Exploratory Data Analysis of Intensive Care Unit Patients using MIMIC-III Database

by

Abhishek Arya
B.E., Rajiv Gandhi Technical University, India, 2009

A Project Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Supervisory Committee

Exploratory Data Analysis of Intensive Care Unit Patients using MIMIC-III Database

by

Abhishek Arya
B.E., Rajiv Gandhi Technical University, India, 2009

Supervisory Committee

Dr. Daniel M. German, Department of Computer Science
Co - Supervisor

Dr. Alex Kuo, Department of Health Information Science
Co - Supervisor

Abstract

Exploratory data analysis refers to the set of procedures for producing descriptive and graphical summaries of the data. This analysis has been performed on MIMIC (Multiparameter Intelligent Monitoring in Intensive Care) version 3 database. It is a relational database containing data related to intensive care unit patients at Beth Israel Deaconess Medical center and is a freely accessible dataset utilized by researchers all over the world. This project provides a vital outlook of the dataset which includes identification of any anomalies, check for assumptions, describing important variables, understanding relationship among variables, and frame hypotheses. The goal of this project is to conduct a pre-analysis of the database, which includes analysis of the database on the basis of variables of interest, before diving into deeper data analysis such as using machine learning or statistical modeling techniques and to make sure that data is relevant without any major problems to provide a valuable contribution towards further researches on the database.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank **Dr. Alex Kuo** for his major support and guidance during the project and my supervisor **Dr. Daniel M. German** for his valuable mentoring and suggestions at each step.

1. Introduction

Exploratory data analysis is basically a set of procedures which uses graphical methods to get an insight of the dataset and summarize its main characteristics. The sole purpose of this technique is to ensure that the data is ready and is in its most suitable form for further deeper analysis using various statistical models and techniques. There has been various ways to achieve this goal. Following steps [1] provide the most general way to approach this methodology,

1. Get the data and import it in most suitable form.
2. Get the overview and learn further about data (either by metadata or sample data).
3. Identify all the variables of interest in the dataset.
4. Spot any assumptions, missing values, anomalies or outliers in the database by summarizing and visualizing variables.
5. Discover patterns and trends in the dataset and formulate hypotheses for further analysis.

MIMIC-III [2] is a freely accessible dataset developed by the MIT lab for computational physiology utilized by numerous researchers and scholars world wide. The database consists of demographics, vital signs, laboratory tests, medications, caregiver notes, and mortality of approximately 40,000 intensive care unit patients at Beth Israel Deaconess Medical center between 2001 and 2012.

The purpose of this project is to get an insight of dataset using exploratory data analysis and provide valuable information in the form of summarizing its main characteristics, variables of interest, formulating hypotheses, showing general patterns and trends in the MIMIC-III database. This will be very helpful for the other researchers and scholars, in particular for researchers who are new to MIMIC-III database, to get right insight of the database and to identify the readiness of this dataset for specific research or analysis.

2. Background and Related Work

Exploratory data analysis [4] was first introduced by John Tukey in 1970s and he wrote a book called *Exploratory Data Analysis* in 1977 [3] in which he mentions various statistical techniques with focus on statistical hypothesis testing which is also known as confirmatory data analysis. The techniques utilized to serve the purpose of exploratory data analysis has been evolved with time and has become more visual and graphical than descriptive. Various graphical methods such as Box plots, histograms, scatter plots, etc., have been utilized to provide a clear and penetrative understanding of the dataset. The book *Graphical Exploratory Data Analysis* by S. H. C. DuToit [5] explains the importance of portraying the data graphically to have a clearer understanding. He mentions about the realization of using graphical methods in various areas of research, teaching methods and any statistical consultation. There have been numerous research in almost every domain and scholar work where exploratory data analysis has been a vital part to produce results.

Medical Information Mart for Intensive Care also known as MIMIC database was first introduced as MIMIC-II v2.4 and has been an effort driven by several research scholars under the guidance of Prof. Roger Mark at the Laboratory for Computational Physiology, Massachusetts Institute of Technology. This study was approved by Institutional Review of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Individual patient consent requirement was waived as the study did not impact clinical care and all data were de-identified [6].There has been several updates since the first release of the database. The major changes from MIMIC-II to MIMIC-III is the added patient's data to the database. MIMIC-II contains database of patients from year 2001 till 2008 where as the current version of MIMIC-III database contains data from year 2001 till 2012.

Modern intensive care units have sophisticated technologies to handle and generate informative set of live clinical and physiologic data used to guide patient care. Clinicians at those units are challenged to interpret this huge amount of data to improve patient care and also to justify the cost. This enormous amount of data and its poor interpretation leads to much lower level of patient care and unwanted cost. This data was highly required to be integrated and managed to overcome these challenges which ultimately be efficient for further analysis and less time consuming [6].

Over the past years, since the MIMIC database was released, there have been numerous research and scholar works conducted on the database to generate important results and predictions. Some of the research includes prediction of disease and symptoms in patients such as prediction of hyperlactatemia using the MIMIC-II database [7], getting predictive value of ionized calcium in critically ill patients using MIMIC-II database [8] and predicting acute respiratory distress syndrome in intensive care unit patients using MIMIC-II physiological database [9]. There has been other kinds of works using the same database which are not direct analysis or generate results on the database but helpful for further analysis. One such work is a web-based data visualization tool for the MIMIC-II database [10] which is interactive to users and its ease of use can open doors for researchers with even less technical knowledge of database query language such as SQL.

3. Motivation for the project

Better care for patients at intensive care units has been motivation of the project. Patients in Intensive care unit are physiologically fragile. They need significant guidance and care than any other type of patients. To achieve this motivation this study is aligned to subject such as co-morbidity and analysis of factors associated with length of stay of patients.

There has been studies which shows the effects of co-morbidity in patients but there is still space to analyse deeper with patient’s database such as the MIMIC-III.

The study on the impact of comorbid conditions on critical illness at National Center for Biotechnology Information, U.S. National Library of Medicine [16] says that presence of co-morbid conditions in critically ill patients can lead to chronic illness, organ failure or even death. Early identification of co-morbid diseases could be immensely helpful for physicians to prepare necessary measures.

Also, reduced length of stay of patients at hospital facility could remarkably bring down the cost. Long stays have been at the hospital facility has been costly and burdensome to patients, their families, and society. [15]

The study could be well utilized by the below mentioned target audience.

Physicians: Diagnoses that co-occur in patients that have been analyzed in this study would be beneficial for physicians at intensive care units to understand co-morbidity in patients. This could lead to early and prompt actions again the diseases that could occur along with certain conditions.

Independent researchers: This study also provides base to researchers to further analyze about certain diagnoses and other results regarding ICU patients at a deeper level.

Hospital facility: Factors associated with length of stay of patients have been shown and its relative importance which could be utilized by hospital administration and physicians.

4. Overview of MIMIC-III database

MIMIC-III [2] critical care database is an freely accessible database containing de-identified health data associated with approximately 40,000 patients who stayed in intensive care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. There has been several versions of the MIMIC-III database since it was released. MIMIC-III v1.0 was released on 25th August 2015, a preliminary beta version which was not widely publicized. There were no significant updates noted for this version of the MIMIC III. Further, there were three more versions of the database (v1.1, v1.2 and v1.3) which have significant updates such as bug fixes, addition of large amount of data, setting the consistency of the dataset and much more. Current version of the dataset is MIMIC-

III v1.4 released on 2nd September 2016 which majorly focused on enhancement of data quality and providing addition of large amount of data.

The current version of the database contains 40 tables and total of about 728,556,685 rows. The database is majorly divided into three levels of data [2],

1. Patient level data
2. Hospital level data
3. ICU (Intensive care unit) level data

Below is the overview of information present in the Patient level data,

1. Gender information
2. Date of birth
3. Date of death
4. Date of death as per hospital records
5. Date of death as per social security records

Below is the overview of information present in the Hospital level data,

1. Hospital admissions associated with stay of patient in intensive care unit.
2. Information of when patients were ready for discharge and the actual time of their discharge.
3. Information on current procedural methodologies associated with patient.
4. Records on diagnosis relating to an admission in hospital.
5. List of patients admitted to ICU and related records in regards with hospital.
6. Records related to laboratory and microbiology tests performed.
7. Patient's data associated with admission in the hospital.
8. Record on medicines prescribed to patients.
9. Data related to hospital services that were given to patient during their stay in hospital.
10. Record on location of patient during the hospital stay.

Below is the overview of information present in the ICU level data,

1. Date and time of events and fluid events that were recorded which are associated with patient during ICU stay.
2. Notes and description associated with patient during hospital stays.
3. Outputs recorded during ICU stay of patient
4. Procedure start and stop time recorded associated with the patient.

5. Methodology

This section of the report provides detail about the methodology adopted for the work. Following are the generalization of steps which are utilized for the analysis of dataset which is also defined in the *introduction section* of this report,

1. Importing the dataset
2. Understanding the dataset
3. Identifying variables of interest
4. Visualizing and analysing the variables of interest.

These steps are described in detail in the following sub-sections.

4.1 Specifications

Specifications section provides detail about the work conducted such as area of research and the involved data and further information related to the data.

Below mentioned Table 1 is a specifications table in regards to the work conducted,

Table 1: Specifications table

Specific subject area	Health-related data set
Type of data	Record in form of tables
How data was acquired	Online freely accessible MIMIC-III database [2] [11]
Data format	MIMIC database is provided as a collection of comma-separated (CSV) files, zipped in GZ format.
Data source location	Dataset is openly available and could be downloaded from Physionet website [12]
Data accessibility	A training course is required to be completed and a completion report is achieved (See Appendix A). Then, request is required to be sent. [2]

4.2 Approach

The following steps describe the methodology in detail,

1. **Importing the dataset**

The MIMIC-III dataset is not readily available for download and anyone who wants to utilize the dataset in any research needs to request at the official website. The request process [2] also requires the applicant to complete CITI "Data or Specimens Only Research" course. Please refer to *Appendix A* for the certificate of completion for this course. After the course is completed, the copy of completion report needs to be attached with the application. The application requires approval and reference from the project supervisor. Once the application is approved, the link to download the database is available. The files available to download are in CSV format and each file is zipped in GZ format. The files can be imported to various database systems or could be utilized in CSV format. In this project CSV files are imported to PostgreSQL database as well utilized in CSV format for analysis.

2. **Understanding the dataset**

Getting familiar with data and understanding its structure is highly important before analysis. Information on the type of record and overview (Entity relationship diagrams and metadata) of MIMIC-III database is available on the website which is very helpful. It also shows how various tables interact with each other which is necessary to understand relationships. Following steps were taken to understand the database thoroughly,

- Data dictionary is read which is available at the MIMIC-III database official website to understand detailed view, purpose and description of each table present in the database.
- Once the database is imported in PostgreSQL, relationship between each table and attribute reference in the database is understood by describing tables using SQL.
- Database is queried to select rows to understand data and its type from each attribute of a table.

3. Identifying variables of interest

After understanding the structure of the dataset, it’s the right time to identify significant variables of interest around which analysis is build.

In this project, variables of interest are identified and based on the outcome which the study seeks to measure.

For example, one of the analyses of this study is to understand co-occurrence of one group of diagnoses with another. Diagnoses identified in a patient are divided into 18 groups of ICD-9 codes. So, here the main variable of interest is ICD-9 code corresponding to each patient's diagnoses. ICD-9 code is the international statistical classification of diseases, injuries and causes of death in patients. This variable is identified as ICD9_CODE attribute in the database.

4. Visualizing and analysing the variables of interest

This section describes how identified variables of interest are analysed and visualised in this project. Various methodology have been adopted for analysis in this project as mentioned below,

- Descriptive Statistics
- Univariate analysis
- Multivariate analysis
- Cross-tabulation analysis
- Comparative analysis
- Network analysis

Below mentioned is the detailed approach for each outcome of the study.

Approach for cross tabulation analysis of patient's survival on the basis of gender

Below mentioned are the steps followed for this analysis,

1. We read data from the table PATIENTS.
2. Attributes GENDER and EXPIRE_FLAG from this table are analysed to produce a cross tabulation analysis. EXPIRE_FLAG is a binary flag which indicates whether the patient died.
3. First, gender distribution of the patients is checked.
4. Next, we checked how many patients are alive using EXPIRE_FLAG.
5. Two crosstab plots are created using pandas python library to visualise patients' survival on the basis of gender.

Approach for analysis of length of stay of patients in ICU using univariate analysis

Below mentioned are the steps followed for this analysis,

1. We have utilized univariate analysis methodology here. Univariate analysis deals with a single variable at a time. Its purpose is to describe a single variable, find patterns and anomalies.
2. The data is read first from the table ICUSTAYS which defines each of the ICU stay of the patient and the details related to it.
3. Attribute LOS is the variable on which this analysis is conducted. LOS is the length of stay for the patient for the given ICU stay. The length of stay is stored in fractional days in the database.
4. Next, we identified total unique patients and total unique ICU stays.
5. Then, we plot a histogram with axis x as the length of stay in the intensive care unit (ICU), in days and axis y is the number of patients using Matplotlib and Seaborn library in python. The goal of this histogram is to plot a univariate distribution of observations.
6. Finally, we generated a descriptive statistics for length of stay variable with discovered patterns, spotted potential anomalies and outliers.

Approach for understanding relation of variables associated with length of stay of patients

Below mentioned are the steps followed for this analysis,

1. We extract the information recorded during the admission of each patient from table ADMISSIONS and their resulting length of stay in the ICU from table ICUSTAYS. As a result, we receive a table with the following columns: length of stay (’LOS’), 'ADMISSION_LOCATION', 'INSURANCE', 'MARITAL_STATUS', 'DIAGNOSIS’.
2. To analyze dependence between length of stay and any of these diagnoses, we use contingency tables. First, we transform length of stay (initially expressed in day) to a categorical variable to receive more interpretable results. We distinguish the following categories in length of stay: '0-1 days', '2-5 days', '6-10 days', '11-25 days', '26-50 days', '51-100 days', '100-200 days’.
3. Finally, we built a contingency table for each paid of variables using Python function pandas.crosstab. The result can be displayed either as a table or as a heat map. However, this analysis does not reveal any dependence between any pair of variables.
4. Finally we also built a random forest regression model, where length of stay is a target variable, and features are the following ones: ['ADMISSION_LOCATION', 'INSURANCE', 'MARITAL_STATUS', 'DIAGNOSIS’]. We allocate 106414 samples in the training set and 10000 samples in the test set.

   Using the class RandomForestRegressor from library sklearn, we fit a model using the training set and evaluate the accuracy using the test set: mean square error is 32.36.

Approach for comparative analysis of top diagnosis of patients in ICU to satisfy the survival rate hypothesis

Below mentioned are the steps followed for this analysis,

1. We read the data from ADMISSIONS table to have a value count of top three most common diagnosis from the attribute DIAGNOSIS.
2. We get a value count of HOSPITAL_EXPIRE_FLAG to get the count of total patients died and survived. This indicates whether the patient died within the given hospitalization. 1 indicates death in the hospital, and 0 indicates survival to hospital discharge.
3. For the top three most common diagnosis (newborn, pneumonia and sepsis) we get the value counts of death and survival of patients.
4. A comparison matrix is created to calculate percentage of survival in each of the diagnoses to satisfy the hypothesis.

Approach for analysis of diagnoses for its co-occurrence in patients

There are two parts to this analysis: A high level analysis using ICD9 group codes for diagnoses and a detailed analysis using ICD9 codes.

For the high level analysis,

1. First, we read data from the table DIAGNOSES_ICD. This table contains a column with full codes of each diagnoses: ICD9_CODE.
2. Then, we constructed a new column ICD9_GROUP, where we wrote an ICD9 code group for each diagnosis. For example, ICD9 codes between 001 are 139 are “infectious and parasitic diseases” (we denote this group as “001”), ICD9 codes between 140 and 239 are “neoplasms” (we denote this group as “140”), etc. We compute this column according to ICD9 classification based on the first three symbols of the disease code. For example, if ICD9 code for a disease is ‘43021’, we take the first three symbols ‘430’ and locate the corresponding range: ‘390–459: diseases of the circulatory system’. For simplicity, we denote this group as '390'. The detail classification for ICD9 group code is present in the *Appendix D*.

3. During the next steps, we built a co-occurrence matrix that shows how often a diagnosis of each group co-occurs with a diagnosis from another group. The matrix is 18x18 matrix (given that we have 18 groups of diagnoses according to ICD9 codes). The matrix is initialized with zeros, because it is used as a counter.
4. We group the data table by patient ID. This way, we can locate the corresponding list of values of ICD9_GROUP for each patient. For example, a single patient may have diagnoses from 4 different groups.
5. For each patient, we located their list of diagnosis groups (after the grouping operation in the previous step). Then, we updated the co-occurrence matrix as follows: if a patient has a diagnosis from both group x and group y, we update the counter: we add 1 to the corresponding element (x, y) in the co-occurrence matrix.
6. As a result of the previous step, we have a co-occurrence matrix that shows which diagnoses groups co-occur together. We plot this co-occurrence matrix as a heat map, to see which cases were the most frequent.
7. Additionally, we can interpret co-occurrence matrix as an adjacency matrix of a graph (network). This way, we have 18 nodes corresponding to each group of diagnoses, and the matrix showing the strength (weight) of connections between them. We visualized this matrix in a circular layout to see the strongest and the weakest connection between the diagnosis groups.

For the detailed analysis,

1. First, we grouped the table of diagnoses by PATIENT_ID. This way, we can identify a list of diagnoses corresponding to each patient. For example, if patient is diagnosed with three conditions, the corresponding list will contain three elements, where each element represents a diagnosis.
2. Next, we constructed a weighted undirected graph. We go through the list of patients, and for each patient we get the list of corresponding diagnoses. For each pair of diagnoses (x, y) in this list, we add a new edge (x, y) with weight=1 to our graph or, in case this edge already exists, we update its weight by adding 1. This way, the weight of each edge (x, y) will show how many patients were diagnosed with both x and y.

3. During the previous step, we have constructed a graph (network) G that we can now save as a file with extension .gexf that can be exported to Gephi.
4. Now we can open the graph file in Gephi, a tool for graph visualization and analysis. Gephi has several built-in preprocessing tools, including the ones for computing degree of each node (in our case, each node is a diagnoses, and the degree is the number of edges towards other connected diagnoses), and for running a clustering algorithm in the graph that "detects communities" in the set of nodes. For example, in a social network, community detection will identify groups of people who are more closely connected to each other, compared with other people. In our case, this algorithm will identify groups of diagnoses that are more closely connected to each other. For a given graph, Gephi can compute modularity classes for each node. Specifically, Gephi uses Louvain method (*) for computing modularity classes or, on other words, for finding clusters in the data. We can export this data for further analysis.
5. At this step, we have information about modularity class for each node computed using Louvain method. Therefore, we generated subgraphs based on the modularity class: for example, if we consider a modularity class number 2 as computed by Gephi, then we can take all nodes assigned to this class and generate a subgraph based on them. This way, we can see which groups of diagnoses are closer to each other based on co-occurrence. In other words, we can identify which diagnoses co-occur together more often.
6. For each network cluster, we visualized it separately and look for the nodes with the highest degrees (in our case, the most "important" or "well-connected" diagnoses of each cluster). This cluster visualization can be computed in Gephi. Moreover, we looked at each cluster and extracted full-text names of each diagnoses, and created a word cloud in Python to show the most frequently occurring words in each cluster. This way, we can obtain another high-level view of each cluster.

Approach for analysis to compute similarity between diagnoses using ICD-9 group codes

To compute associations between diagnoses, we represented each diagnosis as a vector, using the following steps,

1. First, because of a large number of diagnoses, we consider group codes instead of diagnoses codes. For example, ICD-9 code ‘40301’ will correspond to the group ‘390–459: diseases of the circulatory system’. For simplicity, we will use ‘390’, the starting code of the group, to code the diagnoses corresponding to the group ‘390–459’. To save these values, we create a new column ICD9_GROUP in the table DIAGNOSES_ICD.
2. Diagnoses is grouped (ICD9_GROUP) by patient ID. This way, we can locate the diagnoses associated with each patient.
3. We constructed a new table where each row which represents a patient, and each column represents a diagnosis group. This table will contain 18 columns corresponding to each diagnosis group. For each patient x and diagnosis d, value of the corresponding element of the matrix (x, d) will be equal to 1 if the patient has a diagnosed disease from this group, and 0 otherwise. Therefore, the resulting boolean matrix will show which patients were diagnosed with which types of conditions.
4. Using the obtained boolean table, we computed a measure of similarity between diagnoses. Because each diagnosis is represented by a Boolean vector, then instead of computing using Pearson method we can use metrics of dissimilarity applicable to Boolean variables. We chose Jaccard-Needham dissimilarity function from SciPy designed for Boolean arrays. Alternatively, if we have a computed dissimilarity v, then we can compute Jaccard similarity as 1-v.
5. For each pair of diagnosis groups, we computed the metric of Jaccard-Needham dissimilarity and record the obtained values in a new dissimilarity matrix.
6. Finally, we plot this matrix as a heatmap.

4.3 Tools and Technology

There are certain tools, programming languages and libraries utilized to evaluate the MIMIC-III dataset which are mentioned below,

Python

Python [15] is a powerful yet easy to learn scripting language utilized by programmers, researchers and scholars worldwide. Its strong set of libraries to evaluate and visualize large datasets makes it a right choice for data analysis tasks.

Pandas - python library

Pandas [13] is a library written for python programming language and widely utilized for data manipulation and analysis. It offers range of functionalities to produce operations on wide variety of data structures. This library is utilised in the project to evaluate MIMIC-III dataset which is in CSV format. The data format is not changed and Pandas library is utilised to work directly on the CSV format files.

Matplotlib – Python plotting library

Matplotlib [14] is library for plotting 2d figures and can be used in python scripts. With this library one can generate plots, histograms, bar charts, scatter plots and much more.

Seaborn - Python data visualization library

Seaborn is a python data visualization library and is based on Matplotlib library. Seaborn provides a much attractive and highly informative statistical visuals. Also, this library works closely with Pandas library making much more effective graphics with less effort.

NetworkX – Python package

NetworkX is a powerful python package for creating networks, manipulating, studying and analysing the structure and dynamics of network. It provides tools to study the structure with a standard programming interface and graph implementation that is suitable for many application.

Gephi

Gephi is an open source graph visualization and exploration platform. Its most used applications are in exploratory data analysis and network analysis visualisations. In this project network analysis is done using this platform utilizing its modularity feature.

6. Results and Discussions

This section details about the results found after complete analysis utilizing approach mentioned in *Approach* section and discussion is provided for each of the analysis. As mentioned in *Overview of MIMIC-III database* section, the dataset is divided in three levels of data. The analysis and results are discussed on the basis of these levels and one more section deals with hypotheses formulated.

5.1 Cross tabulation analysis of patient's survival on the basis of gender

Patient level data contains information relevant only to patients at the intensive care units. This includes information such as gender, date of birth, date of death, etc.

The total number of unique patients found in the database is 46250.

Gender distributions of the patients is evaluated which shows the count,

```
Male   26121 patients
Female 20399 patients
```

In the table inside dataset, EXPIRE_FLAG contains binary value which indicates whether the patient is alive or dead. These deaths include both deaths within the hospital (DOD_HOSP) and deaths identified by matching the patient to the social security master death index (DOD_SSN) [2].

When this EXPIRE_FLAG is evaluated, below result count is captured,

```
Alive  30761 patients
Dead   15759 patients
```

A cross tabulation using bar graphs is visualized on the above evaluation,

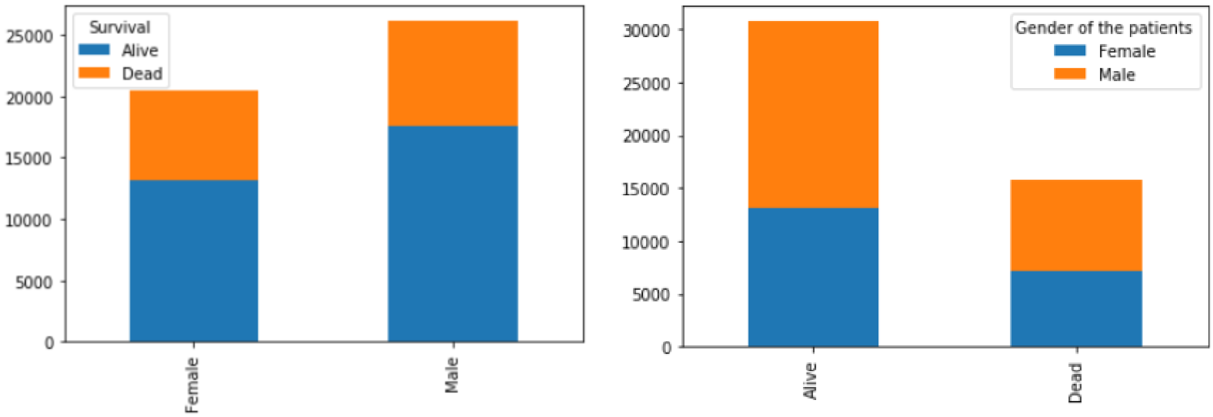


Figure 1: Cross tabulation: Gender vs Survival

Above figure shows the cross tabulation of gender vs survival of the patients. We can clearly see the following points from it,

1. Most of the patients are alive.
2. From the patients who are alive, there are more males than females.

5.2 Analysis of length of stay of patients in ICU using univariate analysis

Length of stay is quite important variable to evaluate when analysing data related to patients and hospital. In the case when we analyse patient's data related to intensive care unit, length of stay of patient is quite important. This will generally show the influence on the cost of an inpatient stay in regards with daily supplies and procedures.

A univariate analysis is conducted on the variable which shows length of stay of patient in the ICU. Please refer to *Approach* section for steps followed

Total of 46476 unique patients were found with 61532 unique ICU stays.

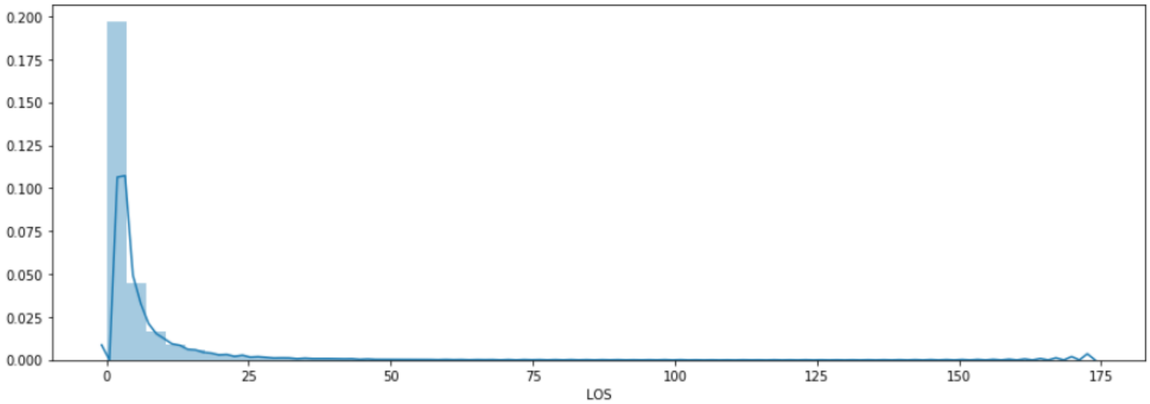


Figure 2: Histogram to show distribution of length of stay of patients in intensive care unit

The above figure shows a histogram to show univariate visualization of patient's length of stay in ICU. The graph shows length of stay (in days) vs the number of patients. Also, this is a "long tail" distribution: there is a large number of observations far away from the central part of the distribution.

To further analyse this variable descriptive statistics are produced as show in below table,

Table 2: Descriptive Statistics for Length of stay (in days) variable

Count	61522
Mean	4.9
Standard deviation	9.6
Min	0.0001
25%	1.1
50%	2.0
75%	4.5
Max	173.0

5.3 Understanding relation of variables associated with length of stay of patients

This analysis tells about how length of stay of a patient in ICU relates to associated variables mentioned below,

- ADMISSION_LOCATION
- INSURANCE
- MARITAL_STATUS
- DIAGNOSES

The details of approach for this analysis is provided in *Approach* section. Below is visualization and analysis of dependence of above mentioned variable on length of stay of patients.

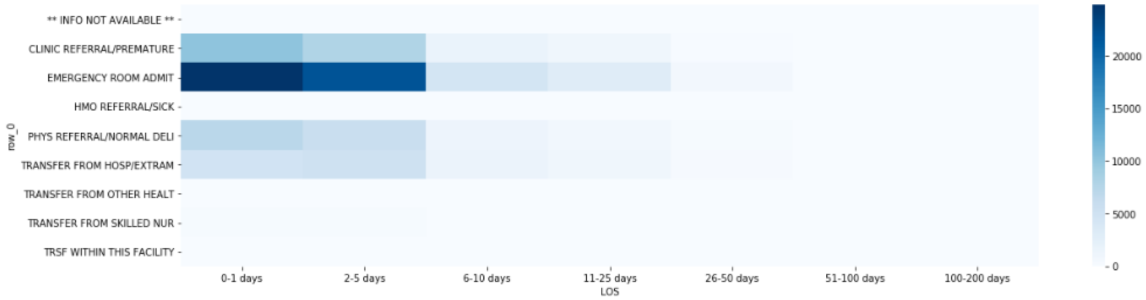


Figure 3: Heatmap to show cross tabulation of admission location and length of stay

For all types of admissions, most ICU stays were 5 days or below. Cross-tabulation additionally shows that most admissions were from the emergency room.

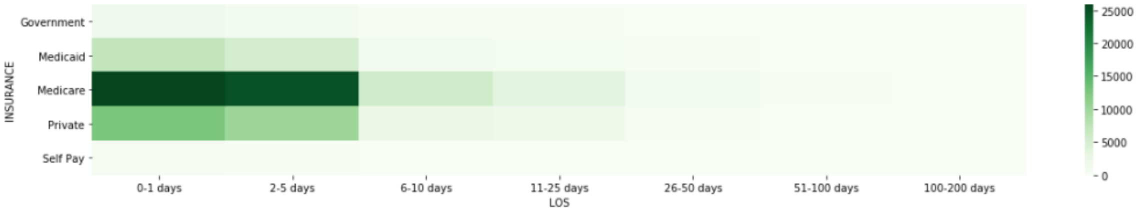


Figure 4: Heatmap to show cross tabulation of insurance and length of stay

Most admitted patients were on Medicare insurance. Cross-tabulation does not show any significant link between the type of insurance and the length of stay.

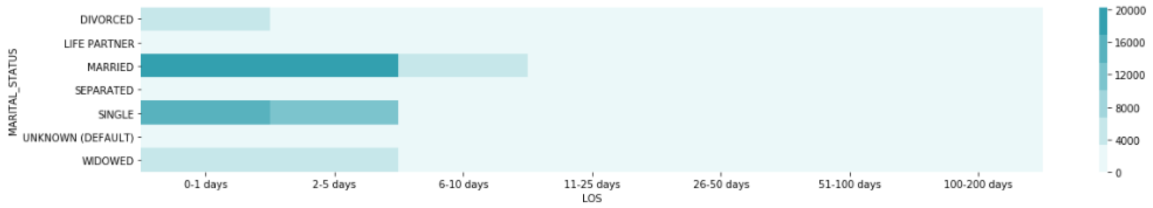


Figure 5: Heatmap to show cross tabulation of marital_status and length of stay

Cross-tabulation between marital status and length of stay does not show any dependence between these two variables. Most admissions were of the category "married", followed by "single".

Further, a random forest regressor is utilized to estimate the variable importance. The details for this method is provided in *Approach* section. Below graph shows the result to compare the variable importance,

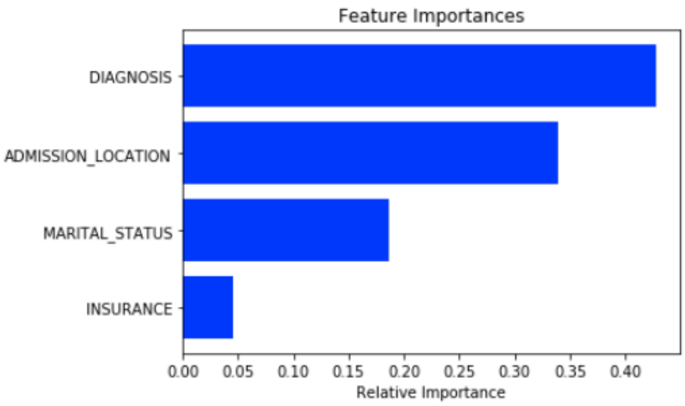


Figure 6: Graph to show feature importance generated using random forest regressor

We can clearly see here that diagnosis and admission location related to patient are more important (influential) than among other two factors for length of stay. The random forest regressor computes relative importance of each feature for the model (this is Gini importance, or mean decrease in impurity during the contraction of the random forest).

5.4 Comparative analysis of top diagnosis of patients in ICU to satisfy the survival rate hypothesis

Hypothesis formulation is part of exploratory data analysis to ensure that research remains scientific and reliable, to either confirm some observation or disapprove it. After looking at the top most common values for diagnosis of patients, following visualization is plotted,

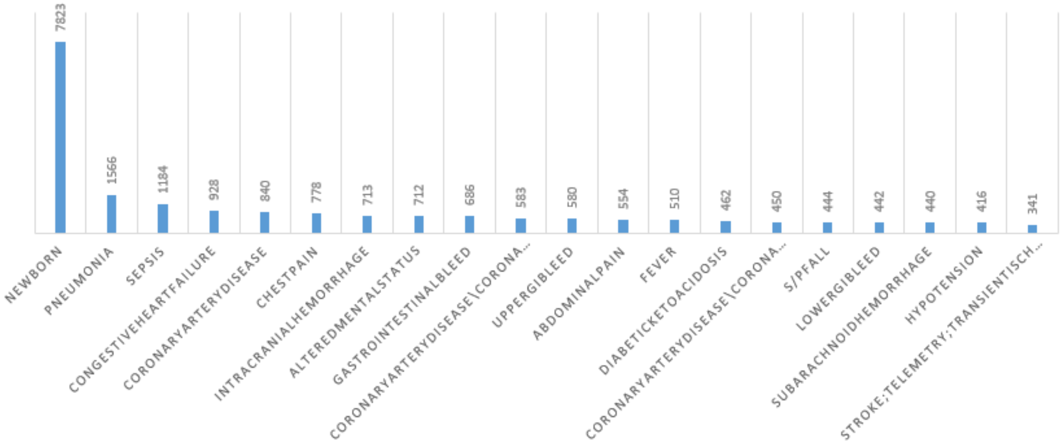


Figure 7: Graph to show number of patients with various diagnosis

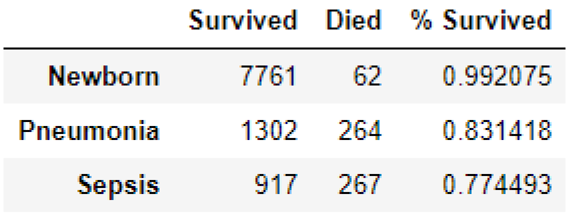
As we can see in the above figure the top three values are Newborn, Pneumonia and Sepsis. Another observation that is checked in the dataset is how many of the patients admitted died which is shown by HOSPITAL_EXPIRE_FLAG in the dataset. It was found that total 5854 patients died.

A comparative analysis is conducted between top most three common diagnoses i.e. newborns, pneumonia and sepsis patients to prove the hypothesis that *Newborns have a higher survival rate than other two most common diagnoses*.

Observations on each of the diagnosis type for the rate of survival was captured and compared for the comparative analysis to prove the hypothesis.

Following observations are captured,

Comparison: Newborn - Pneumonia – Sepsis



	Survived	Died	% Survived
Newborn	7761	62	0.992075
Pneumonia	1302	264	0.831418
Sepsis	917	267	0.774493

Figure 8: Table to show comparison percentage values for survival rate of top diagnoses

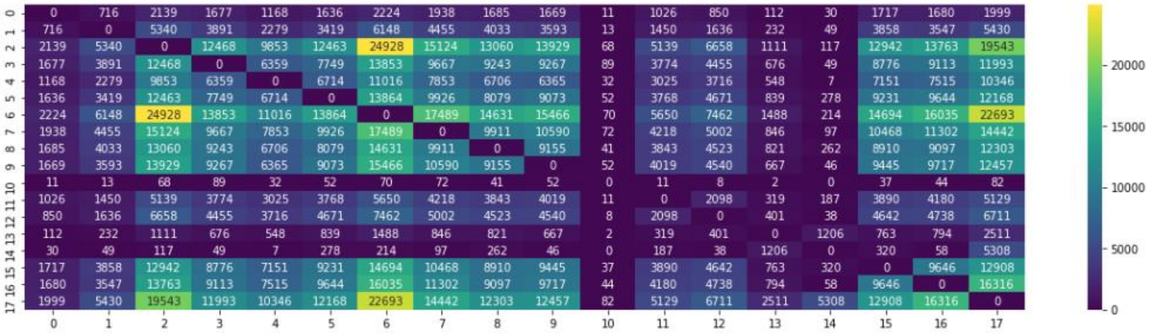
We can clearly see here that 99.2% of the newborns have survived, which confirms the hypothesis that mortality for pneumonia and sepsis was higher than newborn.

5.5 Network analysis of diagnoses for its co-occurrence in patients

The diagnoses information related to patient is present in the database as ICD-9 diagnoses codes with a total of 651,047 records. A network analysis is conducted on the diagnosis data present in MIMIC to evaluate and explore the properties of diagnoses which are connected to other diagnoses. This is very useful information to understand which diagnoses could lead to another one or co-occur with each other.

This analysis is divided into a high level analysis and a detailed analysis. For detailed information on this analysis please refer to *Approach* section.

The high level analysis is based on ICD-9 group codes and its co-occurrence in patients as a group type.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0	716	2139	1677	1168	1636	2224	1938	1685	1669	11	1026	850	112	30	1717	1680	1999
1	716	0	5340	3891	2279	3419	6148	4455	4033	3593	13	1450	1636	232	49	3858	3547	5430
2	2139	5340	0	12468	9853	12463	24928	15124	13060	13929	68	5139	6658	1111	117	12942	13763	19543
3	1677	3891	12468	0	6359	7749	13853	9667	9243	9267	89	3774	4455	676	49	8776	9113	11993
4	1168	2279	9853	6359	0	6714	11016	7853	6706	6365	32	3025	3716	548	7	7151	7515	10346
5	1636	3419	12463	7749	6714	0	13864	9926	8079	9073	52	3768	4671	839	278	9231	9644	12168
6	2224	6148	24928	13853	11016	13864	0	17489	14631	15466	70	5650	7462	1488	214	14694	16035	22693
7	1938	4455	15124	9667	7853	9926	17489	0	9911	10590	72	4218	5002	846	97	10468	11302	14442
8	1685	4033	13060	9243	6706	8079	14631	9911	0	9155	41	3843	4523	821	262	8910	9097	12303
9	1669	3593	13929	9267	6365	9073	15466	10590	9155	0	52	4019	4540	667	46	9445	9717	12457
10	11	13	68	89	32	52	70	72	41	52	0	11	8	2	0	37	44	82
11	1026	1450	5139	3774	3025	3768	5650	4218	3843	4019	11	0	2098	319	187	3890	4180	5129
12	850	1636	6658	4455	3716	4671	7462	5002	4523	4540	8	2098	0	401	38	4642	4738	6711
13	112	232	1111	676	548	839	1488	846	821	667	2	319	401	0	1206	763	794	2511
14	30	49	117	49	7	278	214	97	262	46	0	187	38	1206	0	320	58	5308
15	1717	3858	12942	8776	7151	9231	14694	10468	8910	9445	37	3890	4642	763	320	0	9646	12908
16	1680	3547	13763	9113	7515	9644	16035	11302	9097	9717	44	4180	4738	794	58	9646	0	16316
17	1999	5430	19543	11993	10346	12168	22693	14442	12303	12457	82	5129	6711	2511	5308	12908	16316	0

Figure 9: Heatmap to show co-occurrence matrix for 18 ICD-9 group codes

This co-occurrence matrix in above figure shows 18 groups of ICD-9 groups (between 0 and 17) on both axes, and the count of patients who had this diagnoses from both of these groups. As we can see on the heatmap: The most frequent co-occurrence is between group 2 and group 6. ICD-9 group code of group 2 is 240 and group 6 is 390. Please refer to *Appendix D* for the legend of group codes for ICD-9.

Further, as we can see in the table of ICU-9 codes:
240–279: endocrine, nutritional and metabolic diseases, and immunity disorders
390–459: diseases of the circulatory system
Both diagnoses are relatively frequent, and therefore their co-occurrence does not mean causation.
The least frequent co-occurrences with other diagnosis groups is for group 10: 630–679: complications of pregnancy, childbirth, and the puerperium.

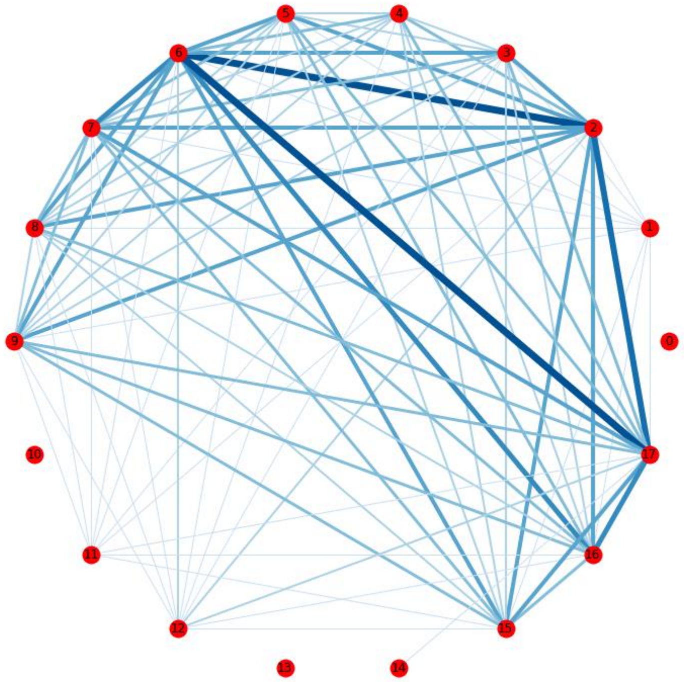


Figure 10: Network of 18 ICD-9 group codes to show co-occurrence between the groups

From the above figure we can see that each node is a diagnosis group. We see a strong co-occurrence between 6 and 2, and 6 and 17.

That means we see strong co-occurrence of,

- 390–459: diseases of the circulatory system and 240–279: endocrine, nutritional and metabolic diseases, and immunity disorders and
- 390–459: diseases of the circulatory system and E and V codes: external causes of injury and supplemental classification.

For detailed analysis for co-occurrence,

The network model represents the following,

- Nodes of the graph: one node = one diagnosis
- Diagnoses x and y are connected if there is at least one patient with both diagnoses
- The edge (x, y) has a weight W if there are exactly W patients which share both diagnoses, x and y

The above mentioned model is used to build network because our goal is to create a co-occurrence of diagnoses, so the best way is to model each diagnosis as a node in a graph. We have total 6985 unique diagnoses in patients.

In the below visualization, colors represent clusters of different diagnoses. This is calculated with Gephi using its modularity feature. One modularity class represents one cluster of diagnoses. Further, we can look at the nodes with the highest degree in each modularity class.

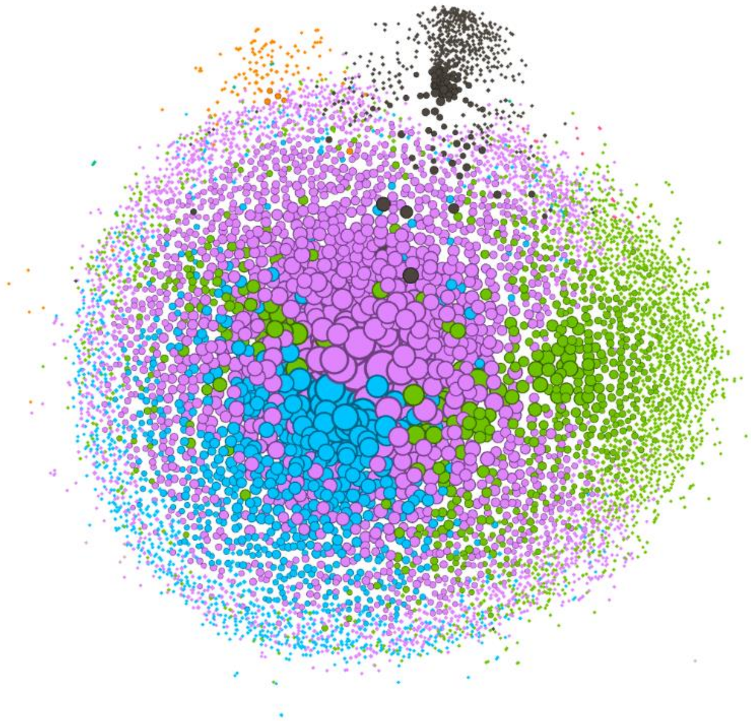
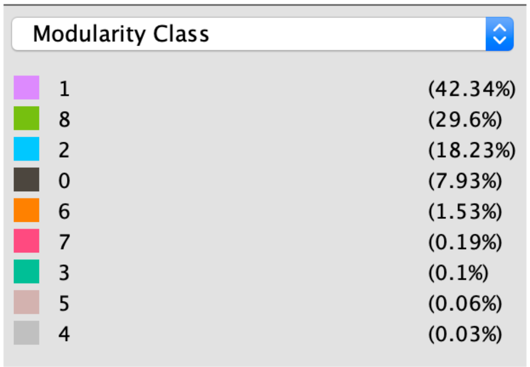


Figure 11: The figure shows total 6985 unique diagnoses clustered into 9 clusters using Gephi

Below table shows the various modularity class and its coverage for the above visualization. There are three most important modularity class (1, 8 and 2) with other smaller groups of diagnoses.



Modularity Class

🟪	1	(42.34%)
🟩	8	(29.6%)
🟦	2	(18.23%)
⬛	0	(7.93%)
🟧	6	(1.53%)
🟥	7	(0.19%)
🟩	3	(0.1%)
🟫	5	(0.06%)
⬜	4	(0.03%)

Figure 12: Legend for modularity class of visualization generated for network of diagnoses

Word cloud and visualization of ICD-9 codes for top three modularity class is shown below. We can notice that these bigger clusters are related to chronic and other serious conditions. However, there are smaller clusters of diagnoses related to pregnant women and childbirth. Additionally, there are cluster for accidents (open fracture, open wound). Please refer to *Approach* section for details on how below mentioned word cloud and cluster view is computed.

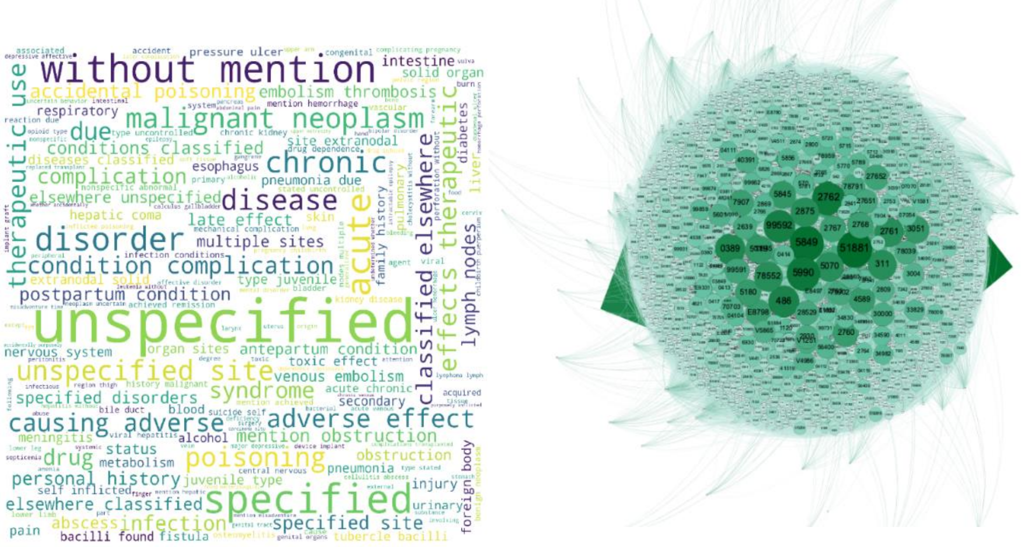


Figure 13: Word cloud and visualisation of network of diagnoses for cluster 1

As above figure shows, the most important nodes in cluster 1 are serious conditions such as acute kidney failure (5849), nephrotic syndrome (51811) that may be a sign of renal dysfunction, and urinary tract infection (5990). Therefore, we can see that some of the central nodes in this cluster are from the group of ICD-9 diagnoses 580–629: diseases of the genitourinary system. However, there are many conditions from other groups that co-occur with the above conditions: other important nodes in Figure 10 include pneumonia (486), acidosis (2762), severe sepsis (99592). In conclusion, cluster 1 shows conditions that co-occur with diseases of the genitourinary system in our dataset.

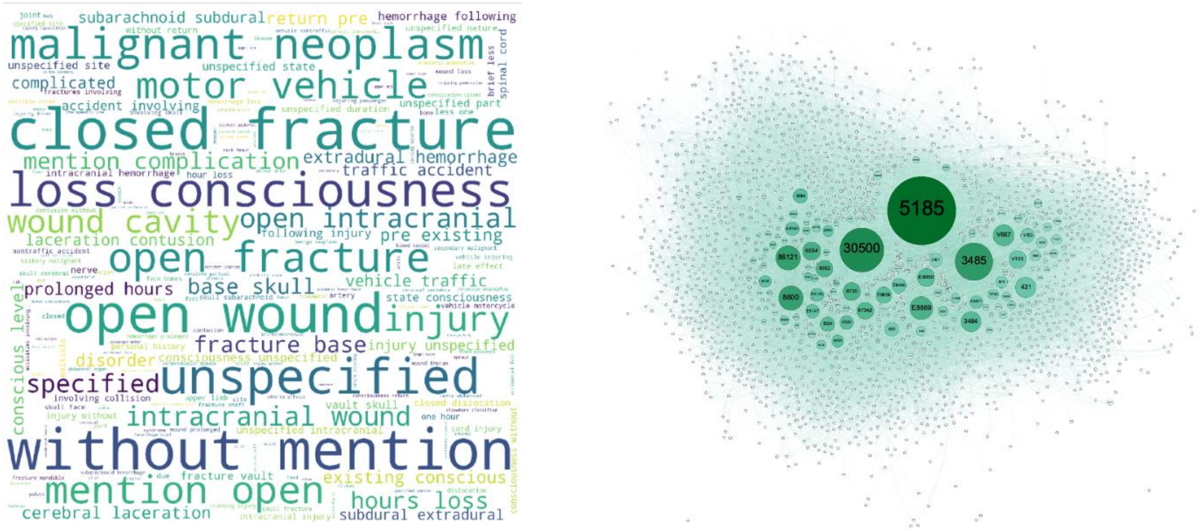


Figure 14: Word cloud and visualisation of network of diagnoses for cluster 8

According to the word cloud in above figure, this cluster includes wounds, fractures, and traffic accidents, as well as malignant neoplasms. The network in figure shows that the most well-connected node in cluster 8 is acute respiratory distress syndrome (5185), a type of respiratory failure that is connected to multiple other conditions, including sepsis, pancreatitis, trauma, pneumonia. Other important elements include alcohol abuse (30500), cerebral edema (3485), as well as contusion of lung (86121), traumatic pneumothorax (8600), closed fracture of lumbar vertebra (8054), and unspecified fall (E8889). In conclusion, the most important nodes in this cluster mostly include accidents and co-occurring conditions.

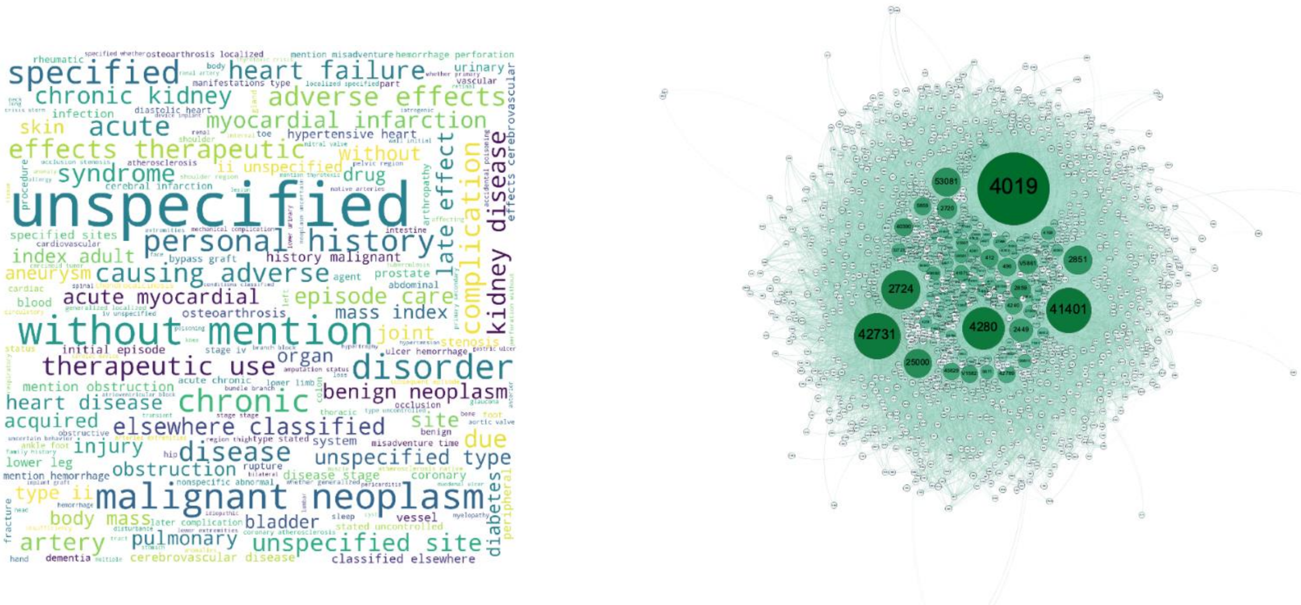


Figure 15: Word cloud and visualisation of network of diagnoses for cluster 2

As above figure shows, the most connected nodes in this cluster include hypertension (4019), coronary atherosclerosis of native coronary artery (41401), diseases of pericardium (42371), and congestive heart failure (4280) — in other words, those are diagnoses from ICD-9 group 390–459: diseases of the circulatory system. However, we can see diagnoses from other groups that are relatively important in this cluster and co-occur with the diagnoses above. Specifically, prominent nodes include hyperlipidemia - elevated number of lipids or lipoproteins in the blood (2724), acute post hemorrhagic anemia (2851), diabetes mellitus (25000). The word cloud in figure shows other conditions such as myocardial infarction, malignant neoplasm, heart failure, and heart disease. In conclusion, the most important nodes in this cluster mostly include conditions co-occurring with hypertension and other diseases of the circulatory system.

Also, as an example, we can see, the most unpopular cluster “4” contains only two diagnoses that co-occur with each other:

'Orbital granuloma'

'Orbital myositis'

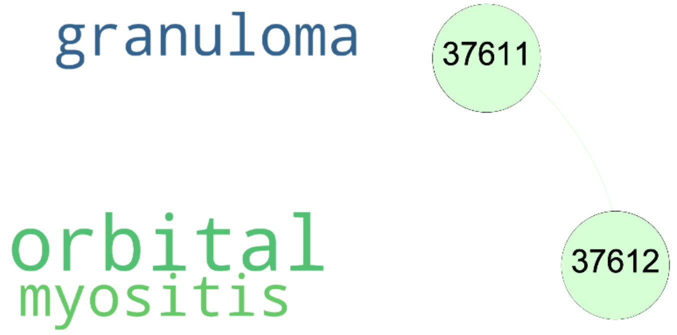


Figure 16: Word cloud and visualisation of network of diagnoses for cluster 4

As above figure shows, this cluster includes only two rare conditions: orbital granuloma (37611) and orbital myositis (37612). Both are in ICD-9 code “Disorders of the orbit” in the group 360-379 “Disorders of the eye and adnexa”. We can conclude that, according to our dataset, these conditions do not co-occur with other conditions often. Therefore, the clustering algorithm assigned them to a separate cluster.

5.6 Analysis to compute similarity between diagnoses using ICD-9 group codes

This section shows the result computed as a heat map from matrix for similarity between diagnoses groups.

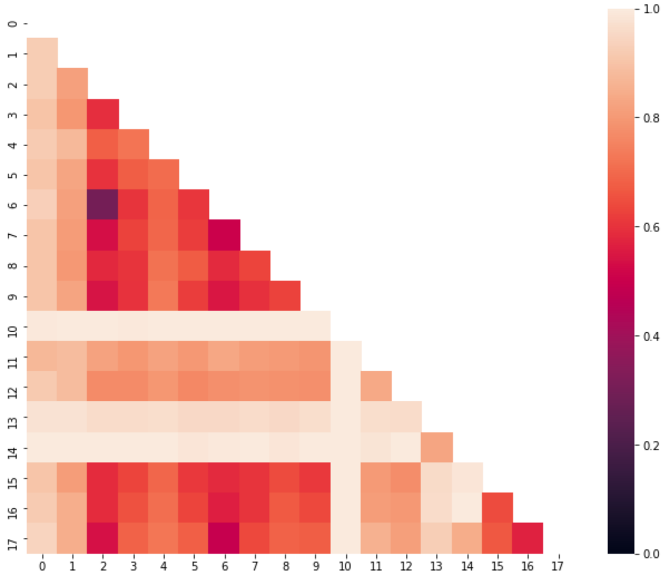


Figure 17: Heatmap to show similarity between diagnoses using ICD-9 group codes

Above figure shows values of Jaccard-Needham distance as a measure of dissimilarity between the diagnosis groups.

As Figure shows, the strongest dissimilarities with other diagnoses groups show the following diagnoses:

10 ('630'): 630–679: complications of pregnancy, childbirth, and the puerperium

14 ('760'): 760–779: certain conditions originating in the perinatal period

13 ('740'): 740–759: congenital anomalies (birth defects)

We can interpret is as follows: in the analyzed dataset, diagnoses from the listed groups co-occur with other conditions (such as chronic and age-related conditions) less frequently.

Additionally, as Figure shows, the strongest similarity is between groups 6 and 2:

6 ('390'): 390–459: diseases of the circulatory system

2 ('240'): 240–279: endocrine, nutritional and metabolic diseases, and immunity disorders

Therefore, according to the metrics we used, these diagnosis groups are certainly correlated to each other.

5.7 Potential outliers, anomalies, assumptions and other important results

This section summarizes the remaining areas of the database in form of potential outliers, anomalies, assumptions and may other important statistics and results.

- **Potential outliers and anomalies detected in the database**

There are some observations which shows that there could be potential anomalies and outlier values in the database related to length of stay.
From the graph described in section *Analysis of length of stay of patients in ICU using univariate analysis* we can see that most patients have stayed in the ICU for less than 25 days. However, there is a very low number of patients who stayed in the ICU for 150+ days. This is a set of potential outlier values.
This is described as below,

- The minimum stay was less than a day: 0.000100 day.
- The maximum stay was 173.07 days.
- The number of patients who stayed less than 1/1000 of the day is 8.
- The number of patients who stayed longer than 150 days is 7.

It is also been observed that data shows some of the patients have spent very few times in the ICU. For example, the patient 4226 was admitted at 20:38:21 and released at 20:38:51 on the same day. This can be a potential anomaly to look into.

We can also see that all dates are set in the future, which can be explained by the fact that the dataset is anonymized.

Also, it is also observed that data shows some patients have spent a lot of time in the ICU, longer than 150 days. We can also see that all dates are set in the future, which can again be explained by the fact that the dataset is anonymized.

- **Statistics and observations from various other important areas of the database**

Variable ADMISSION_TYPE speaks about the type of admission such as elective, urgent, newborn and emergency. This record represents under which category patient was pre-diagnosed and admitted. Emergency and urgent represents category when patient is admitted for unplanned medical care. Elective indicated a previous planned admission in hospital and newborn is for patient's birth.

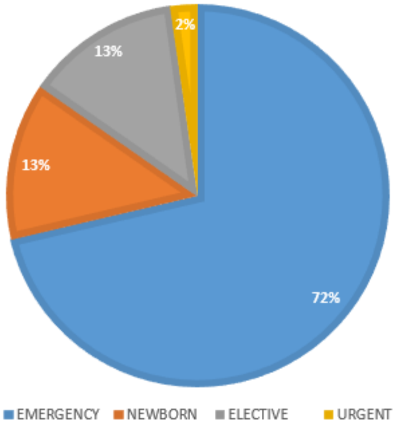


Figure 18: Pie chart to show distribution of patients with certain admission type

As we can see from above figure, 72 percent patients are admitted under emergency category which is usual for intensive care unit patients. Newborn and elective category share same amount of patient got admitted. A very small portion of patients goes under urgent type of admission.

Transfers from one care unit to another in an important variable to analyze. It reveals which are the most common units from there patient take transfers.

Below mentioned pie chart reveals what are top most category of units from where patient takes transfer (this does not include patients who are finally discharged or had not been provided any unit).

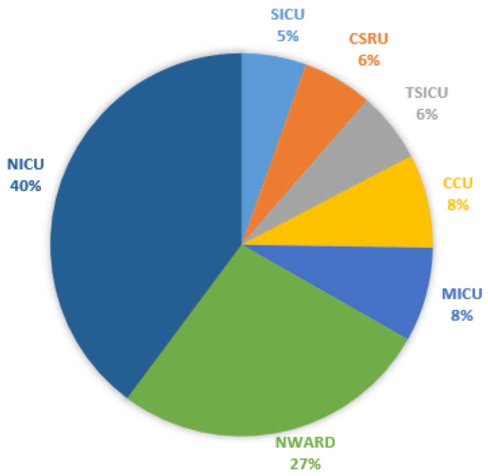


Figure 19: Pie chart to show distribution of patients admitted to certain units in hospital

We can clearly see here that NICU and NWARD shared the major part while other categories are minor ones. NICU and NWARD both are for neonatal care. This also reveals that majority of patients are admitted related to child birth or related procedures.

There is high possibility that a patient in the hospital is admitted under MICU care unit but services give to the patient is different and not being cared by the team at MICU. This could happen because of number of reasons such as bed shortage. It is good to know what services were given to a patient irrespective of admitted unit.

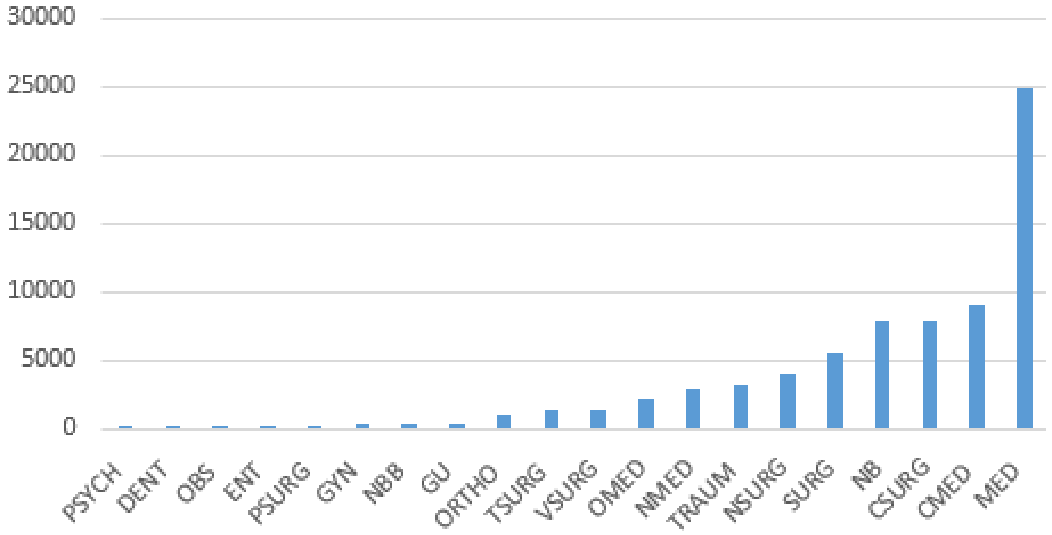


Figure 20: Bar chart to show distribution of patients provided with services in hospital

Above mentioned graph shows distribution of kind of services that were give to patient admitted to the hospital. We can see here that top four categories are MED, CMED, CSURG and NB. It’s obvious most patients to be under general medical service (MED) and shows the highest in the graph. The next two categories (CMED and CSURG) are related to cardiac services for medical and surgical. This also reveals that high number of patients receive heart or its closely related issues. NB is for new born patient’s which are admitted.

7. Conclusion

MIMIC-III is a very informative database for the researchers all around the world. It provide details of over 40,000 intensive care unit admissions which is sufficient to answer many questions related to critical patients who stay in intensive care units. The analysis done in this project has provided valuable results which would definitely help further deep researches using this database. The project has covered all the important aspects of exploratory data analysis.

At patient level of data, the results show the survival of patients admitted at intensive care units on the basis of gender of patients. This is basic but important factor to know what percentage of patients survive which could be first step in knowing the quality of care provided and other relevant factors.

At hospital and ICU level data, various important factors are analysed such as length of stay of patients, diagnosis network, and various factors associated with length of stay, admission type, and transfer from units and results are extracted.

A summary of results extracted is mentioned below,

- Length of stay of patients is affected by two important factors: diagnosis and admission location.
- Further, multivariate analysis of length of stay of patients shows relation with admission location and most admissions were in emergency room.
- Length of stay of patients in the database does contain some outlier values which could be possible anomaly as well.
- Newborns have better survival rate than patients diagnosed with Pneumonia or Sepsis.
- High-level analysis of network of ICD-9 group codes for Co-occurrence of diagnoses shows most co-occurrence between group 240 and 390.

- A detailed analysis of network for co-occurrence of diagnoses is created which shows most co-occurred and associated among patients were related to chronic, heart related and other serious conditions.
- Most patients were admitted under emergency category than newborn, elective or urgent.
- Most patients were initially admitted to NICU unit and takes most transfers.
- Most patients were provided general medical services with cardiac and natal services to follow.
- The database contains potential outliers and anomalies due to the reason that the database is de-identified.

Overall, MIMIC-III has been found to be very robust and easy to use database for further researches. It contains most required and relevant information which could be really helpful to analyze further. However, because certain data in MIMIC-III is anonymized to protect patient confidentiality, there is high possibility some of the information could not be extracted or misleading. Information such as patient's date of birth, date of death, date of admission, date of discharge etc. are de-identified which limits researchers to get some important analysis such as mortality of patients over a period or in a particular year. Also, there has been outliers which are also potential anomalies in the database. All these information in the database should be cleansed or an assumption should be required to make before diving deep for further researches.

8. Future Work

This project has provided a strong base for deep researches on the database. Going ahead analysis are to be conducted which provides holistic view of quality of services and other relevant factors in intensive care units. Below mentioned analysis are to be conducted,

- Analysing mortality of patients diagnosed with diseases acquired during stay in intensive care units such as ventilator associated pneumonia.
- Analysing factors associated with drug prescriptions provided to patients during stay which would be helpful in getting information regarding quality of medication.

References

[1] Patil, Prasad. What is Exploratory Data Analysis? [Internet]. March 2018. Available from: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[2] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635

[3] Tukey, John W. (1977). Exploratory Data Analysis. Pearson. ISBN 978-0201076165.

[4] Leinhardt, G., Leinhardt, S., Exploratory Data Analysis: New Tools for the Analysis of Empirical Data, Review of Research in Education, Vol. 8, 1980 (1980), pp. 85–157.

[5] S. H. C. DuToit,A. G. W. Steyn,R. H. Stumpf (1986) Graphical Exploratory Data Analysis. Springer ISBN 978-1-4612-9371-2

[6] Clifford, Gari D. Scott, Daniel J. Villarroel, Mauricio Villarroel. User Guide and Documentation for the MIMIC II Database. Feb. 2012. https://mimic.mit.edu/archive/mimic-ii-guide.pdf

[7] Dunitz, Max, George Verghese, and Thomas Heldt. "Predicting Hyperlactatemia in the MIMIC II Database." Conference Proceedings : ...Annual International Conference of the IEEE Engineering in Medicine and Biology Society.IEEE Engineering in Medicine and Biology Society.Annual Conference, 08/01/2015, pp. 985-988, doi:10.1109/EMBC.2015.7318529

[8] Zhang, Zhongheng, et al. "Predictive Value of Ionized Calcium in Critically Ill Patients: An Analysis of a Large Clinical Database MIMIC II." PloS One, vol. 9, no. 4, 2014, pp. e95204, doi:10.1371/journal.pone.0095204

[9] A. Taoum, F. Mourad-Chehade, H. Amoud and Z. Fawal, "Predicting ARDS using the MIMIC II physiological database," 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), Beirut, 2016, pp. 47-51.doi: 10.1109/IMCET.2016.7777425

[10] Lee, Joon, et al. "A Web-Based Data Visualization Tool for the MIMIC-II Database." BMC Medical Informatics And Decision Making, vol. 16, Feb. 2016, p. 15. EBSCOhost, doi:10.1186/s12911-016-0256-9

[11] Pollard, T. J. & Johnson, A. E. W. The MIMIC-III Clinical Database http://dx.doi.org/10.13026/C2XW26 (2016)

[12] Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. *Scientific Data* (2016)

[13] "Python Data Analysis Library – pandas: Python Data Analysis Library". pandas. Retrieved 13 November 2017

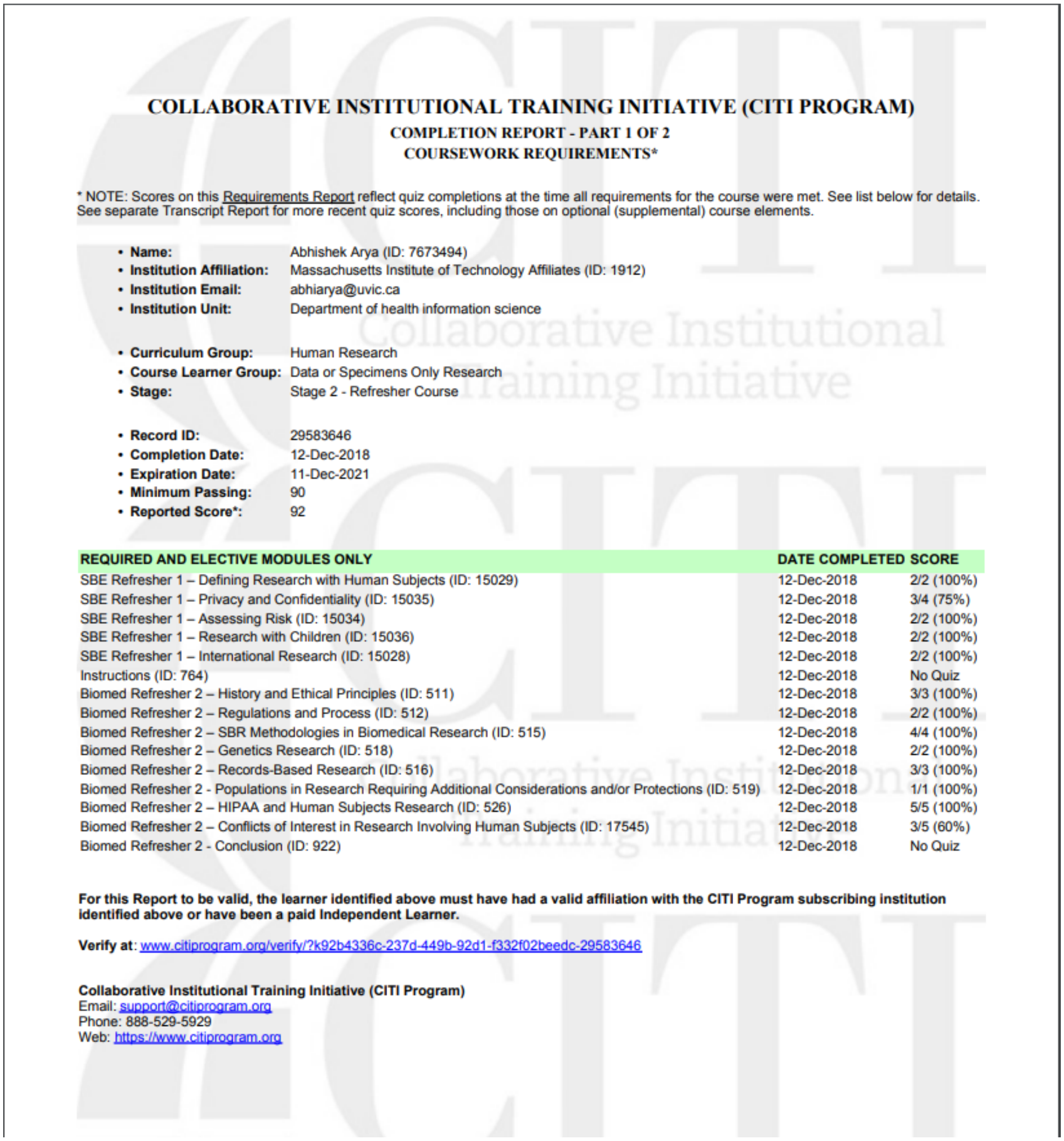
[14] "Matplotlib 2.2.2 documentation". matplotlib.org. Retrieved 2018-04-11

[15] "Overview of Hospital Stays in the United States", 2012, Audrey J. Weiss, Ph.D. and Anne Elixhauser, Ph.D

[16] "The impact of comorbid [corrected] conditions on critical illness", Esper AM1, Martin GS., Crit Care Med. 2012 Mar;40(3):1043

Appendix A

Below are the completion report and certification from CITI (collaborative institutional training initiative) before conducting the research on MIMIC-III database.



COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

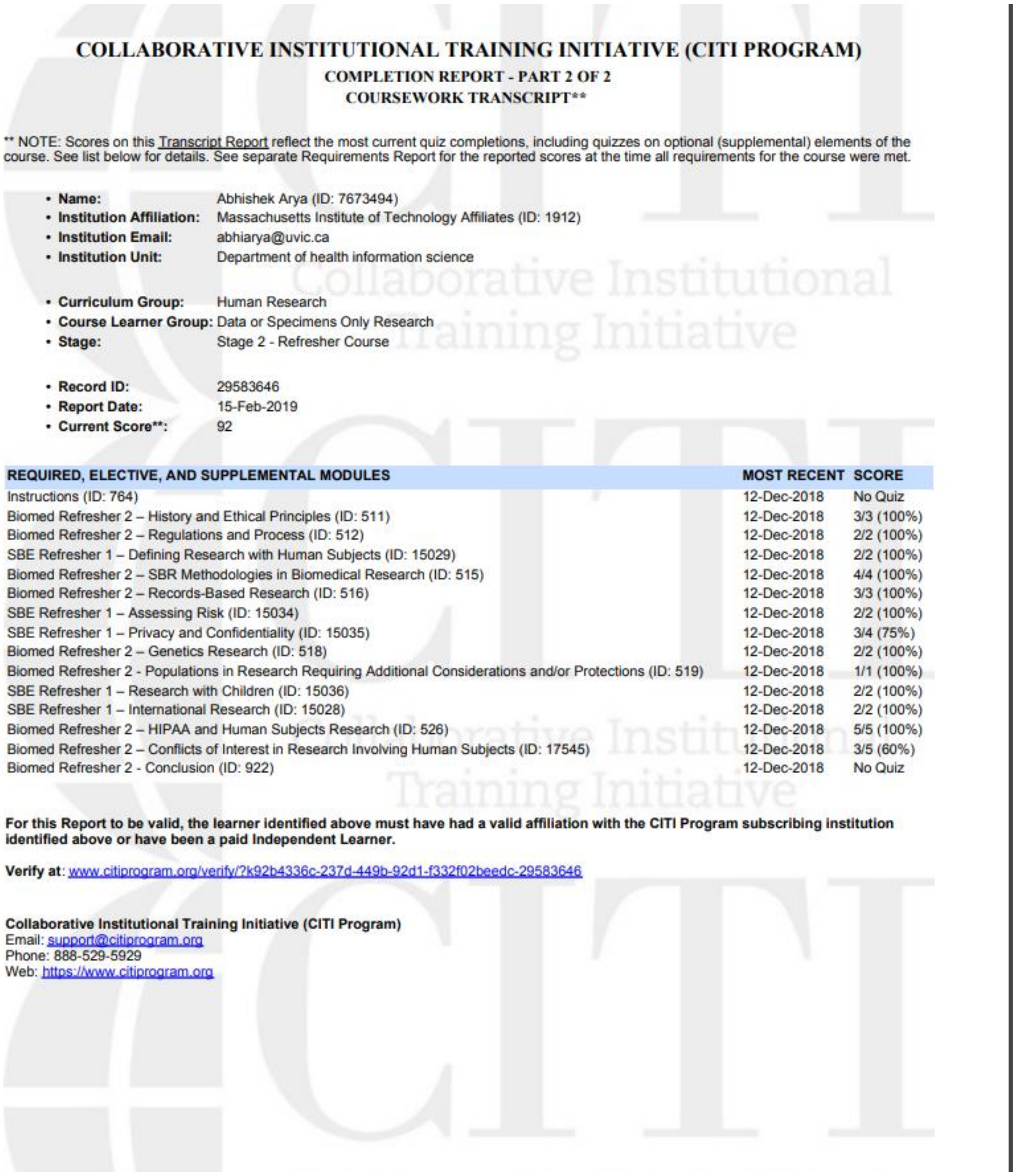
• **Name:**	Abhishek Arya (ID: 7673494)
• **Institution Affiliation:**	Massachusetts Institute of Technology Affiliates (ID: 1912)
• **Institution Email:**	abhiarya@uvic.ca
• **Institution Unit:**	Department of health information science
• **Curriculum Group:**	Human Research
• **Course Learner Group:**	Data or Specimens Only Research
• **Stage:**	Stage 2 - Refresher Course
• **Record ID:**	29583646
• **Completion Date:**	12-Dec-2018
• **Expiration Date:**	11-Dec-2021
• **Minimum Passing:**	90
• **Reported Score*:**	92

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
SBE Refresher 1 – Defining Research with Human Subjects (ID: 15029)	12-Dec-2018	2/2 (100%)
SBE Refresher 1 – Privacy and Confidentiality (ID: 15035)	12-Dec-2018	3/4 (75%)
SBE Refresher 1 – Assessing Risk (ID: 15034)	12-Dec-2018	2/2 (100%)
SBE Refresher 1 – Research with Children (ID: 15036)	12-Dec-2018	2/2 (100%)
SBE Refresher 1 – International Research (ID: 15028)	12-Dec-2018	2/2 (100%)
Instructions (ID: 764)	12-Dec-2018	No Quiz
Biomed Refresher 2 – History and Ethical Principles (ID: 511)	12-Dec-2018	3/3 (100%)
Biomed Refresher 2 – Regulations and Process (ID: 512)	12-Dec-2018	2/2 (100%)
Biomed Refresher 2 – SBR Methodologies in Biomedical Research (ID: 515)	12-Dec-2018	4/4 (100%)
Biomed Refresher 2 – Genetics Research (ID: 518)	12-Dec-2018	2/2 (100%)
Biomed Refresher 2 – Records-Based Research (ID: 516)	12-Dec-2018	3/3 (100%)
Biomed Refresher 2 - Populations in Research Requiring Additional Considerations and/or Protections (ID: 519)	12-Dec-2018	1/1 (100%)
Biomed Refresher 2 – HIPAA and Human Subjects Research (ID: 526)	12-Dec-2018	5/5 (100%)
Biomed Refresher 2 – Conflicts of Interest in Research Involving Human Subjects (ID: 17545)	12-Dec-2018	3/5 (60%)
Biomed Refresher 2 - Conclusion (ID: 922)	12-Dec-2018	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k92b4336c-237d-449b-92d1-f332f02beedc-29583646

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org



COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 2 OF 2

COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Abhishek Arya (ID: 7673494)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** abhiarya@uvic.ca
- **Institution Unit:** Department of health information science

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 2 - Refresher Course

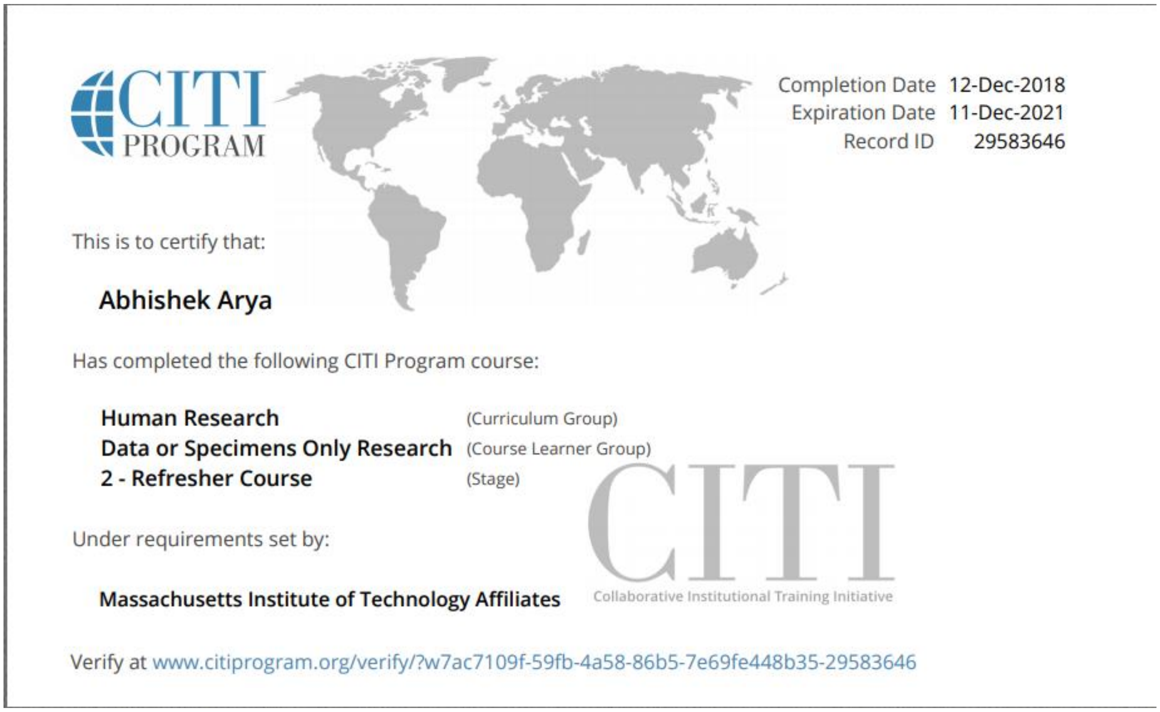
- **Record ID:** 29583646
- **Report Date:** 15-Feb-2019
- **Current Score**:** 92

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
Instructions (ID: 764)	12-Dec-2018	No Quiz
Biomed Refresher 2 – History and Ethical Principles (ID: 511)	12-Dec-2018	3/3 (100%)
Biomed Refresher 2 – Regulations and Process (ID: 512)	12-Dec-2018	2/2 (100%)
SBE Refresher 1 – Defining Research with Human Subjects (ID: 15029)	12-Dec-2018	2/2 (100%)
Biomed Refresher 2 – SBR Methodologies in Biomedical Research (ID: 515)	12-Dec-2018	4/4 (100%)
Biomed Refresher 2 – Records-Based Research (ID: 516)	12-Dec-2018	3/3 (100%)
SBE Refresher 1 – Assessing Risk (ID: 15034)	12-Dec-2018	2/2 (100%)
SBE Refresher 1 – Privacy and Confidentiality (ID: 15035)	12-Dec-2018	3/4 (75%)
Biomed Refresher 2 – Genetics Research (ID: 518)	12-Dec-2018	2/2 (100%)
Biomed Refresher 2 - Populations in Research Requiring Additional Considerations and/or Protections (ID: 519)	12-Dec-2018	1/1 (100%)
SBE Refresher 1 – Research with Children (ID: 15036)	12-Dec-2018	2/2 (100%)
SBE Refresher 1 – International Research (ID: 15028)	12-Dec-2018	2/2 (100%)
Biomed Refresher 2 – HIPAA and Human Subjects Research (ID: 526)	12-Dec-2018	5/5 (100%)
Biomed Refresher 2 – Conflicts of Interest in Research Involving Human Subjects (ID: 17545)	12-Dec-2018	3/5 (60%)
Biomed Refresher 2 - Conclusion (ID: 922)	12-Dec-2018	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k92b4336c-237d-449b-92d1-f332f02beedc-29583646

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org



CITI PROGRAM

Completion Date 12-Dec-2018
Expiration Date 11-Dec-2021
Record ID 29583646

This is to certify that:

Abhishek Arya

Has completed the following CITI Program course:

Human Research	(Curriculum Group)
Data or Specimens Only Research	(Course Learner Group)
2 - Refresher Course	(Stage)

Under requirements set by:

Massachusetts Institute of Technology Affiliates

CITI
Collaborative Institutional Training Initiative

Verify at www.citiprogram.org/verify/?w7ac7109f-59fb-4a58-86b5-7e69fe448b35-29583646

Appendix B

Below mentioned is the legend related to nodes with higher degree representing ICD-9 codes and its corresponding diagnosis long title for each cluster.

Cluster 0

'V290', 'Observation for suspected infectious condition'

'7742', 'Neonatal jaundice associated with preterm delivery'

'V053', 'Need for prophylactic vaccination and inoculation against viral hepatitis'

'V3000', 'Single liveborn, born in hospital, delivered without mention of cesarean section'

'V3001', 'Single liveborn, born in hospital, delivered by cesarean section'

'769', 'Respiratory distress syndrome in newborn'

Cluster 1

'5990', 'Urinary tract infection, site not specified'

'51811', Nephrotic syndrome, unspecified

'486', 'Pneumonia, organism unspecified'

'2762', 'Acidosis'

'99592', 'Severe sepsis'

'5849', 'Acute kidney failure, unspecified'

Cluster 2

'42371', 'Other diseases of pericardium’

'41401', 'Coronary atherosclerosis of native coronary artery'

'2724', 'Other and unspecified hyperlipidemia'

'25000', 'Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled'

'4280', 'Congestive heart failure, unspecified'

'4019', 'Unspecified essential hypertension'

Cluster 3

'63381', 'Other ectopic pregnancy with intrauterine pregnancy'

'63391', 'Unspecified ectopic pregnancy with intrauterine pregnancy'

'6392', 'Damage to pelvic organs and tissues following abortion or ectopic and molar pregnancies'

'6396', 'Embolism following abortion or ectopic and molar pregnancies'

'63320', 'Ovarian pregnancy without intrauterine pregnancy'

'V230', 'Supervision of high-risk pregnancy with history of infertility'

'6391', 'Delayed or excessive hemorrhage following abortion or ectopic and molar pregnancies'

Cluster 4

'37611', 'Orbital granuloma'

'37612', 'Orbital myositis'

Cluster 5

'68884', 'Other diseases of the skin and subcutaneous tissue'

'64914', 'Obesity complicating pregnancy, childbirth, or the puerperium, postpartum condition or complication'

'64294', 'Unspecified hypertension complicating pregnancy, childbirth, or the puerperium, postpartum condition or complication'

'64214', 'Hypertension secondary to renal disease, complicating pregnancy, childbirth, and the puerperium, postpartum condition or complication'

Cluster 6

'64821', 'Anemia of mother, delivered, with or without mention of antepartum condition'

'64822', 'Anemia of mother, delivered, with mention of postpartum complication'

'V270', 'Outcome of delivery, single liveborn'

'64891', 'Other current conditions classifiable elsewhere of mother, delivered, with or without mention of antepartum condition'

'66612', 'Other immediate postpartum hemorrhage, delivered, with mention of postpartum complication'

'64421', 'Early onset of delivery, delivered, with or without mention of antepartum condition'

Cluster 7

'63552', 'Legally induced abortion, complicated by shock, complete'

'65963', 'Elderly multigravida, antepartum condition or complication'

'63512', 'Legally induced abortion, complicated by delayed or excessive hemorrhage, complete'

'63572', 'Legally induced abortion, with other specified complications, complete'

'65583', 'Other known or suspected fetal abnormality, not elsewhere classified, affecting management of mother, antepartum condition or complication'

Cluster 8

'E8889', 'Unspecified fall'

'30500', 'Alcohol abuse, unspecified'

'86121', 'Contusion of lung without mention of open wound into thorax'

'8600', 'Traumatic pneumothorax without mention of open wound into thorax'

'5185', ‘Acute respiratory distress syndrome’

'8054', 'Closed fracture of lumbar vertebra without mention of spinal cord injury'

'3485', 'Cerebral edema'

Appendix C

Below mentioned is the glossary of acronyms and abbreviations used in the report.

Table 3: Glossary for acronyms and abbreviations

CMED	Cardiac Medical - for non-surgical cardiac related admissions
CSURG	Cardiac Surgery - for surgical cardiac admissions
DENT	Dental - for dental/jaw related admissions
ENT	Ear, nose, and throat - conditions primarily affecting these areas
GU	Genitourinary - reproductive organs/urinary system
GYN	Gynecological - female reproductive systems and breasts
MED	Medical - general service for internal medicine
NB	Newborn - infants born at the hospital
NBB	Newborn baby - infants born at the hospital
NMED	Neurologic Medical - non-surgical, relating to the brain
NSURG	Neurologic Surgical - surgical, relating to the brain
OBS	Obstetrics - concerned with childbirth and the care of women giving birth
ORTHO	Orthopaedic - surgical, relating to the musculoskeletal system
OMED	Orthopaedic medicine - non-surgical, relating to musculoskeletal system
PSURG	Plastic - restoration/reconstruction of the human body (including cosmetic or aesthetic)
PSYCH	Psychiatric - mental disorders relating to mood, behaviour, cognition, or perceptions
SURG	Surgical - general surgical service not classified elsewhere
TRAUM	Trauma - injury or damage caused by physical harm from an external source
TSURG	Thoracic Surgical - surgery on the thorax, located between the neck and the abdomen
VSURG	Vascular Surgical - surgery relating to the circulatory system
CCU	Coronary care unit
CSRU	Cardiac surgery recovery unit
MICU	Medical intensive care unit
NICU	Neonatal intensive care unit
NWARD	Neonatal ward
SICU	Surgical intensive care unit
TSICU	Trauma/surgical intensive care unit
ICU	Intensive care unit

Appendix D

Below is the legend for ICD-9 group codes,

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

320–389: diseases of the nervous system and 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 and V codes: external causes of injury and supplemental classification