERSITI TEKNOLOGI PETRONAS

MAY 2015

ABDULRAHMAN EHAB EL SAFTY

B. ENG (HONS) ELECTRICAL & ELECTRONICS MAY 2015

DATA ANALYTICS FOR HOSPITAL OPERATIONS

BY

ABDULRAHMAN EHAB EL SAFTY

14705

Supervised by

Dr. Ho Tatt Wei

Dissertation submitted in partial fulfillment of
the requirements for the

Bachelor of Engineering (Hons)

(Electrical & Electronics)

MAY 2015

Universiti Teknologi PETRONAS,
Bandar Seri Iskandar,
32610, Perak Darul Ridzuan,
Malaysia

CERTIFICATION OF APPROVAL

DATA ANALYTICS FOR HOSPITAL OPERATIONS

PREPARED BY,

ABDULRAHMAN EHAB EL SAFTY
14705

A dissertation submitted to the
Electrical & Electronics Programme
Universiti Teknologi PETRONAS
In partial fulfilment requirement for the
BACHELOR OF ENGINEERING (HONS)
ELECTRICAL AND ELECTRONICS

Approved by,

____

(Dr. Ho Tatt Wei)
Project Supervisor

UNIVERSITI TEKNOLOGI PETRONAS,
TRONOH, PERAK
MAY 2015

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this paper, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

ABDULRAHMAN EHAB EL SAFTY

ACKNOWLEDGEMENT

First and foremost, I would like to take this opportunity to extend the utmost gratitude to those who have contributed either directly or indirectly for their continuous support and contribution in completing this Final Year Project. These 8 months of experience have given me a chance to learn something new and embrace the challenge that has been given to me.

I would like to appreciate the insight encouragement given by my supervisor Dr. Ho Tatt Wei for guiding me with endless patience and cooperation helping me completing this project. I would like to thank him for continuous support and motivation and I sincerely appreciate all the guidance and knowledge he taught me to preparing me to face the real world in the coming years. Henceforth, I am determined to strive for the best so that I could apply all that I have learnt throughout this journey.

Appreciated as well, the help that was provided from my colleagues in the project. Special thanks to Patrick Lee Sheng Siang, and Nur Aqilah Binti Azam for aiding me in various ways, especially when I needed the most.

Finally, I would like to thank my parents, family and friends who helped me a lot in finishing this project within the limited time and gave me endless support and encouragement which enabled me to provide my best for this project. I hope after this program, all the knowledge and experience I gained can be shared with everyone and especially students across the globe.

ABSTRACT

Modern hospitals and clinics produce tons of electronic data every day regarding patients, medications, treatments, and diseases. These amounts of data contain the potential to help humanity understand and analyze the biomedical fields from a statistical and predictive point of view. Throughout the years, research has been concluded to develop methods of interpreting and analyzing this data. Biomedical statistical researchers have experimented algorithms to achieve findings and associations within the aspects of the data given. Medical decision-making is becoming more and more dependent on data analysis, rather than conventional experience and intuition. Hence, this project will look into the feasibility of developing software for hospital data analysis, specifically, the MIMIC-II (Multiparameter Intelligent Monitoring in Intensive Care) data that support a diverse range of analytic studies which extend across epidemiology, clinical decision-rule improvement, and electronic tool development. This statistical software is programmed to run multiple algorithms on the massive datasets in the mean of revealing similarities of items related to sets by focusing on the algorithm based of Market-Basket method. With the aim to assist in showing patterns and associations in hospital big data, the software reveals associations and apprehending patterns inside this data demonstrated as predictive analytics that can assist in handling comparable cases and present clinical and hospital-decisions.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. CHAPTER 1: INTRODUCTION

1.1. Project Background

The industry of healthcare is transforming globally moving towards a value-based industry [1]. Increasing demands from consumers are asking for many things such as, better quality of healthcare, more value, pressurizing the healthcare providers to bring enhanced outcomes. The shortages in physicians and nurses expect overworked professionals to work more efficiently and productively. The increase in life expectancy as well as the spread of continuous diseases is changing the healthcare cost dynamics.

The healthcare industry throughout history has brought up large quantities of data, motivated by keeping records, as well as patient care. Since data was stored as a hard copy form in the past, the present goal is heading towards fast digitization of these vast amounts of data. These huge amounts of data aka ‘big data’ hold the potential of aiding a broad range of medical and healthcare purposes, also others biological decisions assistance, disease monitoring, and improvement of population health. [2].

Performing analytics to obtain more clarified insights is capable of showing value and achieving better results, such as modern treatments and technologies. That information that leads to vision aids knowledgeable and educated citizens to act more responsibly for their own wellbeing. Analytics have the possibility to maintain better effectiveness and overall efficiency. From handling small details up to huge processes, this analytics is able to explore and discover; help structure and form policy as well as programs; advance in delivering service as well as operations; improve sustainability; reduce risk; also provide ways and methods to evaluate and calculate acute organizational data [3]. Above all, it is able to enlarge gate of improvement to healthcare, connect pay with performance while helping to limit increasing healthcare costs.

1.2. Problem Statement

Healthcare industry is facing challenges to enhance and improve the quality of their services. With the increased complexity of the data generated by the hospital records, patients’ treatments, and medical history, it is becoming even harder to process this amount of data to deliver new means of decision making. It is becoming a new challenges and opportunity for the healthcare industry to utilized the readily available data to model treatment delivery for new medical cases.

Statistical algorithms can be developed to obtain similarities or relations between patients’ electronic data to actualize data-driven decision making and improved judgement in healthcare. Different algorithms are to be tested to evaluate their functionality and efficiency in providing fast results. By revealing associations and apprehending patterns inside this data, data analytics provides the chance to develop care, rescue lives and cut down on costs.

1.3. Objectives

The objectives here for this study will be as follows:

- To develop a software for MIMIC II data analysis, and related databases.
- To apply big data analytics techniques on the data at hand.
- To test and validate the algorithm and demonstrate visualized output.

1.4. Scope of Study

The scope of study will be covering the following aspects:

- Interpretation of hospital data sets.

  Big data analysis in healthcare industry is becoming a trend in modeling new patient-oriented treatment and medical decision of diagnosis-oriented medical care. The big-data initiative, which combined information collected from patients' medical history, claims, demographics, laboratory results, pharmacy use and patients' self-description of their health status, and performing sets of algorithm on it to create predictive analysis to prevent the diseases and to manage diagnosis, medical care and health condition.

This project will look into analysis on MIMIC II data sets which encompasses of a diverse and very large data of ICU patients containing diagnosis, medication prescription, lab results, electronic documentation, and bedside monitor trends and waveforms.

- Classification of data using R programming language.

  This project focuses on the development of statistical software using R programming languages on precedent clinical data sets. R is a powerful language for comprehensive statistical analysis that incorporates all of the standard statistical tests, models, and analyses, as well as providing the capability to manage and manipulate data. However, this project will focus on predicting pattern and matching trends and relations of the clinical data sets, independent of its timeframe and time series aspect which eliminates the forecasting scope. Forecasting on the clinical data sets, though powerful and advantageous, is unfeasible for the development of the hospital data analysis software in this project.

- Developing an algorithm for data analysis and statistical results.

  Market-basket model is chosen as the algorithm for data analysis to be programmed in the software for this project due to its capability to form associations in the objects extracted from the clinical data sets.

2. CHAPTER 2: LITERATURE REVIEW

2.1. Advantages of Big Data Analytics To Healthcare

Big data within healthcare indicates the massive and multifaceted health datasets that are challenging (nearly impossible) to process using old-fashioned software or hardware; nor simply managed by general data management means and tools.

The size of big data that exists in healthcare is devastating due to the multiplicity of types of data as well as its size, adding to that the speed at which it needs to be processed. The entirety of data linked to patient wellbeing and healthcare events generate “big data” in the healthcare business. It contains clinical information such as physicians’ orders, notes, prescriptions, chart events, and other machine generated data, such as monitoring vital signals [3].

Performing analytics in order to achieve more explained conclusions will show value as well as obtaining clarified results, heading towards up-to-date treatments and tools. Data that leads to insight is able to assist informed and well-educated consumers behave more responsibly towards their health. Analytics is able to develop efficiency and effectiveness. Ranging from small details up to big processes, analytics are able to help examination and findings; develop the delivery of services and operations; improve sustainability; reduce risk; and offer a way to measure and evaluate important structural data. Most important of all, analytics can enlarge access to healthcare and reduce the growth of healthcare costs [4].

Analytics in big data is possible to change the method used by healthcare providers to make decisions out of their clinical and other data sources from complicated technologies to simpler forms. In the future, fast and general applications of big data analytics among healthcare will be noticed. Regarding that matter, several challenges shall be addressed, issues like privacy, security protection, constructing authority as well as standards [5].

2.2. Data Mining

The general meaning of “data mining” is the detection of “models” intended for data. It is also known as obtaining useful info from large datasets, then the analysis of observational data to seek unsuspected associations while reviewing data in new ways to be useful and understandable by data owner [6].

2.3. Association Rule Mining

Association rule learning is a prominent and well researched method for revealing interesting relations between variables in big data analysis. It is directed to classify strong rules discovered in big data using different measures of interestingness.

2.3.1. Finding Similar Items

A general goal in data-mining is to look for “similar” objects or items. An example is finding near-duplicate Web pages by observing a group of pages online. The result can be plagiarisms, or mirrors that contain the same material but different host information. It phrases the similarity problem as in discovering big intersections between data sets. Then, to find textually similar documents it starts with a technique called “shingling” that turns it into a set problem. Hereafter, another method known as “minhashing” is used to compress huge sets to deduct the similarity of the original sets from their compacted versions [6].

A different issue that arises in searching for similar items is the existence of huge amounts of pairs of items to be tested in their level of similarity, even if it is easy to compute the similarity of just one pair. That issue encourages a new technique named “locality-sensitive hashing”, that focuses on pairs that have a high chance of being similar.

2.3.2. The Market-Basket Model

The market-basket classification of data describes a general way of many connections among two types of items. On one side, there are items, and on the other there are baskets, often named “transactions.” Every basket contains an itemset, and it is assumed that the amount of items inside a basket is few – much fewer than the amount of items. The amount of baskets is often presumed to be huge, bigger than the size the main memory can handle. The data shall be represented in a file made up of an order of baskets. In terms of the arranged file, the objects of the file are baskets, and “set of items” is the type of each basket.

Let’s assume the items:

$$I = \{i_1, i_2, i_3, \dots, i_n\}$$

The transactions or basket is a set of items that re-occur together.

$$t_n = \{i_i, i_j, i_k, \dots, i_l\}$$

Rules are statement of the form

$$\{i_1, i_2, \dots\} \longrightarrow \{i_l\}$$

The rules that enable us to analyze the data acquired are basically identified by the support, confidence and the lift of the market basket analysis.

The support of an item or an item set is the probability of finding that item or item set within our data set that includes this specific item or item set. Generally, it is preferable to identify rules which possess high support, as they shall be applied to a relatively big number of transactions or baskets. In hospital analytics it depends on the item set that is compared to the item, as in, if the library of the diseases for the patients tested, the support should be high. If the list of diseases is small, then it is less likely to obtain similarity; therefore the support should be decreased.

$$\text{support } (\text{A}\rightarrow\text{B}) = P(A \cup B)$$

The confidence regarding the rule shall be the possibility of occurrence of a new transaction that contain an item that it is true for a new transaction that contains the items on the LHS of the rule.

$$\text{confidence (A}\rightarrow\text{B) = support } (A \cup B)\text{/ support (A)}$$

The lift is identified as the ratio of the support on the LHS co-occurring with the RHS items then divided by the possibility of both A & B co-occurring with them being independent.

$$\text{lift (A}\rightarrow\text{B) = support } (A \cup B) \text{ / (support(A)} \times \text{support(B)}$$

If the lift happens to be more than 1, it indicates that the existence of items on the LHS did enlarge the possibility the items on the RHS happening in this transaction. If the lift is less than 1, it indicates that the existence of the items on the left hand side shall set the probability of the items on the RHS that will be part of the transaction less. If the lift is equal to 1, it will suggest that the existence of items both, on the LHS and RHS certainly are independent: by understanding that the items on the left hand side are existing adds no change to the possibility that items will happen for the RHS.

On performing market basket analysis, we are searching for rules that have a lift greater than one. Rules with greater confidence are rules such the possibility of an item appearing on the right hand side is high specified the existence of the items on the left hand side. It is better (higher value) to execute rules that have greater support - as they will be relevant to a bigger amount of transactions.

2.4. The MIMIC II Database

The Multi-parameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database contains thorough material for patients who stay in the intensive care unit, this database includes laboratory data, therapy involvement data as vasoactive medication input amounts and ventilator preferences, nursing action notes, discharge summaries, reports of radiology, entry data of the provider, International Classification of Diseases, 9th

Revision codes, as well as, for a set of patients, high-resolution trends for vital signs and waveforms of 25,328 intensive care unit patient stays [9]. This project will look into analysis on MIMIC II data sets.

Although other clinical research databases exist, such databases are often privately owned and have highly restricted access or require fees for access. MIMIC-II has been fully de-identified in a Health Insurance Portability and Accountability Act (HIPAA) compliant manner and is available free of charge for public use, subject to completion of an appropriate online human-subjects training course and signing of a data use agreement.

The database, which was made freely available in February 2010, is available via PhysioNet, a web-based resource for the study of physiologic data along with a detailed user's guide and a collection of data processing tools [10]. It establishes a new public-access resource for critical care research, supporting a diverse range of analytic studies spanning epidemiology, clinical decision-rule development, and electronic tool development.

2.4.1. Data Mining on MIMIC II Database

The large-scale ICU databases, MIMIC II have been effective resources to understand risk factors and natural histories of critical illness as well as the efficacy of various treatment strategies. Several data mining project has been done by researches on this databases which shall be the benchmark for the development of the software in this project. Some notable data mining project includes,

- Computer-assisted de-identification of free text in the MIMIC II database [11].

An evaluation of methods for computer-assisted removal and replacement of protected health information (PHI) from free-text nursing notes collected in the intensive care unit as part of the MIMIC II project. This study develops a semi-automated method to allow clinicians to highlight PHI on the screen of a tablet PC and to compare and combine the selections of different experts reading the same notes.

- Estimating cardiac output from arterial blood pressure waveforms: a critical evaluation using the MIMIC II database [12].

An evaluation of 11 well-known Cardiac Output (CO) estimators using clinical radial ABP waveforms from the multi-parameter intelligent monitoring for intensive care II (MIMIC II) database, using thermodilution CO (TCO) as reference for comparison.

- Predictive Value of Ionized Calcium in Critically Ill Patients: An Analysis of a Large Clinical Database MIMIC II [13].

This research studies the association of Ionized Calcium (iCa) with mortality by using MIMIC II database. In critical illness, heart failure and hyperadrenergic states are the most commonly seen disorders that have been proven to be associated with calcium derangements. iCa measurements in multiple MIMIC II patients' ICU stays records was used to establishes said association.

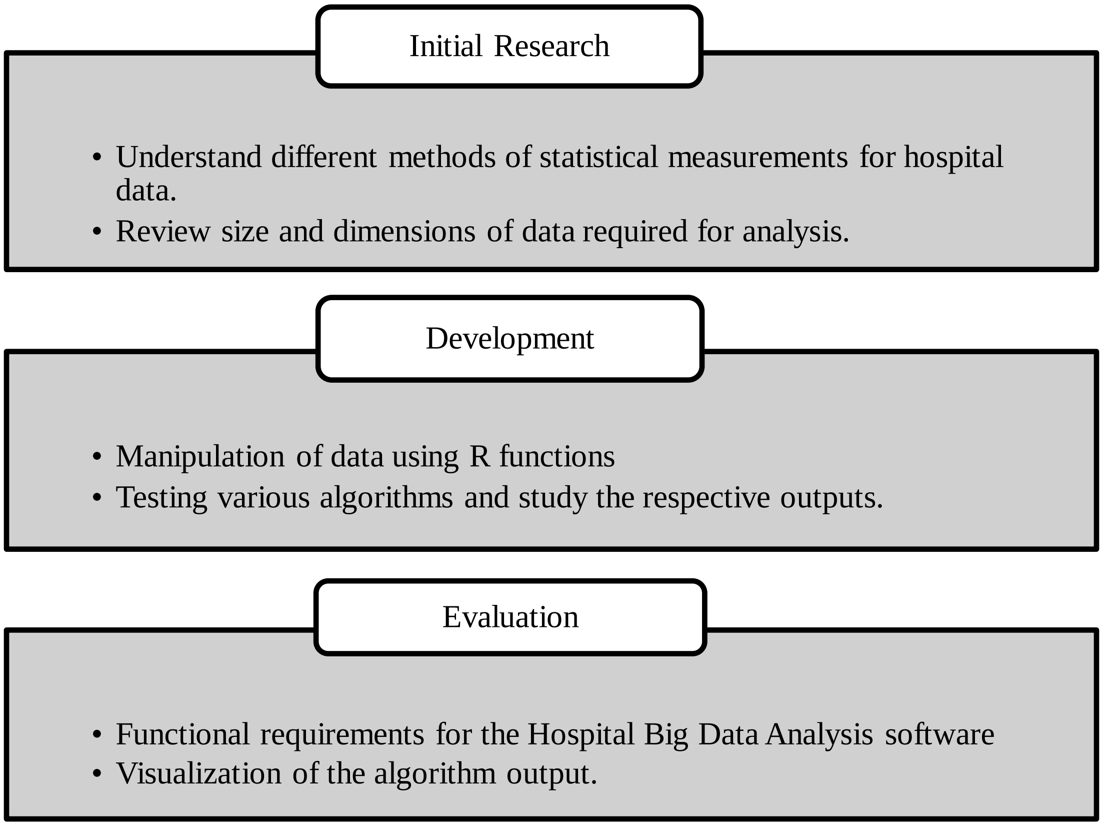
- Accessing the public MIMIC-II intensive care relational database for clinical research [10].

This paper presents the two major software tools that facilitate accessing the relational database: the web-based QueryBuilder and a downloadable virtual machine (VM) image. QueryBuilder and the MIMIC-II VM have been developed successfully and are freely available to MIMIC-II users.

3. CHAPTER 3: METHODOLOGY

3.1. Research Methodology

This study will be based on developing an algorithm to detect similarities or associations within the elements of the present hospital data. Starting with a simple software that can detect similarities within given sample data, a software will be developed that is able to analyse large datasets that stream every day from hospitals, and be able to present facts about that given data. The algorithm development will include presenting visual results to help simple understanding of the results. These algorithms can be integrated in R software that is able to aid modern healthcare analysis. The flow of the project study can be summarized in figure 1 below.



Initial Research

- Understand different methods of statistical measurements for hospital data.
- Review size and dimensions of data required for analysis.

Development

- Manipulation of data using R functions
- Testing various algorithms and study the respective outputs.

Evaluation

- Functional requirements for the Hospital Big Data Analysis software
- Visualization of the algorithm output.

Figure 1: Flow of the project

3.1.1. Initial Research

- **Understand different methods of statistical measurements for hospital data.**

The primary focus of this project is to understand different methods of statistical measurements for hospital data. Explosive growth in the generation and collection of hospital data has created abundance of opportunities for research and statistical analysis in healthcare industry.

Prior to development of the hospital big data analysis software, understanding of the subject matter and its data types is crucial in order to decide on the method of statistical analysis to be performed. The diversity of data types in MIMIC II supports the development and evaluation of automated detection, prediction, and estimation algorithms [15]. Having high temporal resolution and being multiparameter in nature made MIMIC II data suitable for developing clinically useful and robust algorithms [9, 15]. MIMIC II is selected for this study due to its ability to simulate a real-life ICU in offline mode that enables inexpensive evaluation of developed algorithms without the risk of disturbing patients and clinical staff.

- **Review size and dimensions of data required for analysis.**

The MIMIC-II is a massive intensive care unit (ICU) research database with a high-resolution diagnostic and therapeutic data from a large, diverse population of adult ICU patients. Thus, initial research of this project includes reviewing the size and dimensions of MIMIC II data required for analysis. Sets of less than five (5) patients per algorithms run were decided for this study to ensure efficient query processing and memory management. Extractions of the required data within each MIMIC II patient ICU stay logs were done to avoid heavy processing and to allow faster results generation.

3.1.2. Development

- **Development with R programming language**

R is an open source programming language and software environment for statistical computing and graphics. Widely used among statisticians and data miners for developing statistical software [16][17] and data analysis, the R language includes virtually every data manipulation, statistical model, and chart that the modern data scientist could ever need. R has become the most popular language for data science and an essential tool for Finance and analytics-driven companies such as Google, Facebook, and LinkedIn.

In this section, the viability of developing the hospital big data analysis software using R programming language is addressed.

3.1.3. Evaluation

- **Testing and visualization of the algorithm**

The testing and visualization of the algorithm is done using R studio, an integrated development environment (IDE) for the R programming language. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. The visualization of results from the algorithm run is display in the console and plot section of the R studio. These visualizations are provided within this paper in Chapter 4, Results and Discussions.

- **Functional Requirements of the Hospital Big Data Analysis software**

Function 1	The software shall provide data import capabilities
Description	The Hospital Big Data Analysis software should be able to perform data import for the MIMIC II datasets and its tag dictionary.

Function 2	The software shall include data extraction capabilities
Description	The software must be able to extract all the data to be analyse and formed in into the proper data structure.

Function 3	The software shall enable user to run algorithms for data analyzation
Description	The software should be able to perform sets of analytical algorithm on the MIMIC II datasets.

Function 4	The software shall provide results visualization capabilities
Description	The software shall output the results of the algorithms run to the user on the console and provide graphical representation of the performed output.

Table 1 : Functional Requirements of the Hospital Big Data Analysis software

3.2. Gantt Chart

The Gantt charts for Final Year Project I and Final Year Project II are shown in Figures 2 and 3 below:

Week	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Selection of project title	🟩	🟩	🟩	🟩										
Initial research of work				🟩	🟩	🟩								
Extended Proposal Submission							🟩							
Proposal Defence								🟩						
Continuation of research							🟩	🟩	🟩	🟩				
Data interpretation and function testing.											🟩	🟩	🟩	
Submission of interim report														🟩

Figure 2: Gantt chart for FYP I

Week	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Interpreting Data	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥						
Progress Report Submission								🟥						
Running Algorithms									🟥	🟥	🟥	🟥		
Pre-SEDEX											🟥			
SEDEX												🟥		
Dissertation Report Submission (Soft Copy)													🟥	
Viva														🟥
Dissertation Report Submission “Hard Copy”														🟥

Figure 3: Gantt chart for FYP II

3.3. Key Milestones

Key milestones completed so far:

• Project Research kick-off	Week 4 (FYP I)
• Extended proposal submission	Week 7 (FYP I)
• Proposal defence	Week 8 (FYP I)
• Interim report submission	Week 14 (FYP I)
• Progress Report	Week 8 (FYP II)

Key milestones to be completed:

• Pre-SEDEX	Week 10 (FYP II)
• Obtaining Visual Results and Debugging	Week 11 (FYP II)
• Dissertation Report Submission	Week 12 (FYP II)
• Viva	Week 14 (FYP II)

3.4. Project Work

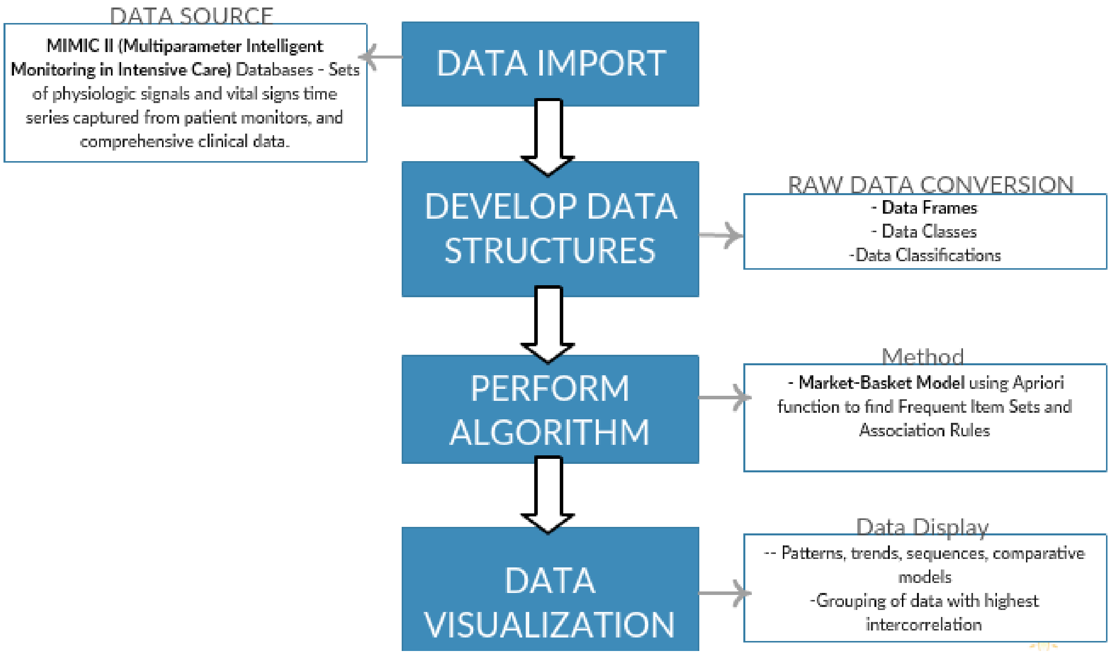
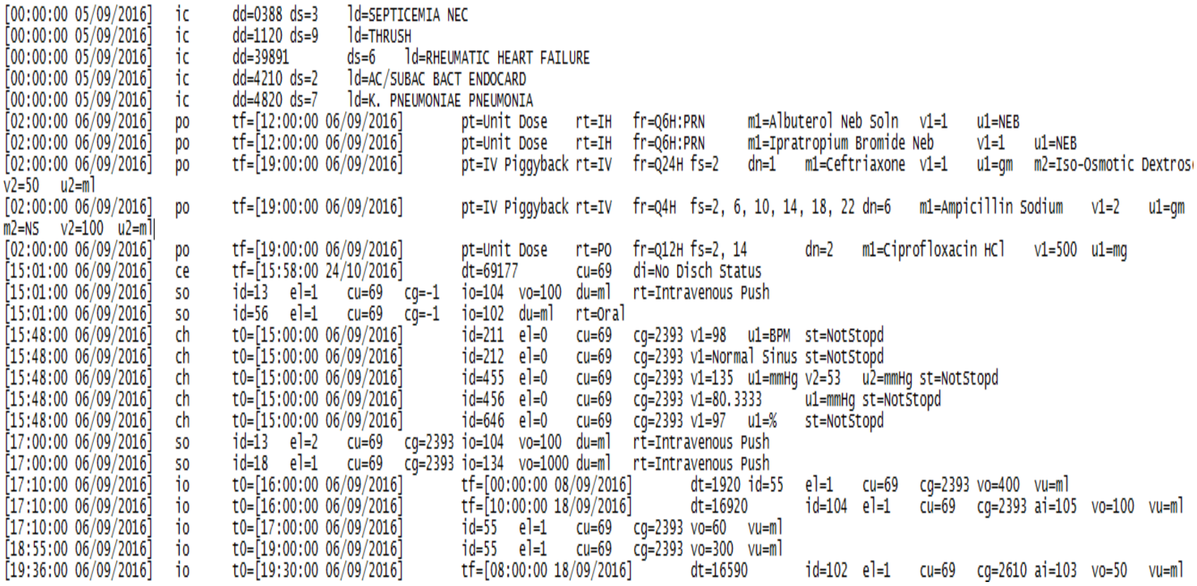

DATA SOURCE
MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care) Databases - Sets of physiologic signals and vital signs time series captured from patient monitors, and comprehensive clinical data.
DATA IMPORT
DEVELOP DATA STRUCTURES
RAW DATA CONVERSION
- Data Frames
- Data Classes
-Data Classifications
PERFORM ALGORITHM
Method
- Market-Basket Model using Apriori function to find Frequent Item Sets and Association Rules
DATA VISUALIZATION
Data Display
-- Patterns, trends, sequences, comparative models
-Grouping of data with highest intercorrelation

Figure 4 : Project Work Methodology for Software Development

3.4.1. Data Import & Extraction

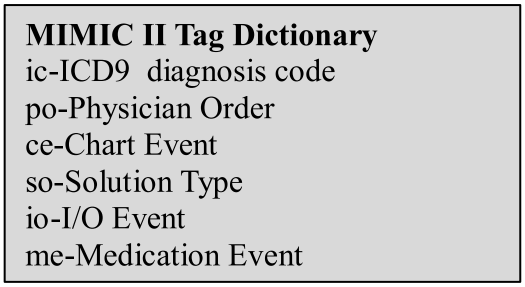
Using the physionet download URL, it has been possible to download the MIMIC II database for a 100 patients, each patient has a record of events and incidents as well as the diagnostics and treatment given, all in plain text file. A sample is given in the following figure:



```
[00:00:00 05/09/2016]   ic      dd=0388 ds=3    ld=SEPTICEMIA NEC
[00:00:00 05/09/2016]   ic      dd=1120 ds=9    ld=THRUSH
[00:00:00 05/09/2016]   ic      dd=39891        ds=6    ld=RHEUMATIC HEART FAILURE
[00:00:00 05/09/2016]   ic      dd=4210 ds=2    ld=AC/SUBAC BACT ENDOCARD
[00:00:00 05/09/2016]   ic      dd=4820 ds=7    ld=K. PNEUMONIAE PNEUMONIA
[02:00:00 06/09/2016]   po      tf=[12:00:00 06/09/2016]        pt=Unit Dose    rt=IH   fr=Q6H:PRN      m1=Albuterol Neb Soln   v1=1    u1=NEB
[02:00:00 06/09/2016]   po      tf=[12:00:00 06/09/2016]        pt=Unit Dose    rt=IH   fr=Q6H:PRN      m1=Ipratropium Bromide Neb      v1=1    u1=NEB
[02:00:00 06/09/2016]   po      tf=[19:00:00 06/09/2016]        pt=IV Piggyback rt=IV   fr=Q24H fs=2    dn=1    m1=Ceftriaxone  v1=1    u1=gm   m2=Iso-Osmotic Dextros
v2=50   u2=ml
[02:00:00 06/09/2016]   po      tf=[19:00:00 06/09/2016]        pt=IV Piggyback rt=IV   fr=Q4H  fs=2, 6, 10, 14, 18, 22 dn=6    m1=Ampicillin Sodium    v1=2    u1=gm
m2=NS   v2=100  u2=ml
[02:00:00 06/09/2016]   po      tf=[19:00:00 06/09/2016]        pt=Unit Dose    rt=PO   fr=Q12H fs=2, 14        dn=2    m1=Ciprofloxacin HCl    v1=500  u1=mg
[15:01:00 06/09/2016]   ce      tf=[15:58:00 24/10/2016]        dt=69177        cu=69   di=No Disch Status
[15:01:00 06/09/2016]   so      id=13   el=1    cu=69   cg=-1   io=104  vo=100  du=ml   rt=Intravenous Push
[15:01:00 06/09/2016]   so      id=56   el=1    cu=69   cg=-1   io=102  du=ml   rt=Oral
[15:48:00 06/09/2016]   ch      t0=[15:00:00 06/09/2016]        id=211  el=0    cu=69   cg=2393 v1=98   u1=BPM  st=NotStopd
[15:48:00 06/09/2016]   ch      t0=[15:00:00 06/09/2016]        id=212  el=0    cu=69   cg=2393 v1=Normal Sinus st=NotStopd
[15:48:00 06/09/2016]   ch      t0=[15:00:00 06/09/2016]        id=455  el=0    cu=69   cg=2393 v1=135  u1=mmHg v2=53   u2=mmHg st=NotStopd
[15:48:00 06/09/2016]   ch      t0=[15:00:00 06/09/2016]        id=456  el=0    cu=69   cg=2393 v1=80.3333      u1=mmHg st=NotStopd
[15:48:00 06/09/2016]   ch      t0=[15:00:00 06/09/2016]        id=646  el=0    cu=69   cg=2393 v1=97   u1=%    st=NotStopd
[17:00:00 06/09/2016]   so      id=13   el=2    cu=69   cg=2393 io=104  vo=100  du=ml   rt=Intravenous Push
[17:00:00 06/09/2016]   so      id=18   el=1    cu=69   cg=2393 io=134  vo=1000 du=ml   rt=Intravenous Push
[17:10:00 06/09/2016]   io      t0=[16:00:00 06/09/2016]        tf=[00:00:00 08/09/2016]        dt=1920 id=55   el=1    cu=69   cg=2393 vo=400  vu=ml
[17:10:00 06/09/2016]   io      t0=[16:00:00 06/09/2016]        tf=[10:00:00 18/09/2016]        dt=16920        id=104  el=1    cu=69   cg=2393 ai=105  vo=100  vu=ml
[17:10:00 06/09/2016]   io      t0=[17:00:00 06/09/2016]        id=55   el=1    cu=69   cg=2393 vo=60   vu=ml
[18:55:00 06/09/2016]   io      t0=[19:00:00 06/09/2016]        id=55   el=1    cu=69   cg=2393 vo=300  vu=ml
[19:36:00 06/09/2016]   io      t0=[19:30:00 06/09/2016]        tf=[08:00:00 18/09/2016]        dt=16590        id=102  el=1    cu=69   cg=2610 ai=103  vo=50   vu=ml
```

Figure 5: Data Sample

In the above figure, is shown a sample of the data obtained. The information is divided by tab "/t" which make it easier to import to programming languages. The first column describes the time stamp at which the events take place. The two characters in the second column are a source code; it identifies the entry, whether its information in the patient's ICU chart, a laboratory test, or a physician's order. The rest of the columns explain the information for measurements taken, ICD 9, or name and quantity of medication. Figure below outlines the tag dictionary of MIMIC II data sets which will be used in this project.



MIMIC II Tag Dictionary
ic-ICD9 diagnosis code
po-Physician Order
ce-Chart Event
so-Solution Type
io-I/O Event
me-Medication Event

Figure 6 : MIMIC II Tag Dictionary

With such a huge amount of clinical data inside a single record of MIMIC II patients' ICU stay, the next question would be how to conceive clinical research ideas. The type of clinical research using big data that is chosen for this project is the Prediction Model which aims to fit a model for pattern matching and future predictions.

This study investigate the association of different data entry in the MIMIC II data sets of 10 different patients' ICU stay with the goal to explore the diagnosis, physician order, chart event, I/O event and medication event matches between these patients data sets. Performance of statistical description for these datasets, by using traditional central tendency (mean, median) and discrete tendency (full range, 95% confidence interval) shall provide graphical demonstration of the event matches. Contour plot help to explore associations between these variables. The study protocol was adapted accordingly as the functional requirement of the Hospital Big Data Analysis software.

Considering the basis that the software to be developed is a small scale analysis software, this project investigate simple and easily obtainable parameters within the MIMIC II sheets. Hence, only a collection of selective variables are extracted from the data sets to run the programmed algorithm on.

3.4.1.1. Data Import

Data import functionalities in R uses the function read.*file-type* within the *utils* package in R. *utils* package is preinstalled in R software and resided in R system library. Using the code below the selected MIMIC II data sets will be imported as a single data frame name patient*n* (*n* depicting patient MIMIC II data sets 1,2….). The tag dictionary which consist of ID and label for chart events, input/output events and medicine events is only imported using the same method as per depict below.

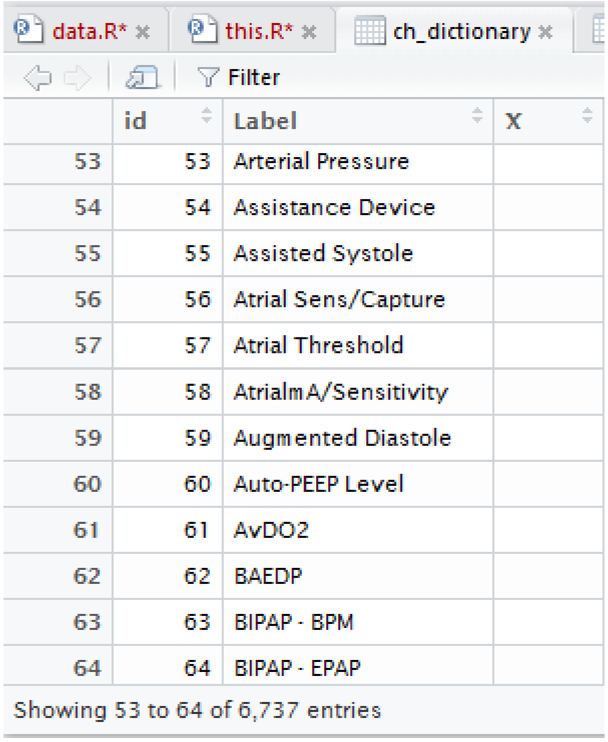
```
library(utils)
#Import MIMIC II Patient's ICU Stay Data
patient1 <- read.csv("C:/Users/Desktop/CSV/1.csv", header=FALSE)
View(patient1)
patient2 <- read.csv("C:/Users/Desktop/CSV/2.csv", header=FALSE)
View(patient2)
patient3 <- read.csv("C:/Users/Desktop/CSV/3.csv", header=FALSE)
View(patient3)
patient4 <- read.csv("C:/Users/Desktop/CSV/4.csv", header=FALSE)
View(patient4)

#Import Chart Event(ch), Input/Output(io), & Medication(me) Dictionary
ch_dictionary <- read.csv("C:/Users/Desktop/Data acquire/tag
dictionary/chdic.csv")
io_dictionary <- read.csv("C:/Users/Desktop/Data acquire/tag
dictionary/iodic.csv", header=FALSE)
names(io_dictionary) <- c("id","Label","X")
me_dictionary <- read.csv("C:/Users/Desktop/Data acquire/tag
dictionary/medic.csv", header=FALSE)
names(me_dictionary) <- c("id","Label","X")
```

3.4.1.2. Data Extraction

Since the imported data sets comprises of a very large data records for a single patient – one MIMIC II sheets can contains thousands of entry lines, data extraction is vital in order to extract only the needed variable for analysis by the developed software. Using the R function *subset* in the *arules* package installed under user library that provides function for mining association rules and frequent itemsets, the code below extract entries in MIMIC II data sets, namely, diagnosis, chart event, I/O event, medication event and physician order event. Several user defined functions namely, `get_ch()`, `get_io()`, `get_me()` and `get_po()` were created to extract the entries from MIMIC II datasets.

```
install.packages("arules")
library(arules)

#Filter patient record to extract ic,ch,io,me,po
#Extract diagnosis data record
get_ic <- function(patient){

  names(patient)<- c("Timeframe","Entry")
  data_diagnosis <- subset(patient, Entry == "ic")

  names(data_diagnosis)<- c("Timeframe","Entry","ID")
  ic <- unique(data_diagnosis[,"ID", drop=FALSE])

  ic <- as.data.frame(sapply(ic,gsub,pattern="dd=",replacement=""))

  return (ic)
}
```

MIMIC II data sets' variables come in the form of unique 'ID' which will then be translated into layman terms from the MIMIC II data sets tag dictionary which has been imported as per previous section with the developed software. Figure below shows an excerpt of chart event tag dictionary.



	id	Label	X
53	53	Arterial Pressure	
54	54	Assistance Device	
55	55	Assisted Systole	
56	56	Atrial Sens/Capture	
57	57	Atrial Threshold	
58	58	AtrialmA/Sensitivity	
59	59	Augmented Diastole	
60	60	Auto-PEEP Level	
61	61	AvDO2	
62	62	BAEDP	
63	63	BIPAP - BPM	
64	64	BIPAP - EPAP	

Showing 53 to 64 of 6,737 entries

Figure 7 : Chart Event (ch) of MIMIC II Data Sets Tag Dictionary

3.4.2. Development of Data Structures

In order to run analytical functions for mining association rules and frequent itemsets, the extracted data from MIMIC II data sets must be converted and built in a form of transactions data frame. Once the data has been coerced to transactions the data is ready for mining itemsets or rules. Association Rule Learning uses the transaction data files available in R.

The code excerpt below depicts the required packages and libraries for the development of transactions data structure for data mining in the software. Sample given is the development of transaction data structure, `patientSet_ch` through user defined function `get_patientset_ch()` which contains the sets of chart event (ch) entries of four (4) different MIMIC II patients' ICU stay data sets.

`ch`*`n`* (*n* depicting patient MIMIC II data sets 1,2….) contains the extraction of only the 'ID' of chart event entries in the patient record excluding its time frame and other details. Due to the fact that multiple entries of the same event varies in time intervals existed within a single MIMIC II data sets, the function *unique* from *arules* package is run onto `ch`*`n`* to remove redundancy of event ID.

`patientSet_ch` is then created containing the list of chart event ID for four(4) different patient record. This list is then coerced as transaction data type as depicted below. The same methodology is done on the other entries of the MIMIC II datasets.

```
install.packages("plyr")
install.packages("matrix")
library(arules)
library(plyr)
library(matrix)
library(stats)
library(graphic)
library(base)

#Get patients chart events (ch) transactions data set
get_patientSet_ch <- function (){

ch1 <- as.matrix(get_ch(patient1))
ch2 <- as.matrix(get_ch(patient2))
ch3 <- as.matrix(get_ch(patient3))
ch4 <- as.matrix(get_ch(patient4))

ch1 <- unname(ch1)
ch2 <- unname(ch2)
ch3 <- unname(ch3)
ch4 <- unname(ch4)

ch1 <- unlist(ch1)
ch2 <- unlist(ch2)
ch3 <- unlist(ch3)
ch4 <- unlist(ch4)

patientSet_ch <- list(ch1,ch2,ch3,ch4)

## set transaction names
names(patientSet_ch) <- paste("Tr",c(1:4), sep = "")
patientSet_ch

trans <- as(patientSet_ch, "transactions")

return (trans)
}
```

3.4.3. Development and Application of Algorithm

After acquiring the data needed in the type required, manipulation is essential in order to compare certain vectors inside the dataset we have. Each algorithm function in R has its own arguments and a certain outcome. To be able to use these functions properly, a wide variation of packages shall be installed and learnt throughout the project.

3.4.3.1. Equivalence Class Transformation

Equivalence Class Transformation (Eclat) algorithm can be used to do mining on itemsets. This allows the user to find repeated patterns within the data. These patterns are included in the association rules and are applied in different application fields.

Basically, Eclat algorithm uses tidset (transaction identifier itemset) intersections then computes the support of an itemset without generating the subsets not included in the pre-set tree.

The Eclat algorithm is explained as a recursive function. The first run computes the single items and their tidsets. In every recursive run, each itemset-tidset pair is verified {X,t(X)} next to the other pairs {Y,t(Y)} to produce candidates:N_{XY}, if it is found that the candidate is frequent, it is added to P_X set. Then the function recursively finds in the X branch all the frequent itemsets.

The Eclat model to run on the variables between different patients is programmed into the software as below.

```
install.packages(“arules”)
install.packages(“arulesViz”)
library(arules)
library(arulesViz)

#function get_eclat is built in order to run different entries datasets
on function call

get_eclat <- function(patientSet_data){

Results <- eclat(patientSet_data, parameter = list(supp = 0.5))

Results.top5 <- sort(Results)[1:5]
inspect(Results.top5)
entry_item <- as(items(Results.top5), "list")

names(entry_item)<- c("1","2","3","4","5")

return (entry_item)

}

#get sets of ic,ch,io & me entries for 4 different patients
patientSet_ic <- get_patientSet_ic()
patientSet_ch <- get_patientSet_ch()
patientSet_io <- get_patientSet_io()
patientSet_me <- get_patientSet_me()

#eclat run on sets of patients medication event (me)
me_item<-as.data.frame(get_eclat(patientSet_me))
```

3.4.3.2. A-Priori

The A-Priori Algorithm is developed to decrease the amount of pairs that has to be counted, with the action of doing two passes on the data, not just one pass. Schematics are illustrated in figure 5.

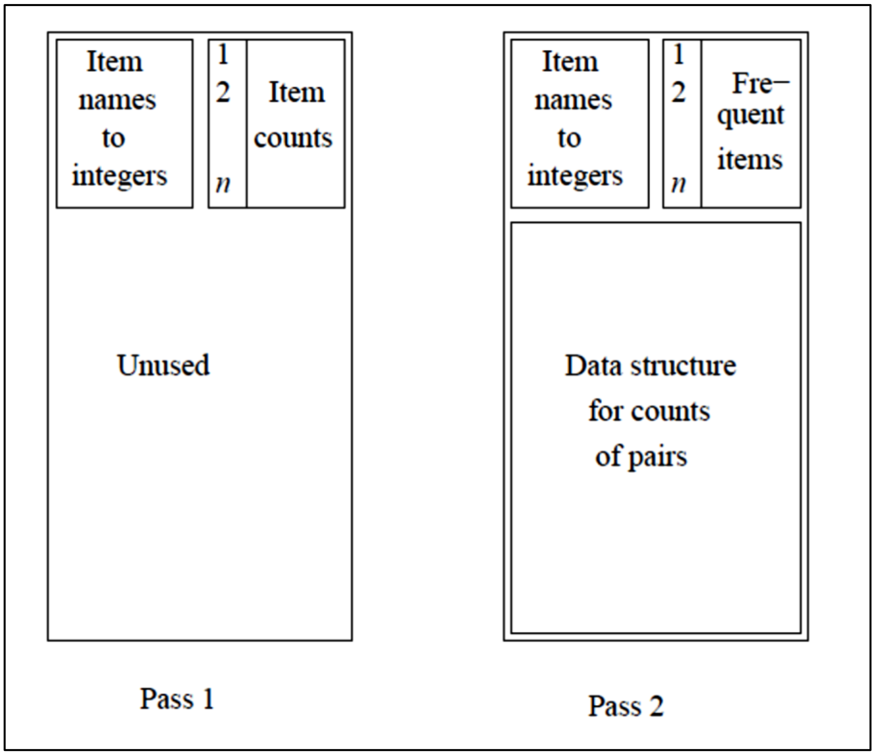


Figure 8: Schematic of main-memory use during the two passes of the A-Priori Algorithm

In the first pass, two tables are created. Item names are translated into integers if needed in the first table. The other table counts the occurrences of the items in the data. Then counts are examined to determine any frequent singletons. In order not to get too many frequent sets, the support *s* should be 1%. For the next pass, frequent items are counted in numberings and stored in the frequents-items table.

In the second pass, all pairs that consist of two frequent items are counted; a pair cannot be frequent unless both members are frequent.

The apriori model to run on the variables between different patients is programmed into the software as below.

```
install.packages(“arules”)
install.packages(“arulesViz”)
library(arules)
library(arulesViz)

#function get_apriori is built in order to run different entries
datasets on function call

get_apriori <- function(patientSet_data){

Results <- apriori(patientSet_data, parameter = list(supp = 0.5,
conf = 0.9, target = "rules"))
summary(Results)

itemsets <- unique(generatingItemsets(Results))
itemsets.df <- as(itemsets, "data.frame")
frequentItemsets <- itemsets.df[with(itemsets.df, order(-
support,items)),]
names(frequentItemsets)[1] <- "itemset"
write.table(frequentItemsets, file = "", sep = ",", row.names =
FALSE)

frequentItemsets

return (Results)

}

#apriori run on sets of patients medication event (me)
apriori_me <- get_apriori(patientSet_me)

Summary(apriori_me)
```

3.4.4. Data Visualization

- Graphical plotting of entries data and its frequency.

```
n <- count(data_chart[,”ID”, drop=FALSE],ID)
as.data.frame(n)

data <- filter(data_chart, n)
plot(data$ID, data$n,
     xlab=”Event ID”,
     ylab=”Frequency”)
```

```
plot(Results, measure=c("support","confidence"), shading="lift")

itemFrequencyPlot(trans, support = 0.1);
```

- Obtaining ICD-9 diagnosis name for diagnosis entries matches between patients.

```
link<-“http://www.hipaaspace.com/Medical_Billing/Coding/ICD-
9/Txt/Diagnosis/”

for(I in 1:50){
  dd<-grepl(‘dd’,x[I,4])
  if(dd == TRUE){
    print(sub(‘dd*\\=’, “”, x[I,4]))

    diagnosis<-sub(‘dd*\\=’, “”, x[I,4])
    link<- paste(s1,diagnosis, sep =””)
    print(link)
  }

  url<-html(link)
  selector_name<-“h1”

  fnames<-html_nodes(url, selector_name) %>%
    html_text()

  head(fnames)
```

- Obtaining entries term from MIMIC II tag dictionary

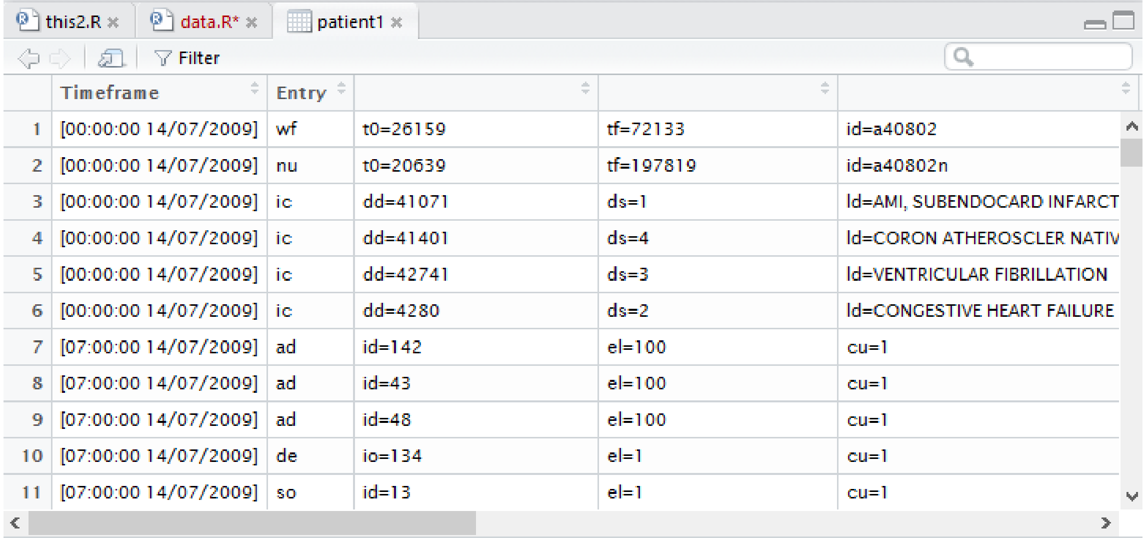
```
me_item<-as.data.frame(get_eclat(patientSet_me))

for(n in 1:5){
  me_id = as.character(me_item[1,n])
  print(subset(me_dictionary, id == me_id , select=c(id, Label)))
}
```

4. CHAPTER 4: RESULTS & DISCUSSION

4.1. Hospital Big Data Analysis Software

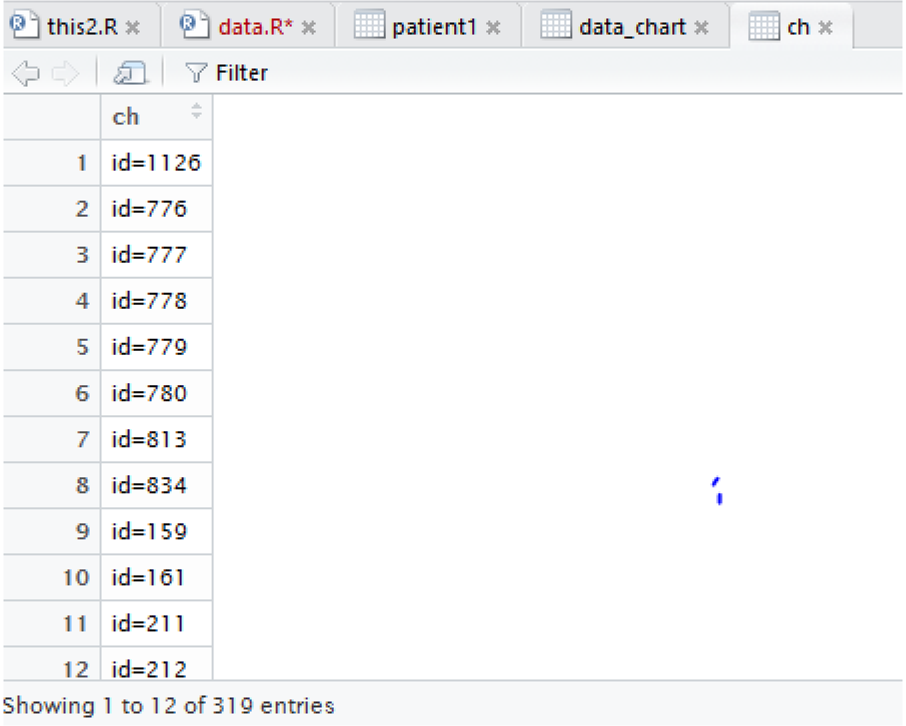
The Hospital Big Data Analysis software was developed with the capability to import, extract, formed data frames and run analysis functions on MIMIC II data sets. The output screenshots of the software run is provided in the sections below.

4.2. Data Extraction



	Timeframe	Entry			
1	[00:00:00 14/07/2009]	wf	t0=26159	tf=72133	id=a40802
2	[00:00:00 14/07/2009]	nu	t0=20639	tf=197819	id=a40802n
3	[00:00:00 14/07/2009]	ic	dd=41071	ds=1	Id=AMI, SUBENDOCARD INFARCT
4	[00:00:00 14/07/2009]	ic	dd=41401	ds=4	Id=CORON ATHEROSCLER NATIV
5	[00:00:00 14/07/2009]	ic	dd=42741	ds=3	Id=VENTRICULAR FIBRILLATION
6	[00:00:00 14/07/2009]	ic	dd=4280	ds=2	Id=CONGESTIVE HEART FAILURE
7	[07:00:00 14/07/2009]	ad	id=142	el=100	cu=1
8	[07:00:00 14/07/2009]	ad	id=43	el=100	cu=1
9	[07:00:00 14/07/2009]	ad	id=48	el=100	cu=1
10	[07:00:00 14/07/2009]	de	io=134	el=1	cu=1
11	[07:00:00 14/07/2009]	so	id=13	el=1	cu=1

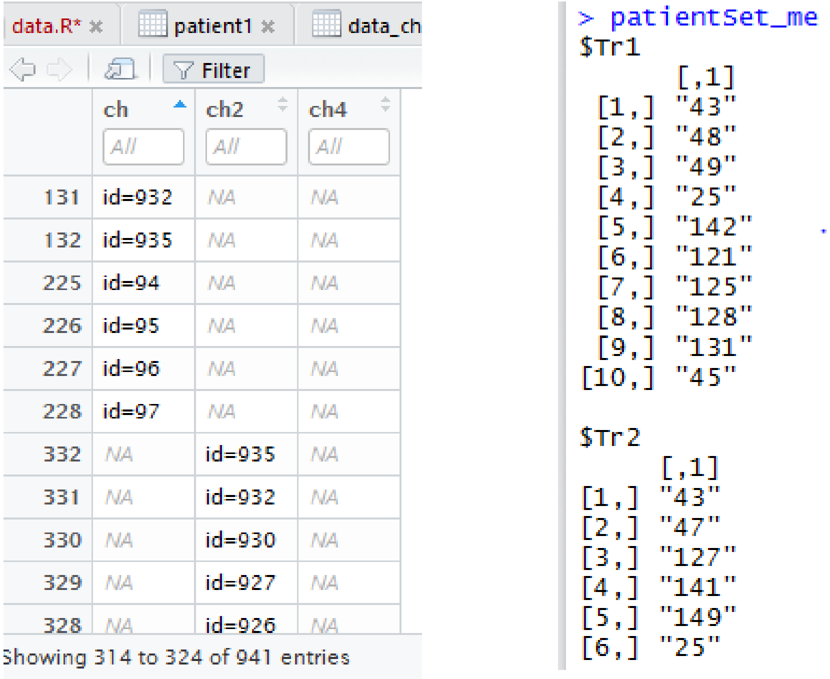
Figure 9: Data Import of MIMIC II patients' ICU Stay Data Sets



	ch
1	id=1126
2	id=776
3	id=777
4	id=778
5	id=779
6	id=780
7	id=813
8	id=834
9	id=159
10	id=161
11	id=211
12	id=212

Showing 1 to 12 of 319 entries

Figure 10 : Chart Event (ch) 'ID' extraction for a single patient data sets



	ch	ch2	ch4
	All	All	All
131	id=932	NA	NA
132	id=935	NA	NA
225	id=94	NA	NA
226	id=95	NA	NA
227	id=96	NA	NA
228	id=97	NA	NA
332	NA	id=935	NA
331	NA	id=932	NA
330	NA	id=930	NA
329	NA	id=927	NA
328	NA	id=926	NA

Showing 314 to 324 of 941 entries

```
> patientSet_me
$Tr1
      [,1] 
 [1,] "43" 
 [2,] "48" 
 [3,] "49" 
 [4,] "25" 
 [5,] "142"
 [6,] "121"
 [7,] "125"
 [8,] "128"
 [9,] "131"
[10,] "45" 

$Tr2
     [,1] 
[1,] "43" 
[2,] "47" 
[3,] "127"
[4,] "141"
[5,] "149"
[6,] "25" 
```

Figure 11: Data Import of MIMIC II patients' ICU Stay Data Sets based on Entry Type

4.3. Data Visualization

4.3.1. Algorithm Visualized Output

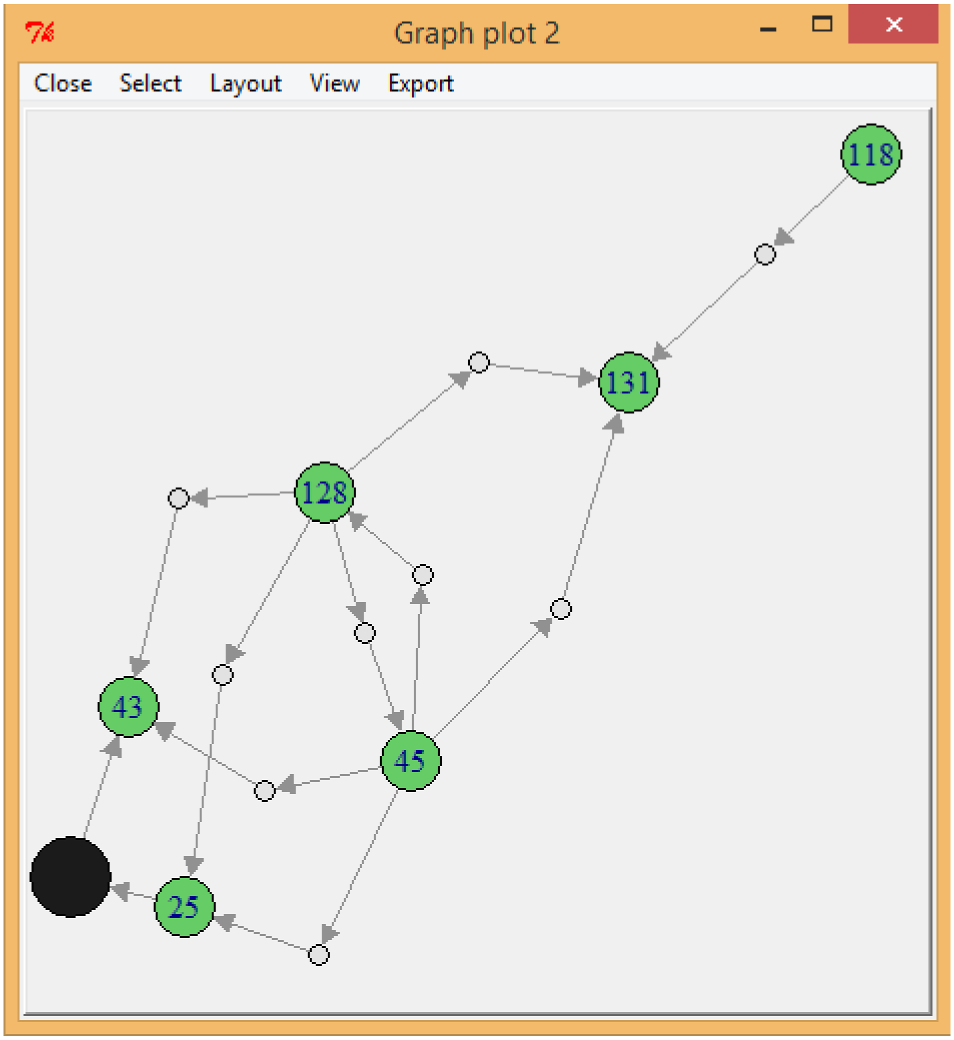


Figure 12: Item frequency plotting from Apriori run on Medication Event (me) entries

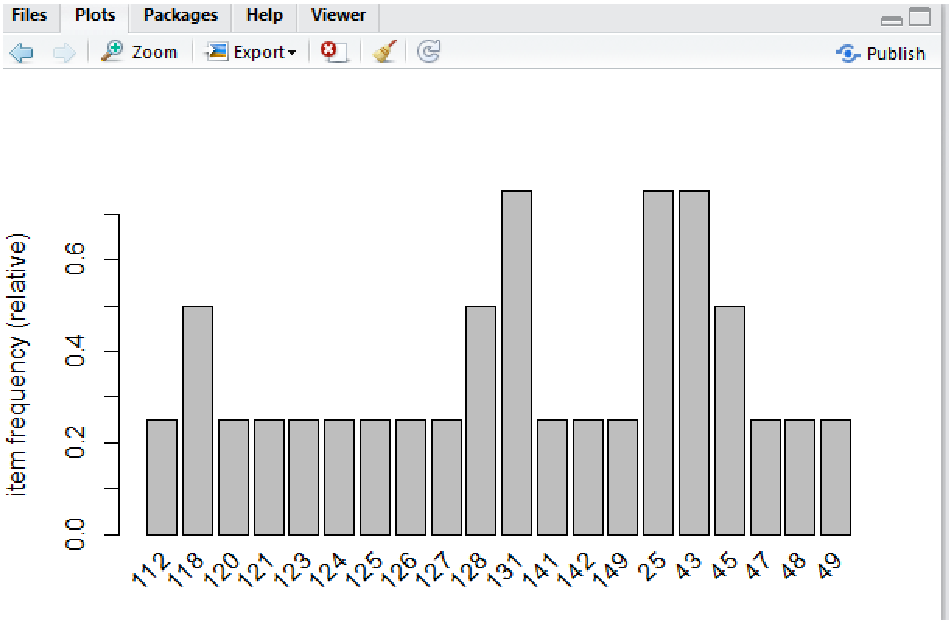
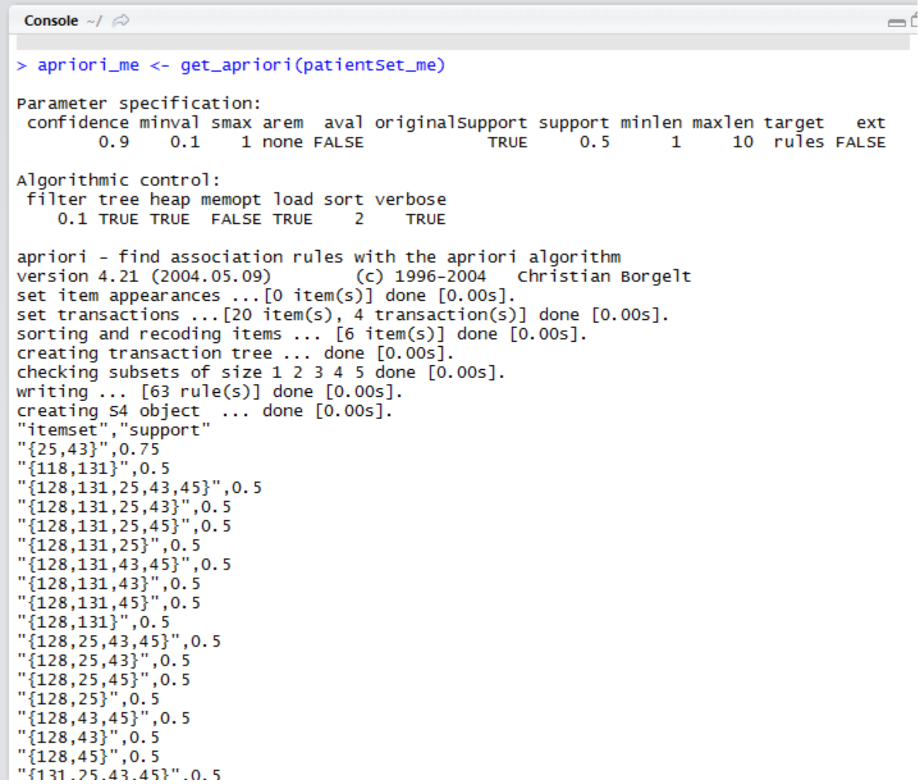


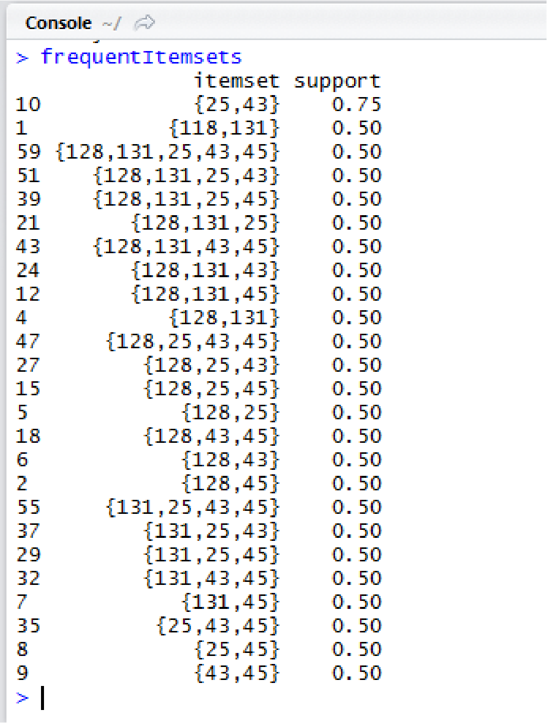
Figure 13: Item frequency plotting from Eclat run on Medication Event (me) entries

4.3.2. Apriori algorithm



```
> apriori_me <- get_apriori(patientSet_me)

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.9    0.1    1 none FALSE            TRUE     0.5      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[20 item(s), 4 transaction(s)] done [0.00s].
sorting and recoding items ... [6 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [63 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
"itemset","support"
"{25,43}",0.75
"{118,131}",0.5
"{128,131,25,43,45}",0.5
"{128,131,25,43}",0.5
"{128,131,25,45}",0.5
"{128,131,25}",0.5
"{128,131,43,45}",0.5
"{128,131,43}",0.5
"{128,131,45}",0.5
"{128,131}",0.5
"{128,25,43,45}",0.5
"{128,25,43}",0.5
"{128,25,45}",0.5
"{128,25}",0.5
"{128,43,45}",0.5
"{128,43}",0.5
"{128,45}",0.5
"{131,25,43,45}",0.5
```

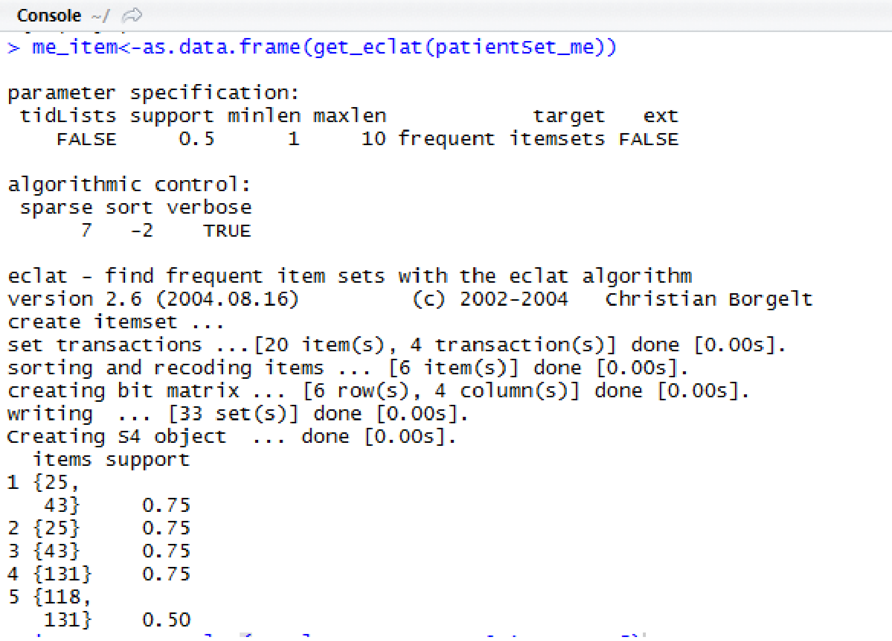
Figure 14: Apriori algoritm results



```
> frequentItemsets
                itemset support
10              {25,43}    0.75
1              {118,131}   0.50
59 {128,131,25,43,45}      0.50
51    {128,131,25,43}      0.50
39    {128,131,25,45}      0.50
21       {128,131,25}      0.50
43    {128,131,43,45}      0.50
24       {128,131,43}      0.50
12       {128,131,45}      0.50
4           {128,131}      0.50
47     {128,25,43,45}      0.50
27        {128,25,43}      0.50
15        {128,25,45}      0.50
5            {128,25}      0.50
18        {128,43,45}      0.50
6            {128,43}      0.50
2            {128,45}      0.50
55     {131,25,43,45}      0.50
37        {131,25,43}      0.50
29        {131,25,45}      0.50
32        {131,43,45}      0.50
7            {131,45}      0.50
35         {25,43,45}      0.50
8             {25,45}      0.50
9             {43,45}      0.50
> |
```

Figure 15: Apriori algoritm frequent item sets

4.3.3. Eclat Algorithm



```
> me_item<-as.data.frame(get_eclat(patientSet_me))

parameter specification:
 tidLists support minlen maxlen            target   ext
    FALSE     0.5      1     10 frequent itemsets FALSE

algorithmic control:
 sparse sort verbose
      7   -2    TRUE

eclat - find frequent item sets with the eclat algorithm
version 2.6 (2004.08.16)         (c) 2002-2004   Christian Borgelt
create itemset ...
set transactions ...[20 item(s), 4 transaction(s)] done [0.00s].
sorting and recoding items ... [6 item(s)] done [0.00s].
creating bit matrix ... [6 row(s), 4 column(s)] done [0.00s].
writing  ... [33 set(s)] done [0.00s].
Creating S4 object  ... done [0.00s].
  items support
1 {25,
   43}     0.75
2 {25}     0.75
3 {43}     0.75
4 {131}    0.75
5 {118,
   131}    0.50
```

Figure 16: Eclat algorithm results

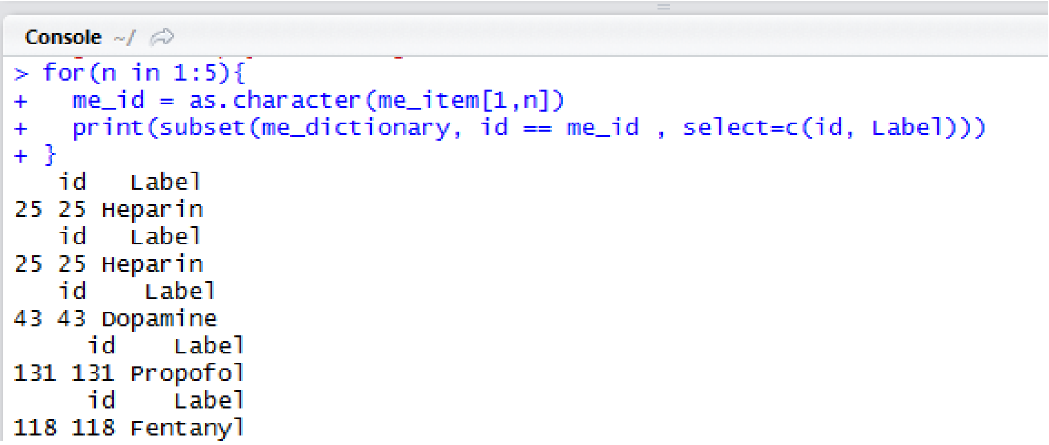
4.3.4. Comparison of Apriori and Eclat Algorithm

Apriori uses a breadth-first search strategy to count the support of itemsets, whereas Eclat uses a depth-first search algorithm using set intersection. Eclat requires less space than apriori if itemsets are small in number. It is suitable for small datasets and requires less time for frequent pattern generation than apriori. However, apriori results are better in terms of its analytical depth and larger item sets produced.

As depicted in the screenshots within the previous sections, the results of Apriori algorithm run through the software is larger than the Eclat algorithm. Apriori would be more feasible for some entry data sets but may be impractical for other entry data sets. This is the case for chart event (ch) entries between different patients' MIMIC II data sets, due to the large and extensive entry logs; it is simply unfeasible to run apriori on these data sets since the odds of having abundance of similar frequent items sets are high.

4.4. Entries Term from MIMIC II tag dictionary

Following the Apriori and Eclat algorithm run on the software, the terms of the IDs of the resulting item sets is extracted from the dictionary to form descriptive output.



```
> for(n in 1:5){
+   me_id = as.character(me_item[1,n])
+   print(subset(me_dictionary, id == me_id , select=c(id, Label)))
+ }
   id   Label
25 25 Heparin
   id   Label
25 25 Heparin
   id     Label
43 43 Dopamine
     id     Label
131 131 Propofol
     id     Label
118 118 Fentanyl
```

Figure 17: Descriptive output for Medication Event (me) patients' data sets algorithms run

5. CHAPTER 5: CONCLUSION & RECOMMENDATION

The study plans to develop different algorithm examples to suit needed outcomes from the analysis of hospital data. These results should assist and aid the study of correlation between different factors relevant to healthcare.

The study has reviewed so far the A-Priori and Eclat algorithm which contains one of the most basic similarity check mechanism. This algorithm has proven its simplicity in finding similar pairs of items in baskets. It will be used as a base method to understand further algorithms and functions to process more complex data.

To be able to obtain required analysis using the market-basket model, it shall be controlled using the support, confidence, lift identities assigned to the rule. It is recommended that other algorithms be tested in correlation with the present one. This paper also advises to discover and understand which data format shall be used in the future to adjust the algorithm and enable it to process data streams or flow.

R language has proven to be the suitable programming software to compute such similarities as its online libraries grow exponentially and its ability to simply manipulate with data and implement various functions.

REFERENCES

[1] Cortada, J., Gordon, D., & Lenihan, B. (2012). *The value of analytics in healthcare*. Retrieved February 22, 2015.

[2] Joachim Roski, George W. Bo-Lin0000n and Timothy A. Andrews: *Creating Value In Health Care Through Big Data: Opportunities And Policy Implications*. Health Affairs, 33, no.7 (2014):1115-1122

[3] Piai, S., & Claps, M. (2013, September). *Bigger Data for Better Healthcare*. Retrieved February, 2015.

[4] Raghupathi and Raghupathi: *Big data analytics in healthcare: promise and potential*. Health Information Science and Systems 2014 2:3.

[5] Sinha, A., Hripcsak, G., & Markatou, M. (2014). Large Datasets in Biomedicine: A Discussion of Salient Analytic Issues. *Perspectives on Informatics*, 16(6), 759-767. Retrieved January 23, 2015, from jamia.bmj.com

[6] Rajaraman, A., Ullman, J., & Leskovec, J. (2012). *Mining of Massive Datasets* (1st ed., Vol. 1, p. 495). New York, N.Y.: Cambridge University Press.

[7] Morton, A., Mengersen, K., Whitby, M., & Playford, G. (2013). *Statistical methods for hospital monitoring with R* (1st ed., Vol. 1, p. 399). John Wiley & Sons.

[8] *Market basket analysis: Identifying products and content that go well together*. (2014, September 1). Retrieved June 12, 2015.

[9] Saeed, M., Lieu, C., Raber, G., & Mark, R. G. (2002, September). *MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring*. In Computers in Cardiology, 2002 (pp. 641-644). IEEE.

[10] Scott, D. J., Lee, J., Silva, I., Park, S., Moody, G. B., Celi, L. A., & Mark, R. G. (2013). *Accessing the public MIMIC-II intensive care relational database for clinical research.* BMC medical informatics and decision making, 13(1), 9.

[11] Douglas, M., Clifford, G. D., Reisner, A., Moody, G. B., & Mark, R. G. (2004, September*). Computer-assisted de-identification of free text in the MIMIC II database*. In Computers in Cardiology, 2004 (pp. 341-344). IEEE.

[12] Sun, J. X., Reisner, A. T., Saeed, M., & Mark, R. G. (2005, September). *Estimating cardiac output from arterial blood pressurewaveforms: a critical evaluation using the MIMIC II database*. In Computers in Cardiology, 2005 (pp. 295-298). IEEE.

[13] Zhang Z, Xu X, Ni H, Deng H (2014) *Predictive Value of Ionized Calcium in Critically Ill Patients: An Analysis of a Large Clinical Database MIMIC II*. PLoS ONE 9(4): e95204. doi:10.1371/journal.pone.0095204

[14] Osborne, R. M., Aronson, A. R., & Cohen, K. B. (2014). *A repository of semantic types in the MIMIC II database clinical notes*. ACL 2014, 93.

[15] Lee, J., Scott, D. J., Villarroel, M., Clifford, G. D., Saeed, M., & Mark, R. G. (2011, August). *Open-access MIMIC-II database for intensive care research*. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 8315-8318). IEEE.

[16] Ihaka, Ross (1998). *R : Past and Future History*. Statistics Department, The University of Auckland, Auckland, New Zealand.

[17] Fox, John and Andersen, Robert (2005, January). *Using the R Statistical Computing Environment to Teach Social Statistics Courses*. Department of Sociology, McMaster University