Technische Universität Berlin

Faculty IV Electrical Engineering and Computer Science
Database and Information Systems

Fakultät IV
DIMA Group
Einsteinufer 17
10587 Berlin
https://www.dima.tu-berlin.de

Master's Thesis

Pattern-based Prediction of Life-Threatening Diseases in Intensive Care

Joud SAYED ISSA

Matriculation Number: 387721
13.09.2019

Supervised by
Prof. Dr. Volker MARKL

Advisors
Dr. Ralf-Detlef KUTSCHE
Anne SCHWERK, Ph.D.

Declaration Of Originality

Hereby I declare that I wrote this thesis myself with the help of no more than the mentioned literature and auxiliary means.

Berlin, 13.09.2019

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich diese These selbst mit Hilfe der oben genannten Literatur und Hilfsmittel selbst verfasst habe.

Berlin, 13.09.2019

. . . .

(Signature [your name])

Abstract

Zusammenfassung

Contents

List of Figures

List of Tables

1 Introduction

This master thesis aims to push forward research in the field of data science, focusing specially on the healthcare domain. Before we start by explaining the clinical scenario and the methods proposed, we want to introduce some fundamental facts and issues related to the same field. Further we will emphasis the need for such a research study on this field.

1.1 Motivation

In recent years, *big data* or datasets with a huge amount of records have had the tendency to be increasing and becoming more common to exist. In a healthcare setting, The move from paper to electronic based patient health records to generate digital medical data has made this industry rich with information helpful for the sake of care continuum. This type of data can quickly add up and produce hundreds of exabytes of healthcare data where it is seen that this amount is expected to be more than yottabytes in the near future [FP16].

In addition to the fact that the healthcare sector contains an enormous amount of data, this produced data can be as structured or unstructured. Structured data is the data gathered from radiology results, laboratory, past medical history, allergy information, or medications lists; while doctor notes on treatments and progress checks are considered unstructured. For storing such data, preexisting hardware and software systems, including relational databases, might not be capable of organizing and handling such large datasets to perform a proper analysis. Especially when data is arriving from different databases or warehouse sources [TK15].

To handle healthcare big data, the methods of *data analytics* is used. With the use of this emerged techniques, different tools are used to extract *patterns* from the complex datasets to make it some how easier to gain insight and knowledge for the purpose of providing better services and better cost efficiency. In such situation, when regular data reports (e.g. daily, weekly, yearly.) on specific parameters of the dataset are needed then *descriptive statistics* are used to compare different Key Performance Indicators (KPI) such as mortality rate or readmission rate. On the other hand, when a prediction is needed then *Predictive analytics* is used on past data events or modelings. Some of the clinical applications, in this case, could be predicting the risk for a patient to have a heart attack, or to predict which patient is more likely to be readmitted to the hospital after doing some type of surgery [AG14]

Over the past few years, the importance of predictive analytics in the settings of healthcare has received a vast amount of interest. The way of gaining knowledge for health and medical data using predictive analytics will change the way medicine is implemented as well as increasing the prevention of the significant potential illnesses and diseases [Alh18]. Usually, it is relatively difficult to analyze real huge datasets in a quick manner in order to be able to obtain decisions related to patient's health. The use of one of the important research area methods such as *data mining* to identify this meaningful information is considered to play a substantial and helpful role in the process of data analytics. This will provide more efficient and valuable information when trying to detect unknown knowledge from the huge heterogeneous healthcare data (e.g. identification of medical treatment methods and detecting unknown diseases) [Mil12]. As of [HDM15], following such methods and techniques in medical research domain have helped a lot in making efficient policies in the domain as well as constructing proper drug recommendation systems, and improving health profiles of patients.

Predictive analytics has developed out of data mining methods to address the specific needs of an organization using designed applications [Fin14]. In today's healthcare organizations we can see how the implementation of predictive analytics has evolved with the aim to start managing and processing the different large data they have, hoping to discover trends, relationships and predict well timed and accurate outcome that could support the delivery of improved healthcare service [Sie13].

Machine learning techniques have become progressively popular in conducting predictive analytics due to their outstanding performance in manage large scale datasets with uniform characteristics and noisy data. Observational studies show that machine learning is appropriate to build predictive models by extracting patterns from large datasets [NI17].

1.2 Objective

Sepsis is a critical clinical disorder described by organ dysfunction caused by a patient's dysregulated reaction to contamination. The septic shock on the other hand is a subset of sepsis with increased mortality described by hypotension. It is characterized as organ dysfunction coming about because of the patient's injurious reaction to disease [PK19]. This disorder is common to arise in an intensive care unit (ICU) which may cause several accompanying organ dysfunctions, kidney is the organ that is said to be more frequent to be affected. Where several studies showed that Acute kidney injury (AKI) may occur in about 19% patients with moderate sepsis, 23% with severe sepsis and 51% with a septic shock when blood cultures are positive [RFPC$^+$95]. AKI may occur prior or after having a septic shock, where we can see that the risk factors are different for each case as well as the pathophysiologic mechanisms may differ. This occurs on a rate range from 11 to 42% [HLV$^+$03] [BGB08] and may at some point reach a high point to 67% in a population with septic surgical [WHB13]. It is stated that for 50% of the cases it is common to develop AKI for critically ill patient in the ICU [UKB$^+$05]. This rate has a correlation of severity with the AKI rate underlying sepsis, where it is noted that septic AKI is associated with length of stay and increased mortality as opposed to patients wiht non-septic shock. Septic AKI is a hallmark of severe sepsis and septic shock and is associated with worse outcomes including prolonged hospital length of stay, fewer

ventilator-free days and increased mortality when compared to patients with non-septic AKI [BGB08] [WHB13] [GMM15].
For the purpose of treating heart disease and related conditions, the drug "Beta blockers" or "beta-adrenergic receptor blockers" are used to reduce blood pressure. It is also used to manage cardiac arrhythmias and are cardioprotective after myocardial infarction (heart attack). The way this drug performs is that it releases the Epinephrine from the adrenal medullary gland where it has influnced the heart and blood vessels. This concept of combination with the adrenoreceptors has set to make a reduction in the strength of the heart muscle and the heart rate [Bla10] [Fri81]. Therefore, this drug is considered to be a performance-enhancing drugs (PEDs) drug and is banned by the World Anti-Doping Agency (WADA)[Hac17].
In the given setting of heart failure and its origin in Acute Kidney Injuries, our central research question in this thesis can be stated as: *In the adult ICU patients with Heart Failure, is the exposure to the beta blocker associated with the development of Acute Kidney Insurey (AKI) than the people who doesn't take beta blocker ?*
As patient's medical records in hospitals has been moved from paper format and digitized into Electronic Health Records (EHR). This has exposed several opportunities for having granular data with lab results, medications and timing of clinical events [SLS$^+$13]. This in turn generates *big databases* which might also have EHR from variety of sources into one clinical system. The nature of this kind of databases can give an in-depth insight about why some interventions have been taken with some patients and why others have not. This puts us to the fact the "Big data" databases, such as the Medical Information Mart for Intensive Care III (MIMIC-III) can help with the process of performing different study types to design different research questions. Furthermore, the granular nature of the data can provide insight as to the reason why one patient received an intervention and another did not which can partly address confounding by indication. Thus, the promise of "big data" is that it contains small, very detailed data. "Big data" databases, such as MIMIC-III, have the potential to expand the scope of what had previously been possible with observational research [MMW16].
The thesis proposed will apply complex data prepossessing techniques and "Big data Analytics" to gain knowledge from a big clinical data set derived from the intensive care unit (ICU) setting: the MIMIC III database. Therefore the solution we are going to follow is to Performe an analysis for EHR data including all patient's diagnosis information, all lab values, vital signs and demographic information, enable suitable prepossessing of heterogeneous data sources, discover patterns of the analyzed data to enable subtyping of patients, apply predictive modeling on historical EHR to verify patterns and at the end establish an end-to-end processing pipeline that allows a high level prediction in the ICU setting.

1.3 Scope - (not finalized)

In order to enable the successful and efficient integration of such heterogeneous and very large data sources in the given setting of heart failure and its origin in Acute Kidney

Injuries (AKI), the thesis proposed will apply complex data preprocessing techniques and 'Big Data Analytics' to gain knowledge from a big clinical data set derived from the intensive care unit (ICU) setting: the MIMIC III database. It includes different data sources and formats, including different data streams (EEG, ECG), demographic information, clinical notes, diagnoses, billing information, and laboratory test results – allowing for complex data modeling [JPS+16].
'Figure 1.1 shows over all picture of the implementation pipeline ...'

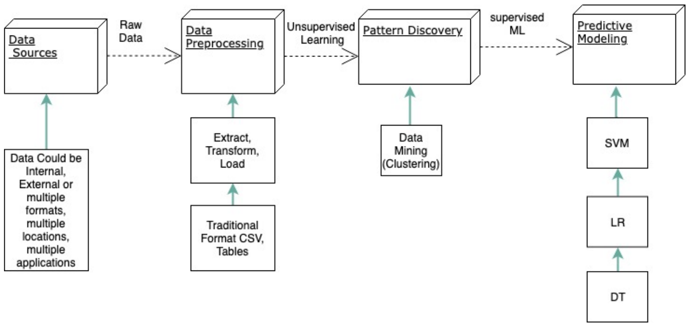

Figure 1.1: Component Based Architecture

To reach our goal we will follow the below steps

- We will use an open public database, the MIMIC-III database [PJI16], which contains deidentified, clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, allowing international researchers to use and replicate those data under a data use agreement.
- The first step during data prepossessing is to select the proper cohort for the proposed study, fol-lowed by the extraction of the related records based on the patient's ID.
- The next prepossessing step involves cleaning the extracted data (e.g. inconsistent record units, multiple values for a variable or handling range variables).
- After prepossessing the raw data, an unsupervised machine learning (ML) method will be applied to all selected patients in the database to identify relevant features.
- Automatic feature selection methods (model-based selection or iterative selection, wrapper method) will then be applied.
- Different clustering methods (e.g. K-means and hierarchical clustering) will be applied to uncover data patterns.
- Then, the selected features will be used in a supervised model for survival/outcome prediction. This is done by comparing logistic regression (LR), support vector machine (SVM) and decision tree (DT) models.

- To evaluate the performance of our model we will calculate the testing error metric and then apply a cross-validation process by using another data set.
- The program Python will be used for ML models and R for simultaneous visualizations.

1.4 Outline

Following we will give a brief introduction into the main chapters of our work. Where this thesis is separated into 7 chapters.

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter ??

2 Background and Fundamentals

In this chapter we will give an introduction about relevant terms and technologies in the field of healthcare and Data Analytics.

2.1 Electronic Health Records

The electronic health record is a key element for assisting the delivery of health care to a patient. It is a dynamic informational entity that continuously monitors the evolution of the health status of a patient. The major advantages of electronic health record over their paper based counterparts (besides the more commonly known provided by the electronic management of data) is that all diagnostic tests, especially medical imagery, become attached to the patients profile and available on electronic format. Most of those tests consist of unstructured media formats, like the images [Spe13].

2.1.1 Basic Components Of An Electronic Health Record

The basic key components of an EHR include:

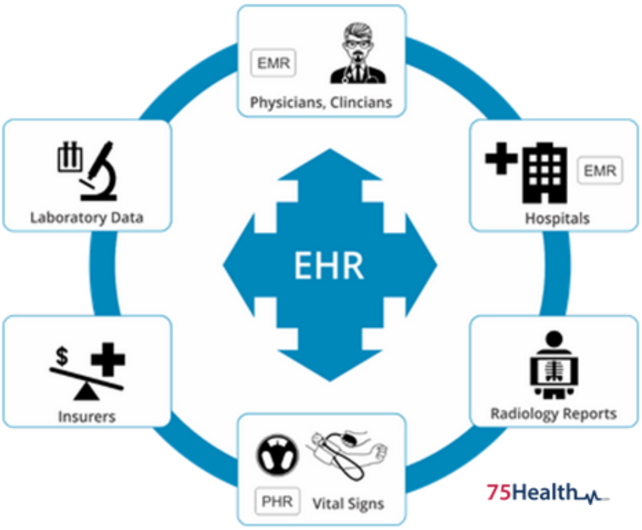

Figure 2.1: Components Of Electronic Health Records [75H17]

Patient Management Component This component is required for patient registration, admission, transfer and discharge (ADT) functionality. Patient registration includes key patient information such as demographics, insurance information, contact information etc. When a patient is registered in an EHR for the first time, a unique ID (often called "Medical Record Number") is generated. Whenever a patient has an encounter with the organization, another unique "encounter" number is generated [Spe13].

Clinical Component This component can house multiple sub-components e.g. Computerized Provide Order Entry (CPOE), electronic documentation, nursing component etc. CPOE allows providers to enter orders that are needed for patient management directly in the computer. This component can make use of clinical decision support tools such as drug-drug, drug-allergy, and drug-diagnoses interactions. This module also allows providers to enter multiple orders from order sets.
Electronic documentation by providers allows them to document notes such as History Physical, consults, discharge summaries, operative notes etc. Multiple tools may be used to enable electronic documentation such as templates, speech recognition and transcription services. The pharmacy system allows for maintaining a drug formulae, filling prescriptions and crosschecking any orders that are placed by providers in the EHR.
Nursing component allows for collection of key patient information such as vital signs, input and output etc. This component also allows for medication administration record (MAR), barcode medication charting and nursing documentation [Spe13].

Laboratory component Lab components are typically divided into two sub-components; 1) Capturing results from lab machines, and 2) Integration with orders, billing and lab machines. The lab component may either be integrated with the EHR or exist as a standalone product [Spe13].

Radiology Information System Radiology information system (RIS) and Picture Archiving Communications System (PACS) are used to manage patient workflow, ordering process, results and the images themselves [Spe13].

Billing System The billing system (hospital and professional billing) is used to capture all charges generated in the process of taking care of patients. These charges generate claims, which is submitted to insurance companies, tracked and completed.
Overlaid on top of the core application layers is generally a data layer, which is fed data by the EHR. This data layer allows healthcare professionals to monitor Key Performance Indicators, view dashboards with relevant business data, and run analytical reports to monitor and improve the performance of the health care organization. One of the successes of using the data captured by EHR's is the ability to track organizational expenses, inventory, and revenue cycle performance. These basic tasks were very complicated in the pre-EHR era.
One of the biggest challenge today around managing electronic health record data is in obtaining clean, discreet data that can be used for analytics. Natural language processing and other tools are being developed to solve some of these problems.
We are in the early years of using this technology in healthcare. Organizations are trying to rebuild paper processes in the EHR, the government regulations are suppressing innovation, and this is leaving healthcare providers frustrated and disenfranchised. It

will be a long time before healthcare is able to reengineer it's processes and adopt this technology to its fullest [Spe13].

2.2 Coding Systems

One of the primary reasons in adopting EHR is to facilitate and promote exchange of information among different healthcare settings. Seamless exchange of information requires coding standards. An interoperable EHR requires standards in four major areas (Reddy and Rahman, 2014): (1) Interaction with users (2) System communication (3) Information processing and management (4) Consumer device integration. Some popular coding standards implemented in most EHR are given below.

2.2.1 International Classification of Diseases

The International Classification of Diseases (often referred as ICD) is the official coding standard introduced by WHO (World Health Organization) to standardize disease and health related information exchange. It is a system of codes that covers diseases and related problems, social circumstances and external causes of injury or disease. The ICD system has gone through various revisions since its introduction. ICD-9 is the most popular version released in 1978. The current revision is ICD-10, which was released by WHO in 1994. ICD-10 covers more diseases and diagnosis codes when compared to its predecessors and the coding scheme is more efficient [RR15]. Australia has its own version of ICD-10 by adding country specific codes. The eleventh revision is ICD-11, is planned for 2018 according to WHO [Gop17].

2.2.2 ICD-9

While the WHO manages ICD, the National Center for Health Statistics (NCHS) and the Centers for Medicare and Medicaid Services (CMS) manage the US's implementation. As such, the ICD-9 source files can be obtained from NCHS or from CMS. The US version is called ICD-9-CM, indicating that they're modified from the original WHO version.

Examples of ICD-9 codes

- E916: Struck accidentally by falling object
- V01.3: Contact with or exposure to smallpox
- V86.0: Estrogen receptor positive status [ER+]
- 250.00: Secondary type II diabetes without mention of complication
- 042: Human immunodeficiency virus [HIV] disease
- 584.9: Acute kidney failure

ICD-9 codes generally consist of a three-digit number, followed by a decimal and up to two more digits. The numbers prior to the decimal indicate the general category of the condition, while any numbers after provide more specific information about the location, severity, or nature of the condition. Additionally, some codes are prefixed by an E or V. These are broader categories of external causes of injury and poisoning (E) or factors influencing health status (V). Each digit is significant and meaningful. That is, the meaning of the ICD-9 code 250 is different from 250.0, which is different from 250.00. The ICD-9 hierarchy is provided in the NCHS files (the DTAB file) [Sch16].

General Structure of ICD-9 Hierarchy

(Numbers)	Diseases and Conditions
001 - 139	Infectious and Parasitic Diseases
140 - 239	Neoplasms
240 - 279	Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders
280 - 289	Diseases of the Blood and Blood-Forming Organs
290 - 319	Mental Disorders
320 - 359	Diseases of the Nervous System
360 - 389	Diseases of the Sense Organs
390 - 459	Diseases of the Circulatory System
460 - 519	Diseases of the Respiratory System
520 - 579	Diseases of the Digestive System
580 - 629	Diseases of the Genitourinary System
630 - 679	Complications of Pregnancy, Childbirth, and the Puerperium
680 - 709	Diseases of the Skin and Subcutaneous Tissue
710 - 739	Diseases of the Musculoskeletal System and Connective Tissue
740 - 759	Congenital Anomalies
760 - 779	Certain Conditions Originating in the Perinatal Period
780 - 799	Symptoms, Signs, and Ill-defined Conditions
800 - 999	Injury and Poisoning
E	External causes of injury and poisoning
V	Factors influencing health status

Table 2.1: General Structure of ICD-9 Hierarchy [Sch16]

2.2.3 ICD-10

In the 15 years between the development of ICD-9 and ICD-10, there were many changes in the uses and requirements of health data. Of course, one of the largest changes was the increase in ubiquity of computers. With the improvement in processing capability, the level of detail contained within ICD-9 was no longer sufficient. For example, when a patient has a burn, the location of the burn is often a notable detail. Given that ICD-9 only allows for up to 5 digits, there is an inherent limit in the amount of codes that can be provided (10ˆ5). ICD-10 solves this by changing the structure of the code to start with an alpha character followed by two digit number. In addition, there are up to three additional alphanumeric characters available for specifying additional clinical details [Sch16].

General Structure of ICD-10 Hierarchy

(Numbers)	Diseases and Conditions
A00 - B99	Infectious and parasitic diseases
C00 - D48	Neoplasms
D50 - D89	Diseases of blood and blood-forming organs
E00 - E90	Endocrine, nutritional, and metabolic diseases
F00 - F99	Mental and behavioral disorders
G00 - G99	Diseases of the nervous system
H00 - H59	Diseases of the eye and adnexa
H60 - H95	Diseases of the ear and mastoid process
I00 - I99	Diseases of the circulatory system
J00 - J99	Diseases of the respiratory system
K00 - K93	Diseases of the digestive system
L00 - L99	Diseases of the skin and subcutaneous tissue
M00 - M99	Diseases of the musculoskeletal system and connective tissue
N00 - N99	Diseases of the genitourinary system
O00 - O99	Pregnancy, childbirth and the puerperium
P00 - P96	Certain conditions originating in the perinatal period
Q00 - Q99	Congenital malformations, deformations and chromosomal abnormalities
R00 - R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
S00 - T98	Injury, poisoning and certain other consequences of external causes
V01 - Y98	External causes of morbidity and mortality
Z00 - Z99	Factors influencing health status and contact with health services
U00 - U99	Codes for special purposes

Table 2.2: General Structure of ICD-10 Hierarchy [Sch16]

Similar to ICD-9, the ICD-10 source files can be obtained from either NCHS or CMS. And, also similar to ICD-9, the file formats vary depending on where you obtain it: NCHS offers the files in XML or PDF format, while CMS offers them in a fixed-width format. While the hierarchy has changed pretty substantially, and there are much more details available ICD-10 than ICD-9, the core way we access the data hasn't changed [Sch16].

2.3 The MIMIC-III Database

The MIMIC database is in its third iteration and is referred to as the "Medical Information Mart for Intensive Care" (previously called the "Multiparameter Intelligent Monitoring in Intensive Care" while in its second iteration - MIMIC-II, [JPS+16]). MIMIC-III is a comprehensive collection of deidentified data from 53,423 distinct critical care hospital admissions from 38,597 distinct adult patients at the Beth Israel Deaconess Medical Center in Boston, Massachusetts [JPS+16]. The data has been compiled into 26 tables which contain, for example, an average of 4579 charted observations and 380 laboratory measurements for each hospital admission as well as a total of 3.8 gigabytes of unstructured textual data from various healthcare provider notes and analyses [Pol14]. An excellent figure from [JPS+16] summarizing the MIMIC-III database is included in figure 2.2. Historically, the MIMIC database has been used in industrial research, quality improvement initiatives, and higher education coursework [Pol14].

2.3.1 Data Acquisition

In 2003, colleagues from academia (Massachusetts Institute of Technology), industry (Philips Medical Systems), and clinical medicine (Beth Israel Deaconess Medical Center, BIDMC) were received NIH (National Institutes of Health) funding to launch the project "Integrating Signals, Models and Reasoning in Critical Care", a major goal of their which was to build a massive critical care research database. The requirement for individual patient is to protected health information and to be de-identified. Each patient record began at ICU admission and ended at final discharge from the hospital. The data acquisition process was continuous and invisible [Mar16]. Three categories of data were collected: **clinical data**, which were aggregated from ICU information systems and hospital archives; **high-resolution physiological data** (waveforms and time series of vital signs and alarms obtained from bedside monitors); and **death data** from Social Security Administration Death Master Files [Mar16]

Clinical Data Bedside clinical data were downloaded from archived data files of the CareVue Clinical Information System (Philips Healthcare, Andover, MA) used in the ICUs. Additional clinical data were obtained from the hospital's extensive digital archives. The data classes included [Mar16]:

- Patient demographics, Hospital administrative data: admission/discharge/death dates, room tracking, billing codes, etc.
- Physiologic: hourly vital signs, clinical severity scores, ventilator settings, etc.
- Medications: IV medications, physician orders
- Lab tests: chemistry, hematology, ABGs, microbiology, etc.
- Notes and reports: Discharge summaries; progress notes; ECG, imaging and echo reports.

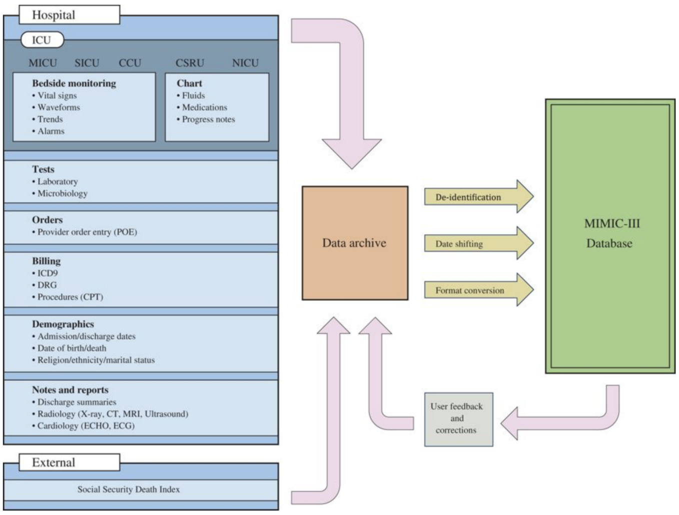

Figure 2.2: Overview of the MIMIC-III critical care database [JPS+16]

2.3.2 MIMIC-III Data Sharing

MIMIC-III is an unprecedented and innovative open research resource that grants researchers from around the world free access to highly granular ICU data and in the process substantially accelerates knowledge creation in the field of critical care medicine. To restrict users to legitimate medical researchers, access to the clinical database requires completion of a simple data use agreement (DUA) and proof that the researcher has completed human subjects training [Mar16].

The MIMIC-III clinical database is available in two forms. In the first form, interested researchers can obtain a flat-file text version of the clinical database and the associated database schema that enables them to reconstruct the database using a database management system of their choice. In the second form, interested researchers can gain limited access to the database through QueryBuilder, a password-protected web service. Detailed documentation and procedures for obtaining access to MIMIC-III are available at the MIMIC-III web site[MIM16].

2.3.3 Data Cleaning

2.3.4 Missing Data

2.4 Pattern Discovery

2.4.1 Why Pattern Discovery?

Pattern Discovery Technologies is a part of the data mining and predictive analytics. Founded in 1997 by distinguished scientist Dr. Andrew Wong as a spin-off from the world renowned Pattern Analysis and Machine Intelligence (PAMI) Lab at the University of Waterloo, Pattern Discovery continues to push the envelope in developing solutions that tackle the most challenging data mining, analysis and plant optimization problems.

2.4.2 Advantages of the Pattern Discovery

Some of the advantages are as follows

- The ability to characterize key performance factors and forecast events using Intelligent Analytics can return crucial dividends. In operations that are complex and multi-faceted, count on Pattern Discovery's state-of-the-art technologies to provide the insight needed to take control of the situation.
- Discover new unique relationships that are not intuitively obvious Optimize processes and resources by understanding and quantifying the key parameters that drive operational efficiencies
- Predict results that take into account the interrelationships of complex multi-dimensional factors (forecast events before they happen)
- Make decisions that can be acted on more confidently and more consistently, using a basis that is squarely supported by the data that drives the operation (free of assumptions and personal biases)
- Leverage existing investments in data collection and aggregation by integrating Discover*e to provide advanced analysis solutions

2.5 Predictive Modeling

There are two types of predictive analytics that we can start of describing

Retrospective analytics This approach helps us analyze history and gain insights from the data. It allows us to learn from mistakes and adopt best practices. These insights and learnings is the purpose of devising better strategy. Not surprisingly, many experts have been claiming that data is the new middle manager.

Predictive analytics This approach unleashes the might of data. In short, this approach allows us to predict the future. Data science algorithms take historical data and spit out a statistical model, which can predict who will buy, cheat, lie, or die in the future.

Here we can come to the definition of predictive analytics as it is an ensemble of statistical algorithms coded in a statistical tool, which when applied on historical data, outputs a mathematical function. It can inturn be used to predict outcomes based on some inputs (on which the model operates) from the future to drive a goal in business context or enable better decision making in general.

Ensemble of statistical algorithms Statistics are important to understand data. It tells volumes about the data. How is the data distributed? Is it centered with little variance or does it varies widely? Are two of the variables dependent on or independent of each other? Statistics helps us answer these questions.

2.5.1 Algorithms

Algorithms are the blueprints of a model. They are responsible for creating mathematical equations from the historical data. They analyze the data, quantify the relationship between the variables, and convert it into a mathematical equation. There is a variety of them:

Logistic Regression

Clustering

Decision Trees

Time-Series Modelling

Naïve Bayes Classifiers

These models can be classified under two classes:

Supervised algorithms These are the algorithms wherein the historical data has an output variable in addition to the input variables. The model makes use of the output variables from historical data, apart from the input variables. The examples of such algorithms include Linear Regression, Logistic Regression, Decision Trees, and so on. **Unsupervised algorithms** These algorithms work without an output variable in the historical data. The example of such algorithms includes clustering.

It is also noted in many other researches that selection of a particular algorithm for a model depends majorly on the kind of data available.

3 Literature Review

In this section we will focus on the scientific background, we will take a look at the related work with use to the same data source we are using. The different medical scenarios that have been researched previously and what are some the techniques they have been using to find answers for the different research questions. We will give a broad and detailed overview of the state of the art in health care predictive analytics and pattern discovery.

3.1 State of the art in MIMIC

3.2 Data collection and Analysis

3.3 Healthcare Data Analytics

3.4 Prediction Models for Clinical Data Analysis

4 Methodology

4.1 Building the Database

4.2 Cohort selection

4.2.1 Building the Views

Demographics

Vital Signs

Lab events

Patients with sepsis

Patients with Beta-blocker treatment

Patients with Acute Kidney Injury

4.2.2 Putting it All Together

4.3 Data Cleaning

4.4 Data Analysis

4.4.1 Pattern Mining

4.4.2 Prediction Models

5 Implementation

5.1 Environment

5.2 Important Implementation Aspects

5.3 Documentation

6 Results

6.1 Performance Measurements

7 Conclusion

7.1 Summary

7.2 Future work

List of Acronyms

KPI	Key Performance Indicator
EHR	Electronic Health Records
HF	Heart Failure
AKI	Acute Kidney Injury
EHR	Electronic Health Record
ICU	intensive care unit
PEDs	performance-enhancing drugs
WADA	World Anti-Doping Agency
MIMIC-III	Medical Information Mart for Intensive Care III
WHO	World Health Organization

Bibliography

[75H17] 75Health: *What Components Constitute an Electronic Health Record?*, 2017.

[AG14] Adams, James and David Garets: *The healthcare analytics evolution: moving from descriptive to predictive to prescriptive.* In *Analytics in Healthcare*, pages 24–31. HIMSS Publishing, 2014.

[Alh18] Alharthi, Hana: *Healthcare predictive analytics: An overview with a focus on Saudi Arabia.* Journal of Infection and Public Health, 11(6):749 – 756, 2018.

[BGB08] Bagshaw, Sean M, Carol George and Rinaldo Bellomo: *Early acute kidney injury and sepsis: a multicentre evaluation.* Critical care, 12(2):R47, 2008.

[Bla10] Black, James: *Reflections on drug research.* British Journal of Pharmacology, 161(6):1204–1216, 2010.

[Fin14] Finlay, Steven: *Predictive analytics, data mining and big data: Myths, misconceptions and methods.* Springer, 2014.

[FP16] Fang, Ruogu and Pouyanfar: *Computational health informatics in the big data age: a survey.* ACM Computing Surveys (CSUR), 49(1):12, 2016.

[Fri81] Frishman, William H: *β-Adrenoceptor antagonists: new drugs and new indications.* New England Journal of Medicine, 305(9):500–506, 1981.

[GMM15] Godin, Melanie, Patrick Murray and Ravindra L Mehta: *Clinical approach to the patient with AKI and sepsis.* In *Seminars in nephrology*, volume 35, pages 12–22. Elsevier, 2015.

[Gop17] Gopakumar, Shivapratap: *Machine learning in healthcare: an investigation into model stability.* Technical Report, Deakin University, 2017.

[Hac17] Hackney, Anthony C: *Doping, performance-enhancing drugs, and hormones in sport: mechanisms of action and methods of detection.* Elsevier, 2017.

[HDM15] Haraty, Ramzi A, Mohamad Dimishkieh and Mehedi Masud: *An enhanced k-means clustering algorithm for pattern discovery in healthcare* *data.* International Journal of distributed sensor networks, 11(6):615740, 2015.

[HLV+03] Hoste, Eric AJ, Norbert H Lameire, Raymond C Vanholder, Dominique D Benoit, Johan MA Decruyenaere and Francis A Colardyn: *Acute renal failure in patients with sepsis in a surgical ICU: predictive factors, incidence, comorbidity, and outcome.* Journal of the American Society of Nephrology, 14(4):1022–1030, 2003.

[JPS+16] Johnson, Alistair EW, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi and Roger G Mark: *MIMIC-III, a freely accessible critical care database.* Scientific data, 3:160035, 2016.

[Mar16] Mark, Roger: *The Story of MIMIC*, pages 43–49. Springer International Publishing, Cham, 2016.

[Mil12] Milovic, Boris: *Prediction and decision making in health care using data mining.* Kuwait chapter of arabian journal of business and management review, 33(848):1–11, 2012.

[MIM16] MIMIC, Physionet: *Requesting access*, 2016.

[MMW16] Mehta, Anuj, Brian Malley and Allan Walkey: *Formulating the research question.* In *Secondary Analysis of Electronic Health Records*, pages 81–92. Springer, 2016.

[NI17] Nithya, B and V Ilango: *Predictive analytics in health care using machine learning tools and techniques.* In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 492–499. IEEE, 2017.

[PJI16] Pollard, TJ and AEW Johnson III: *The MIMIC-III Clinical Database http://dx. doi. org/10.13026.* C2XW26, 2016.

[PK19] Poston, Jason T and Jay L Koyner: *Sepsis associated acute kidney injury.* BMJ, 364, 2019.

[Pol14] Polese, Giuseppe: *A decision support system for evidence based medicine.* Journal of Visual Languages & Computing, 25(6):858–867, 2014.

[RFPC+95] Rangel-Frausto, M Sigfrido, Didier Pittet, Michele Costigan, Taekyu Hwang, Charles S Davis and Richard P Wenzel: *The natural history of the systemic inflammatory response syndrome (SIRS): a prospective study.* Jama, 273(2):117–123, 1995.

[RR15] Rahman, Rajiur and Chandan K Reddy: *Electronic Health Records: A Survey.* Healthcare Data Analytics, 36:21, 2015.

[Sch16] Schrom, John: *Introduction to electronic health records.* O'Reilly Media, Inc., 2016.

[Sie13] Siegel, Eric: *Predictive analytics: The power to predict who will click, buy, lie, or die.* John Wiley & Sons, 2013.

[SLS+13] Scott, Daniel J, Joon Lee, Ikaro Silva, Shinhyuk Park, George B Moody, Leo A Celi and Roger G Mark: *Accessing the public MIMIC-II intensive care relational database for clinical research.* BMC medical informatics and decision making, 13(1):9, 2013.

[Spe13] Speculation, Accidental: *Basic Components Of An Electronic Health Record*, 2013.

[TK15] Trujillo, George and Kim: *Virtualizing hadoop: how to install, deploy, and optimize hadoop in a virtualized architecture.* VMware Press, 2015.

[UKB+05] Uchino, Shigehiko, John A Kellum, Rinaldo Bellomo, Gordon S Doig, Hiroshi Morimatsu, Stanislao Morgera, Miet Schetz, Ian Tan, Catherine Bouman, Ettiene Macedo et al.: *Acute renal failure in critically ill patients: a multinational, multicenter study.* Jama, 294(7):813–818, 2005.

[WHB13] White, Laura E, Heitham T Hassoun and Bihorac: *Acute kidney injury (AKI) is surprisingly common and a powerful predictor of mortality in surgical sepsis.* The journal of trauma and acute care surgery, 75(3), 2013.

Annex