**Assessment and Improvement of predictive value of complexity based features for sepsis prediction**

Priyanka Gupta
IIIT-D-MTech-CS-DE-15-047

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science

to

Indraprastha Institute of Information Technology Delhi

July, 2017

# Certificate

This is to certify that the thesis titled **"Assessment and Improvement of predictive value of complexity based features for sepsis prediction"** submitted by **Priyanka Gupta** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by her under our supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The results contained in this thesis have not been submitted in part or full to any other university or Institute for the award of any degree/diploma.


**Dr. Tavpritesh Sethi**

**Dept. of Computational Biology**

**IIIT Delhi**

**Dr. Vikram Goyal**

**Dept. of Computer Science**

**IIIT Delhi**

# Abstract

Data analytics is not embedded into critical environments such as Intensive Care Unit. It is now realized that there is an untapped potential of Big-data in improving patient care. However, this is not taken into practice due to lack of communication between the doctors and the data-analysts. To bridge this gap, we worked on Intensive Care Unit (ICU) data from two sources, one being numeric data from version 3.1 of the Multiparameter intelligent monitoring in intensive care (MIMIC II) waveform Matched subset and second one being Vital Monitoring data from the Pediatric Intensive Care Unit at the All India Institute of Medical Sciences (AIIMS), New Delhi.

This thesis presents unsupervised modelling on Numeric data of MIMIC Matched Subset, called Permutation Distribution Clustering (PDC). It is a complexity-based approach to cluster time series data. Clusters obtained by clustering are validated for their clinical potential and examined in order to determine the enrichment of sepsis. It is calculated using SIRS score which uses numeric and clinical data of MIMIC.

This work also involves development of a doctor-friendly dashboard that carries out analytics on the Vital Monitoring data from the Pediatric Intensive Care Unit at the All India Institute of Medical Sciences (AIIMS), New Delhi. It displays key patient indices calculated from the continuous data. The dashboard is updated regularly in an hourly fashion. The dashboard has deployed Quality Check, Basic Analytics and Advanced Analytics modalities. This dashboard is designed for Quality Improvement through display of basic Quality Check mechanisms such as patient information checks, technical anomalies and fidelity of sensor data. It is also designed to give care-related feedback to Intensivists and Nurses.

# Acknowledgements

I sincerely thank my supervisor Dr. Tavpritesh Sethi, IIIT Delhi for giving me the opportunity to work on this project and also his constant supervision and valuable guidance during the course of the thesis. I would also like to thank my co-supervisor Dr. Vikram Goyal, IIIT Delhi for his valuable insights and assistance. I would like to acknowldedge Dr. Rakesh Lodha, professor-in-charge of Pediatric ICU at AIIMS for giving us access to the ICU and for allowing us to test our pipelines live in the hightly critical ICU environment. A special thanks to Aditya Nagori, Ph.D. candidate at IGIB for his help and support in understanding and deriving patient cohorts during the duration of my thesis.
I would also like to thank my parents for providing me the motivation, support and encouragement during the entire duration of my thesis.

# Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Overview and Problem definition

The data in healthcare is rapidly increasing day by day, with increasing computerization of hospitals and digitization of records. Every piece of patient data can be valuable contribution to a learning healthcare system. The future of healthcare research lies in analysing and visualizing this data to improve early care of the patients. One large barrier to analysing of these data is inaccessibility to researchers. Even if available, there is a lack of trained professionals to understand and deal with electronic health data. A few databases have been made publicly after anonymization and can be accessed after fulfilling regulatory terms and conditions. These data available to clinicians and researchers offer a tremendous opportunity for research in medical or healthcare domain. However, as said earlier, these are not without challenges due to specialized requirements for understanding medical jargon, scoring criteria and well as codes such as International Classification of Diseases (ICD 10) [1].

Despite these challenges, many studies have demonstrated excellent potential of deriving insights from these data. For example, disease specific work has been done for acute kidney injury [2], acute respiratory distress syndrome [3] and septic shock [4]. The ability to utilize the EHR would allow learning systems. The possibility that patient specific data i.e. electronic medical record and clinical data can be fed into a population based database and provide real-time decision support for individual patients based on data from similar patients is very exciting. Clinicians and patients can take better

decisions with data resources and analytic pipelines in place [5].

Though EHR can support decision making capabilities, many clinicians fail to use it to its full ability due to inability to use traditional data processing methods with large datasets. As a solution to the increased complexity associated with this type of research, investigators work in collaboration with multidisciplinary teams including data scientists, clinicians and biostatisticians would be great contribution in healthcare research. Thus the objective of this thesis is in (i) taking away some of this complexity by providing dashboards for our research on hard-core computational analysis requiring data-science expertise, (ii) conceptualizing a scientific approach to derive complex features in the physiological signals for early detection of critical conditions In particular, this thesis focuses on a complexity based feature relying on permutation distributions of a time series for performing clustering and determine the enrichment of sepsis in obtained clusters.

## 1.2 Challenges

- **Massive dataset:** This is a typical example of "Big-data" where it grows in volume everyday and is complex and heterogenous and are computationally expensive.

- **Missing value: Being observational data,** the data from ICUs contain missing values. These may result from dislodgement of a probe. However, there are only a few probes in which data is missing more while others are completed. Fortunately, data such as heart rate and respiratory rate are more complete whereas, others like Oxygen saturation (Partial Pressure of O2,$SpO_2$), are missing more. Since permutation patterns are relatively robust against missing data, we did not interpolate missing data as medical data are known to be complex and nonlinear.

- **Accessibility:** As the patient data is confidential, the data is not shareable. However, Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) data is freely available for research work. The data are available with surrogates date and time. The patient's information are not shown, unique admission can be identified with subject_id and icu_stay_id in the data.

## 1.3 Contribution

The key contributions are:

- Unsupervised modelling of patient's time for determining the enrichment of sepsis.
- Fast implementation of pair-wise distance metric calculation between permutation distribution of time series.
- Development of a doctor-friendly prototype dashboard that carries out analytic on the Vital Monitoring data from the Pediatric Intensive Care Unit at the All India Institute of Medical Sciences (AIIMS).

## 1.4 Thesis Outline

In the next chapter, the background is discussed.In chapter 3, Unsupervised modelling on patient's numeric records and obtained results are being discussed. In chapter 4, Dashboard for safe-ICU is discussed. Finally, in the last chapter, conclusions along with the work is summarized with the future work.

# Chapter 2

# Literature Review

The literature is surveyed keeping an eye on two different viewpoints:

## 2.1 Studies on sepsis

Heart rate provides a way to look at autonomic nervous system function in the human body. There are small accelerations and decelerations of heart rate in healthy state. Disturbance in autonomic nervous function and abnormal heart rate variability (HRV) is a signal of pathology. Sometimes, the decreased HRV and short-term heart rate decelerations occurs simultaneously. There abnormal heart rate characteristics (HRC), decreased HRV and decelerations also occur in neonates with sepsis [8]. Also in [9] authors describes instant changes in HRV are correlated with various disease that require critical care, HRV estimation can be used to report the status of ill patients and also can be used to predict events resulting to any critical illness. However, these abnormality, short-term changes in the heart rate i.e. for fraction of seconds or abnormal HRC cannot be recognized trivially by monitoring the heart rate or electrocardiogram (ECG) signals. These events passes very fast from the screen of the ordinary monitor which cannot be observed by hospital staff (doctor or nurses).

University of Virginia's researcher developed a monitor (HeRO monitor) that analyses the bedside monitor's ECG and heart rate signals and find patterns that can predict forthcoming symptoms of sepsis. A hero score was presented on the monitor for clinical decision making. Data of 300 very low birthweight (VLBW) infants with over 100 sepsis events was used for defining

a mathematical algorithm which consist of three components [10-11,12-13] i.e. Standard deviation of inter-heartbeat (RR) intervals, sample asymmetry and sample entropy. The classical method of measure HRV is standard deviation of RR intervals which is one component of HeRO score. However, other work showed that only analysis of low and high frequency HRV does not enhance on the analysis of time series for the detection of the neonatal sepsis[14] and also failed to predicting early stages of sepsis. Consequently two other measures were also incorporated into the HeRO score algorithm. Thought the researchers developed score using infants data from one NICU, the HeRO score was also validated for infants form second NICU to detect early stage sepsis[8,9]. The HeRO monitor was approved by the US Food and Drug Administration in 2003 and by Europe in 2012 for measurement of HRC in infants and children in NICU. The monitor is updated on the hourly basis and score is calculated on the previous 12 hour of heart rate and ECG recordings. The monitors provides individual patient and multi patient's view. Individual view displays current HeRO score, the heart rate during previous 30 minutes, the trends in the score during the previous five days. Multi patient's view displays the above mentioned metrics for the cluster of patients in NICU.

The testing of the HeRO monitor was conducted from 2004 to 2010 [7] on VLBW infants from nine NICUs in US. The monitor testing involved randomly displaying or non-displaying the HeRO score of the infants (with the consent of parents) to the clinicians. The study says that there were no compulsory intervention in the treatment for the particular score or rising in the score, whereas the clinicians were well known about the development of HeRO score and evaluations on score rising events. After randomisation of 120 days, the primary outcome was days alive and not on mechanical ventilation and secondary outcome was mortality. Study with 3,003 infants resulted wtih 22% relative reduction in mortality (from 10.2% to 8.1%, p=0.04) for infants whose HeRO score was displayed. These results were not only statistically significant but also clinically important. Even outcomes of the HeRO monitoring were significant there were disadvantage of continuously displaying the risk of sepsis as a score on the monitor. It could make infants engage in more antibiotics intakes and medical evaluations.

There are disadvantage of HeRO monitoring, continuously displaying a number indicating risk for sepsis could result in infants undergoing more medical examinations and receiving more antibiotics. In the randomised clinical trial (RCT), infants in the HeRO display group had 10% more blood cultures obtained (1.8 compared with 1.6 blood cultures per month, p=0.05)

and 5% more days on antibiotics (15.7 compared with 15.0 days, p=0.31). Also in study for septicaemia[15], Authors describes that from 3003 infants randomized displaying or not displaying score, there are 700 infants who had Late-onset septicemia (LOS). The infants with one or more events of LOS in group of displaying HRC had more alive days compared to the group of non-displaying (76.4 vs 69.7 days, p=0.04). However, infants with LOS events with in group of displaying had more days of antibiotics intake compared to the non-displaying group (32 vs 29 days).

We decided to make our scores more specific by relying on stable patterns of longer lengths and to assess their value in predicting sepsis, thus motivating the work in the rest of the thesis.

## 2.2 Application of Permutation Distribution Clustering (PDC)

PDC is a complexity-based clustering approach that addresses how time series can be clustered on the basis of their permutation complexity patterns. In [16], author presents some application on of PDC on various datasets.

1. **Time Series Segmentation on Accelerometer Data:** The task was to detect continuous subsequences of time series that differ from the remaining segments. Accelerometer recording of two persons walking at different speeds on a treadmill were used for segment. Authors presents, full stops of treadmill were identified and different walking speed segments were also found.

2. **Clustering Electroencephalographic Data:** Permutation entropy applied to biomedical signals, used as the feature for classification task to predict absence seizure (X. Li, Ouyang, Richards, 2007) or the fetal behavioral states from biomagnetic recordings (Frank, Pompe, Schneider, Hoyer, 2006). In [16], Author apply PDC on data from the COGITO study (Schmiedek, Lövdén, Lindenberger, 2009) consists of EEG recordings of five participants in resting state under two condition eye open and eye closed. The task was to examine whether PDC can find partitions of the underlying dynamics of the two conditions in a high-dimensional data set. The result showed that PDC was able to do that and also PDC can operate well on raw data as the data was not artifact-corrected.

3. **Clustering Electrocardiographic Data:** The PDC was applied on two dataset ECG-I and ECG-II from Keogh, Lonardi, and Ratanamahatana (2004) to find utility of PDC for the analysis of biophysical signals . These datasets consist of univariate time series of ECG recordings. First dataset, consists of 20 time series of length 2,000 which were 10 randomly extracted time series sequences from two ECG datasets. The authors ground truth was two type of ECG were splitted on the top level of hierarchical clustering. The same applies to the ECG-II that contains ECG recordings of four different persons, 20 time series of length 10,000. A hierarchical clustering should be able to distinguish these four different sources in the data set. Authors shows that for both dataset ECG I and II, PDC clustering results matches with the results of original author's compression based methods reported by Keogh, Lonardi, and Ratanamahatana (2004).

4. **Clustering on the KDD 2004 Data Set:** Keogh, Lonardi, and Ratanamahatana (2004) also presented an important data set [17] for the evaluation of clustering algorithms. This dataset consist of time series from various source that include ECG data, the number of sunspots per year and yearly power demand of different countries. Author compares the results of PDC using complete linkage and embedding dimension of m = 6 with best clustering approaches reported by the original authors from 51 different approaches they tested. The performance was measured on the basis of **Q-measure** defined by original authors, which counts the number of correctly retrieved pairings at the lowest clustering level divided by the total number of pairs. Thus, the defined measure assess the known structure of the lowest clustering level. A compression-based clustering was clearly superior on the data set, yielding a Q = 1.0, however some significant results were also achieved by the dynamic time warping with Q = 0.33, piecewise linear approximation with Q = 0.33, simple Euclidean distance with Q = 0.27, LPC Cepstra, autocorrelation with Q = 0.16, and LCSS with Q = 0.33. PDC with an embedding dimension of 5 can keep up very well with 15 out of 18 time series in perfect pairs at the lowest levels (Q = 0.78).

5. **Clustering Natural Language Texts:** PDC is applied on a dataset consist of natural language texts in different languages from Keogh, Lonardi, and Ratanamahatana (2004). The original author presents

that compression based clustering obtained the clusters of natural language that are similar with hierarchical classification of languages from linguistics. Author, demonstrate that PDC can also achieve important results in this domain. For applying PDC to natural language texts, the language text needs to be converted into numeric time series. Any method which assume any prior information or add any parameter cannot be considered as the preprocessing step in this approach. Therefore, author convert each in the text with its character encoding number. Text were available Latin 1 format (ISO, 1987) which encodes 191 distinct characters in 8 bits per character. The standard alphabets can be represented as between 65-90 (A-Z) and 97-122 (a-z). The space character is encoded with number 32 which is smaller value compared to other numbers, space character encodes the word boundaries. The encode value of punctuation is between space character encoding and standard character encoding. The special characters have higher encoding value than the standard characters. Authors applied PDC on two dataset, first data set contained 50 chapters of Genesis in the languages Dutch, German, Norwegian, Danish, Spanish, Italian, Latin, French, English, and Maori. The length of the time series was about 130,000. The min entropy was found on dimension 8 for this text data. The second data contained textual content of yahoo portals in different language. The length of time series is 1,600 in this case. The min entropy was found on dimension 6 in second dataset. PDC is able to reveal inherent structure in this data set and is able to cluster the different language text.

Therefore, because of established usefulness of PDC in some of related tasks, we decided to assess its utility as a feature for indicating critical outcomes such as sepsis. The next chapter describes the details of the method and its application to ICU data conducted in this thesis.

# Chapter 3

# Mining Permutation Patterns for Unsupervised Clustering of Multivariate Physiological Time Series

## 3.1 Introduction

In this chapter, we describe the MIMIC dataset and unsupervised approach, permutation distribution clustering (PDC) applied on time series data (vital sign of patient in ICU) to find enrichment of sepsis. Figure 3.1 illustrates Flow diagram of the data analytic on MIMIC dataset.
We now describe the details of the work in the following sections.

### 3.1.1 Dataset

The MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care)[25] Databases contain two database Waveform Databases which consists of physiologic signals and vital signs time series (numeric records) captured from bed-side monitors, and Clinical Database collected from medical information systems of hospitals. The numeric records (time series) were recorded at the resolution of one minute. The dataset consist of tens thousands of Intensive Care Unit (ICU) patients. Data were collected between 2001 and 2008 at the Beth Israel Deaconess Medical Center. Waveform Database records

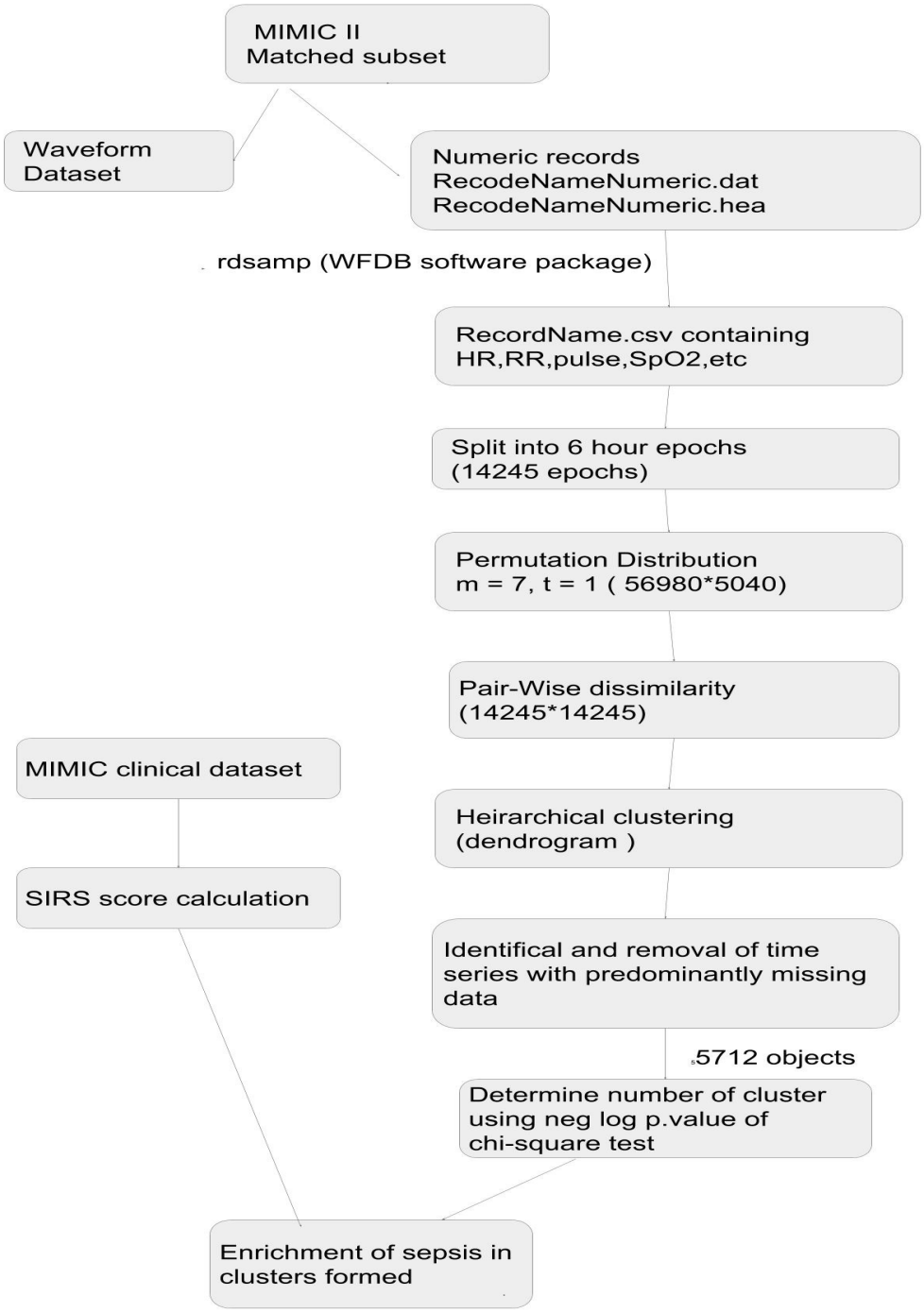


Figure 3.1: Flow Diagram of data analytic on MIMIC dataset to find enrichment of sepsis

of MIMIC II matched with clinical database are available in the MIMIC II matched subset database. Importantly, MIMIC III which is the newer version does not have waveform data and is under the linking process. Hence this thesis work is based on MIMIC II matched subset.

The MIMIC II Waveform Database Matched Subset contains all MIMIC II Waveform Database records that have been associated with MIMIC II Clinical Database records.It is a publicly available dataset containing 4,897 waveform records and 5,266 numeric records matched with 2,809 MIMIC II Clinical Database records. Each record directory with the name record consists of two records a waveform and associated numerics record that have been matched with a single Subject ID in the clinical database. The records are in format of RecordNameNumeric.dat file along with a corresponding header file RecordNameNumeric.hea. Numerics include heart and respiration rates, $SpO_2$, pulse and noninvasive blood pressure (raw as well as systolic, diastolic, and mean). Recording lengths are also different, most are a few days in duration, but some are shorter and others are several weeks long. Numeric records matched to corresponding clinical dataset are used for unsupervised modelling in this work.

MIMIC II Clinical Database is also used in the work. Data or Specimens Only Research course of Collaborative Institutional Training Initiative (CITI) [18] certification is needed to get the access the clinical dataset. The certification is provided after complete the course as an MIT affiliate, which needs to submit on physionet website in dataset access section. After that, Physionet provides the data access for research purpose. The following list gives an overview of the data categories:

- **Demographics:** At the time of patient's admission to hospital, the details like marital status, religion and insurance status are recorded.
- **Medication records:** All medications prescribed during the patient stay is recorded. It is digitalize or manual. Manual prescription is recorded for pharmacy orders.
- **Fluid records:** Fluids withdrawn from/to a patient are also recorded. This information provides physicians an accurate measure of a patient's fluid levels, which can help in better monitoring and treatment.
- **Notes:** MD, nursing notes and discharge summaries are recorded in database.

- **Chart:** A patient's medical chart contain any parameters recorded by the staff: validated physiologic recordings, demographic information, weight, height and ventilator settings are examples of the type of information recorded here.

- **Laboratory tests:** Body fluid test, blood gas are recorded.

### 3.1.2 Permutation Distribution Clustering (PDC)

Permutation distribution clustering (PDC) is a complexity-based clustering approach that addresses how time series can be clustered on the basis of their permutation complexity patterns. As a broad approach, permutation distribution is used as the feature to represent time series. It is defined using the probabilities associated with all possible patterns in the embedded time series. The dissimilarity between permutation distribution describes the dissimilarity between the time series. A metric approximation of the Kullback-Leibler(KL) divergence, squared Hellinger distance is used as metric of dissimilarity. The parameters used in the approach are determined by entropy-based heuristic, which make this possible for user to free from choosing parameter as well as searching best parameters for specific data.
We now describe the details of the permutation distribution and the subsequent distance metric used for clustering.

#### Permutation Distribution (PD)

The PD is interesting candidate for dissimilarity measure for time series due to its several properties.

- Invariant to monotonic transformation, basic mathematical operations applies on the the time series does not change its PD. It also implies that normalization, subtracting the mean and dividing by standard deviation is not needed as the preprocessing step.

- Invariant to phase, the occurrence of patterns are important but not when they occurs in the time series.

- Compressed representation, it allows comparison between time series of different length.

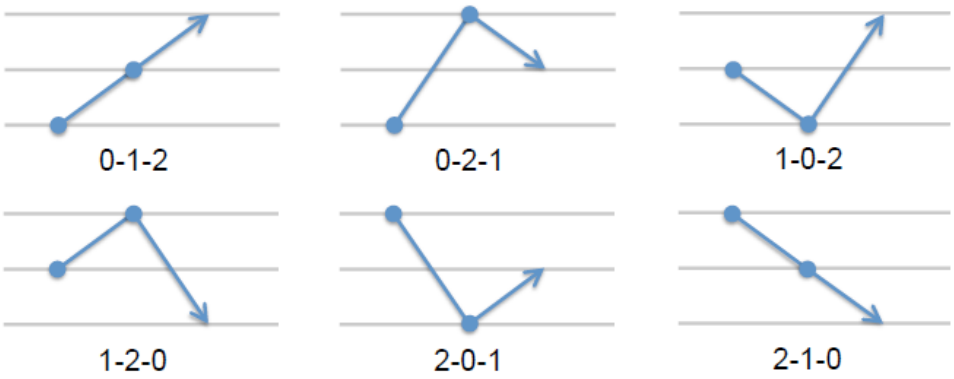


Figure 3.2: All possible patterns with embedding dimension of three.

Briefly, the PD are the probability of certain patterns of the ranks of values which occur in a time series. The time series can be single variate or multivariate (more than one variable). The time series is splitted into fixed length(m) sub-sequence, referred as m-space embeddings. Every $t^{th}$ value in the time series is consider as start of sub-sequence. The rank of values are calculated for sorted observed values for each sub-sequence. The frequencies of distinct rank patterns (also called ordinal patterns) are counted and each frequency is divided by total number of patterns in the time series to form permutation distribution. Each possible rank pattern is a permutation of the values from 0 to m-1. Figure **??** illustrates ordinal patterns for an embedding of m = 3. Given an ordinal time series $X = \{x_t | t = 0...T\}$, the time series embedding in dimension m and time delay t can be represented as,

$$X' = \{[x_t, x_{t+1}, x_{t+2}, ..., x_{t+(m-1)}] | t = 0...T - (m-1)\} \tag{3.1}$$

The rank or ordinal pattern of a subsequence $x' \in X'$ can be obtained by sort m value and if two indices have the same value then keep them in original order. There are m! possibility of original patterns in the time series. The frequencies of these patterns of the elements $x' \in X'$ divided by the total number of possible ordinal patterns present in the time series i.e. m! form permutation distribution.

### Dissimilarity Measure

The dissimilarity between two time series can be defined by dissimilarity between permutation distribution of corresponding time series. The Kullback-Leibler (KL) divergence (Kullback Leibler, 1951; Shannon, 1951), also known as relative Shannon entropy, is often used to determine divergence between probability distributions. Entropy calculated between two probability distribution is always greater than zero or zero (between similar distribution). This is not possible with KL divergence, as it does not follow triangle inequality and also asymmetric in nature. Therefore, KL can not be used as metric to calculate dissimilarity. Therefore, the squared Hellinger distance is used. Let P = (p1, p2, . . . , pn) and Q = (q1, q2, . . . , qn) be two permutation distributions. The squared Hellinger distance can be calculated as the Euclidean norm of the difference of the square root vectors of the distinct probability distributions,

$$D(P,Q) = \frac{1}{\sqrt{2}}||\sqrt{P} - \sqrt{Q}||_2^2 \tag{3.2}$$

Taylor approximation of KL divergence can be used to derived squared hellinger distance. It can be used as metric due to its symmetric nature, non-negative and also it satisfies the triangle inequality. The pair-wise squared Hellinger distances between the set of permutation distributions creates a distance matrix that can be input to a clustering algorithm.

### Clustering

Sequential agglomerative hierarchical non-overlapping clustering (S. Johnson, 1967) forms a tree structure of clusters. Initially, each time series is considered as a cluster with a single member. Based on a chosen dissimilarity measure, closest clusters iteratively merges to form cluster consists of elements of both subclusters until there is only a single top cluster left. This leads to a hierarchy of clusters. Distances between clusters are calculated using linkage function. There are common linkage functions (Halkidi, Batistakis, Vazirgiannis, 2001) are defined below. Let a set of objects O, clusters Ci $\subseteq$ O,

- Average linkage dissimilarity, calculated as the average distance between all pairwise distances of the cluster's objects:

$$d_{\text{average}}(C_{\text{i}}, C_{\text{j}}) = \frac{1}{|C_{\text{i}}||C_{\text{j}}|} \sum_{x \in C_{\text{i}}, y \in C_{\text{j}}} D(x, y) \tag{3.3}$$

- Single linkage dissimilarity, calculated as the minimum distance between objects of the clusters.

$$d_{\text{single}}(C_{\text{i}}, C_{\text{j}}) = \min_{x \in C_{\text{i}}, y \in C_{\text{j}}} D(x, y) \tag{3.4}$$

- Complete linkage similarity, calculates as the maximum distance between objects of the clusters.

$$d_{\text{complete}}(C_{\text{i}}, C_{\text{j}}) = \max_{x \in C_{\text{i}}, y \in C_{\text{j}}} D(x, y) \tag{3.5}$$

The resulting tree structure represents the hierarchical clustering is called as dendrogram. It is binary tree, every inner node in the tree has two children node, represents the merging of two clusters. The leaf node are the initial time series that are to be clustered. The distance between two cluster is represented by the height of merging in the dendrogram.

### 3.1.3 Systemic Inflammatory Response Syndrome (SIRS) score calculation

SIRS score defines the severity of sepsis. SIRS value ranges between 0 and 4. SIRS is defined as 2 or more of the following variables:

- Temperature greater than 38°C (100.4°F) or less than 36°C (96.8°F).
- Heart rate of greater than 90 beats/minute.
- Respiratory rate greater than 20 breaths/minute.
- Abnormal White blood cell count (>12,000/$\mu$L or < 4,000/$\mu$L or >10% immature [band] forms)

SIRS scores were calculated using MIMIC clinical dataset based on above criteria. Median of above mentioned variables are calculated for a time period of patient data. If any of the above condition satisfy, 1 point is added to a maintained counter. After checking every condition, final counter is considered as SIRS score for that time period of patient data.

## 3.2 Methodology

### 3.2.1 Data Cleaning and Preprocessing

As the numeric data was available in .dat and .hea format. Therefore, first step of Data cleaning and preprocessing was to convert this data into integer format. The function name rdsamp from WFDB software package is used to convert these files into csv format containing numeric value of various time series with their corresponding timestamps. The numeric data contains many missing time series for patients. Therefore, only those patients are considered for whom all the four time series data namely, heart rate (HR), oxygen saturation ($SpO_2$), respiratory rate (RR) and pulse rate (PR) are available. Despite this check, there may be intermittently missing data in time series. These were not discarded from the analysis. A peculiar feature of MIMIC dataset is that missing data have been represented as numeric zero (0). Since this is not possible for vital signs such as heart rate and $SpO_2$ in a living person, and may introduce errors, all zeroes in the time series are also replaced by NA (missing value). After implementing these steps, the data for 2081 out of total of 2809 patients are further considered.

Patients with no clinical data available in the MIMIC database are discarded as this data is needed to calculate the Systemic Inflammatory Response Syndrome (SIRS) scores. After implementing all the pre-processing steps, **1916 patients** were carried forward for analysis. This data was split into six hour epochs starting from start-time of the numeric data, which represents the time of admission into the ICU in-time. As numeric data is recorded at the resolution of 1 minute i.e. 60 data point per hour, each epoch consists of 360 data points of HR, PR, RR, $SpO_2$. The epochs with more than 50% missing data are discarded. After implementing above pre-processing, a total of **14,245 epochs** for n = 1,916 patients were taken forward. A 3D matrix was generated using all epochs with dimension $360 \times 14245 \times 4$ accounting for 360 number of data point in each epoch, 14245 total epochs and 4 time series in each epoch. All pre-processing and calculations were performed using R and supporting packages such as dplyr, lubridate etc.

### 3.2.2 Fast implementation of PDC

R package called pdc is available for computing PDC for time series. The main function of the package pdc in R is pdclust, which performs a hierarchi-

cal clustering of a set of time series based on differences in their permutation distributions. This function calls codebook function to generate permutation distribution of time series followed by pdcDist function to calculate the pairwise dissimilarity between the time series using Hellinger distance (symmetric Alpha Divergence). However, the computation of distance matrix from the permutation patterns pdcDist is not efficient. pdcDist function iterates over all objects to calculate pairwise dissimilarity and hence is computationally inefficient. In our implementations, codebook function is used to generate permutation distribution for multivariate time series which is followed by generation of dissimilarity matrix by using a computationally more efficient self-written R code instead of using pdcDist from package pdc. This code uses matrix multiplication, matrix addition and matrix modification like square and square root to calculate pairwise dissimilarity. Finally, hclust function is called for hierarchical clustering analysis.

Let P = (p1, p2, ... , pn) where pi = (wi,xi,yi,zi) be the multivariate permutation distribution and A is matrix consists of n such permutation distributions with dimension of 4n × m!. We improved the speed of the implementation by calculating the pairwise dissimilarity in these n distributions using the following steps:

- Partition the A matrix in four matrix corresponding to all variables (W,X,Y,Z). One matrix now is of dimension of (n×m!).
- For each of four matrix, we calculate
  1. W = (4*(1 - sqrt($WW^T$)))2
  2. X = (4* (1 - sqrt($XX^T$)))2
  3. Y = (4* (1 - sqrt($YY^T$)))2
  4. Z = (4* (1 - sqrt($ZZ^T$)))2
- Add all four matrix and take the square root of matrix.
  A = sqrt(W + X + Y + Z)

A matrix obtained is of size n × n, pairwise dissimilarity matrix. This code is efficient in term of time because matrix operation execute in parallel in the processor, however iteratively code (using for loop) execute sequentially.

### 3.2.3 PDC on multivariate time series using all four variables

After pre-processing the MIMIC data, PDC is applied with m = 7 (embedding dimension), t = 1 (time delay) on the data of dimension 360 × 14,245 × 4. The output of pdc consists of a tree which can be visualized as dendrogram, permutation distribution of all objects (an object corresponds to one patient's six hour epoch which contains multivariate time series) and a dissimilarity matrix between all objects.

### 3.2.4 Shuffling of time series to create a null-distribution of patterns followed by multivariate PDC of all four variables

Permutation is performed on every single time series in each object and PDC is computed again with m=7, t=1. Though the output for permuted data also consists of a tree, permutation distributions and a dissimilarity matrix, however, the results are distinct from the ones obtained using the original time series.

### 3.2.5 Identification and removal of time series with predominantly missing values in multivariate PDC

Dendrogram with shuffled time series should not contain any distribution and should like uniform. The objects which form uniform distribution are extracted from the dendrogram of original time series and SIRS scores are calculated for each object, considered to determine the enrichment of sepsis. Chi-square test is performed between cluster membership of object and SIRS scores of object. Chi-square test is performed for number of clusters (k) $\in$ {2,50}.

### 3.2.6 PDC on bivariate time series and shuffled bivariate time series

The heart rate (HR) and respiratory rate (RR) series are extracted from above created data for each object. The data created is of dimension 360 ×

14245 × 2. PDC is applied again with m = 7 and t = 1 for original and shuffled time series.

### 3.2.7 Identification and removal of epochs with predominantly missing values in bivariate PDC

Again, objects are selected from the dendrogram of bi-variate time series by looking uniform objects in the dendrogram of shuffled bi- variate time series. These selected objects are used to find the enrichment of sepsis.

## 3.3 Results

### 3.3.1 PDC on multivariate time series using all four variables

The output of PDC on multivariate time series with m = 7 is shown as dendrogram in figure 3.3. The y-axis represents the value of distance metric between clusters. For example, if two clusters merge at a height xx, it means that the distance between those clusters is xx. The max height of the tree is 7.9 implies maximum dissimilarity in the dataset is 7.9. Figure 3.4 shows the first 20 objects from the dendrogram, leaf node are patients epochs in the format of subjectId_icuStayId_epochNumber.

### 3.3.2 Shuffling of time series to create a null-distribution of patterns followed by multivariate PDC of all four variables

The output of PDC on shuffled multivariate time series with m = 7 is shown as dendrogram in figure as shown in figure 3.5. For the permuted time series, the entire dendrogram should have represented a uniform distribution as obtained inside the highlighted red box shown in figure 3.5. There are 5712 multivariate time series are present insides the red box. The max height (distance between cluster) is 8.

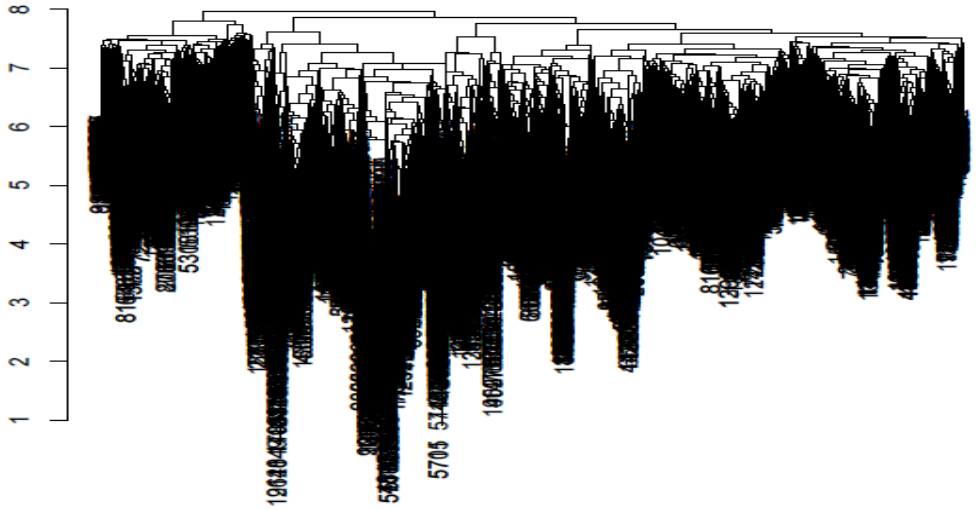


Figure 3.3: Dendrogram on multivariate time series using all four variables with m = 7

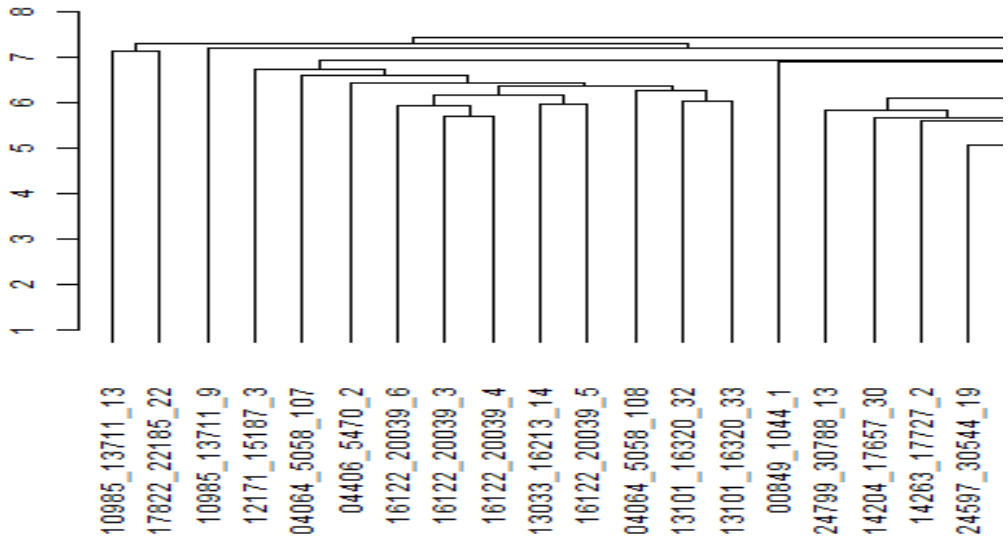


Figure 3.4: First 20 objects form dendrogram on multivariate time series using all four variables with m = 7. Leaf nodes represents patient epoch in format of subjectId_icuStayId_epochNumber

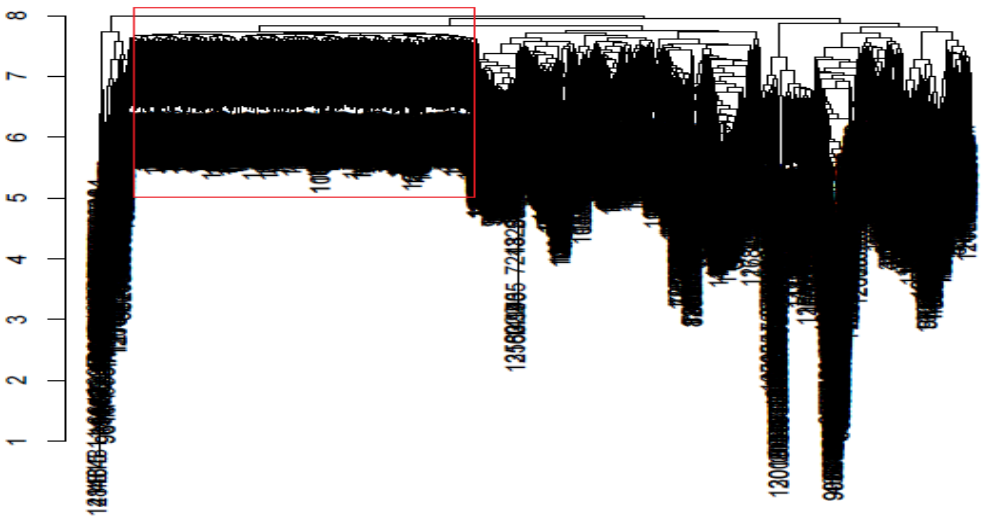


Figure 3.5: Dendrogram on shuffled multivariate time series using all four variables with m = 7.

### 3.3.3 Distribution of percentage of missing data in multivariate PDC

Figure 3.6 shows the distribution of percentage of missing data for objects outside and inside the red box in the dendrogram of permuted time series in figure 3.5. This shows that missing data percentage is large for objects in non-uniform part of the dendrogram as compare to the uniform part.

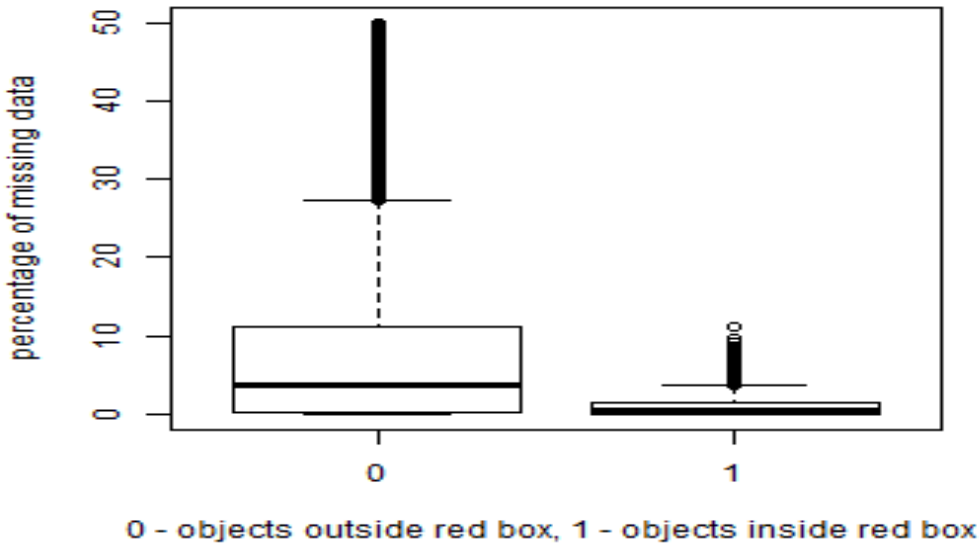


Figure 3.6: Distribution of percentage of missing data for object inside and outside the red rectangle

### 3.3.4 Negative log p.value of chi-square test for identified multivariate time series

Figure 3.7 and figure 3.8 shows negative log of p.value of chi-square test for all k $\in$ {2,50} for original time series and permuted time series. It can be shown from plots that there is no pattern present in case of permuted time series and also the p.value does not go beyond 0.6, however there is clear pattern or enrichment of sepsis is present in case of original time series and p.value reached maximum value of 2.021812 at k = 28.

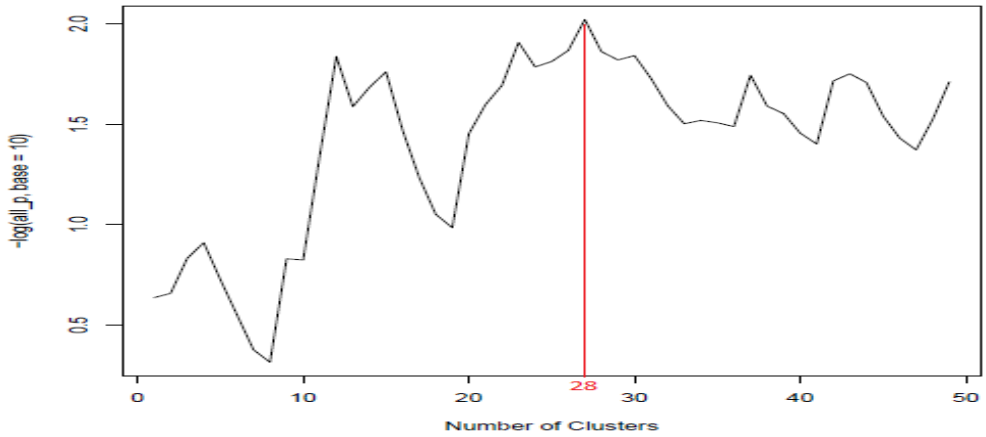


Figure 3.7: Negative log p.value for original time series

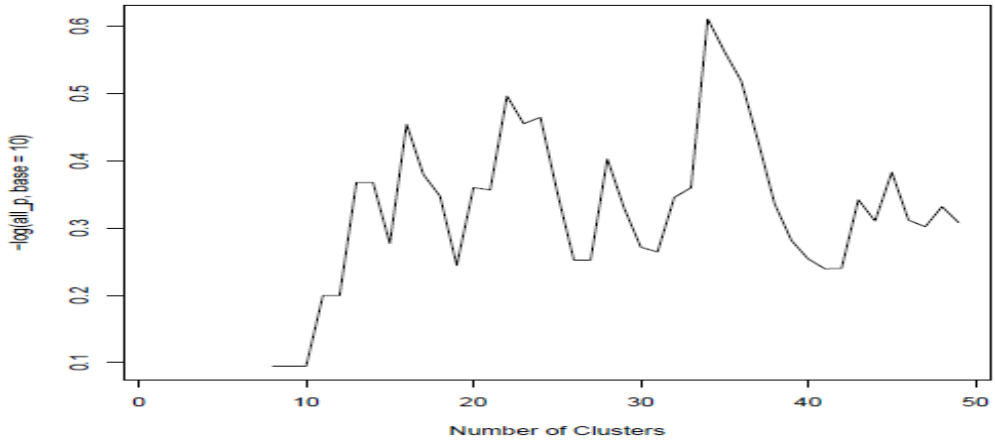


Figure 3.8: Negative log p.value for permuted time series

### 3.3.5 Distribution of SIRS score in multivariate PDC clusters

The distribution of SIRS score in all clusters is shown in Figure 3.9 and proportional to the number of object in one cluster is shown in figure 3.10. The clusters are of mixed SIRS score.

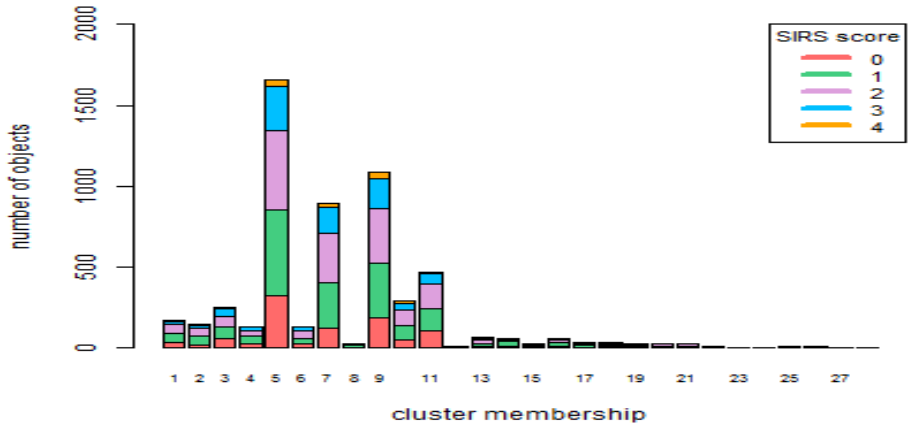


Figure 3.9: Distribution of SIRS score in clusters

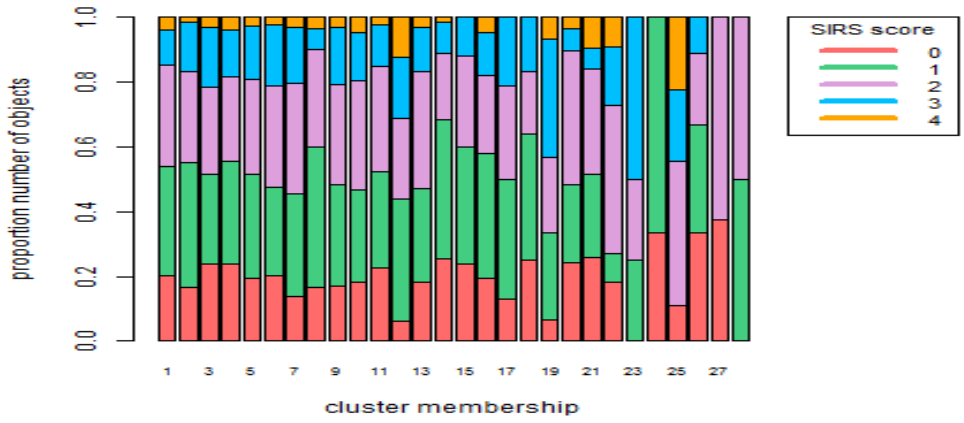


Figure 3.10: Proportional distribution of SIRS score in clusters

### 3.3.6 PDC on bivariate time series

The output of PDC on bivariate time series is shown as dendrogram in figure 3.11. The max height of the tree is 5.6.

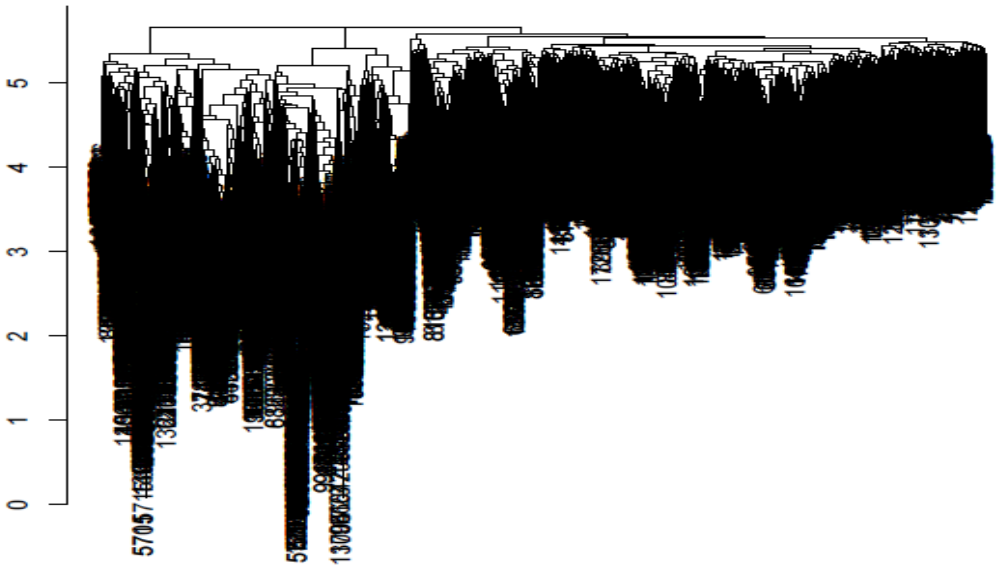


Figure 3.11: Dendrogram on original data with heart rate (HR) and respiratory rate (RR)

### 3.3.7 PDC on shuffled bivariate time series

The output of PDC on shuffled bivariate time series is shown as dendrogram in figure as shown in figure 3.12. There are 10,194 bivariate time series are present insides the orange box. The max height (distance between cluster) is 5.7.

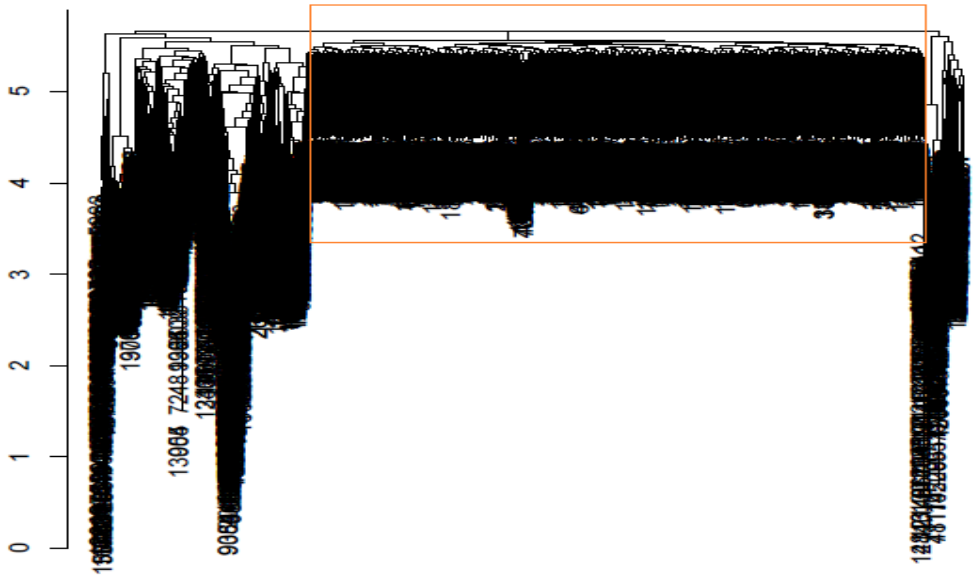


Figure 3.12: Dendrogram on permuted data with heart rate (HR) and respiratory rate (RR)

### 3.3.8 Distribution of percentage of missing data in bivariate PDC

Figure 3.13 shows the distribution of percentage of missing data for objects outside and inside the orange box in the dendrogram of permuted time series fig 3.12. This shows that missing data percentage is large for objects in non-uniform part of the dendrogram as compare to the uniform part.

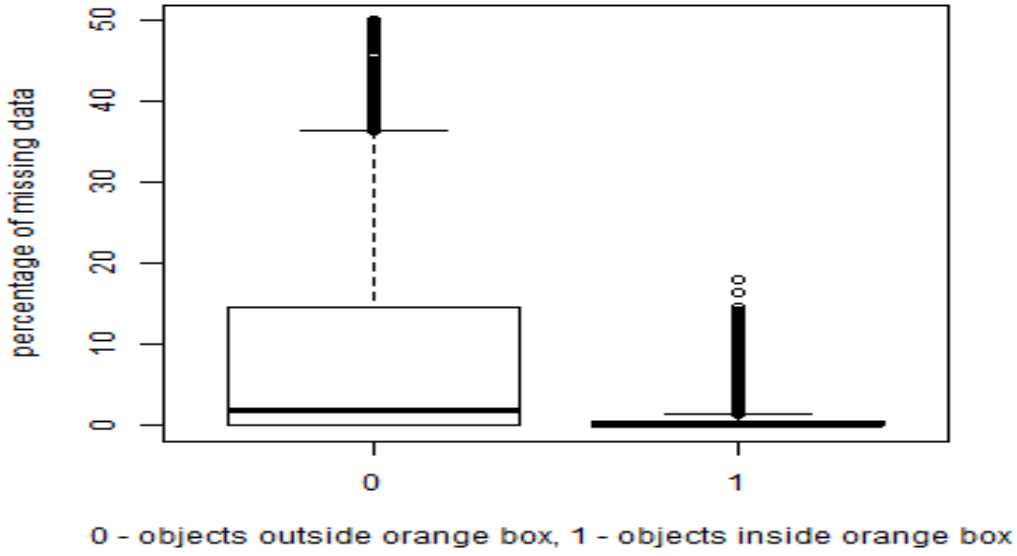


Figure 3.13: Distribution of percentage of missing data for object inside and outside the orange rectangle

### 3.3.9 Negative log p.value of chi-square test for identified bivariate time series

Figure 3.14 and figure 3.15 shows negative log of p.value of chi-square test for all k $\in$ {2,50} for original time series and permuted time series with two time series i.e. HR, RR. In this case with permuted time series the p.value does not go beyond 0.39, however in case of original time series p.value reached maximum value of 2.5 at k = 30.

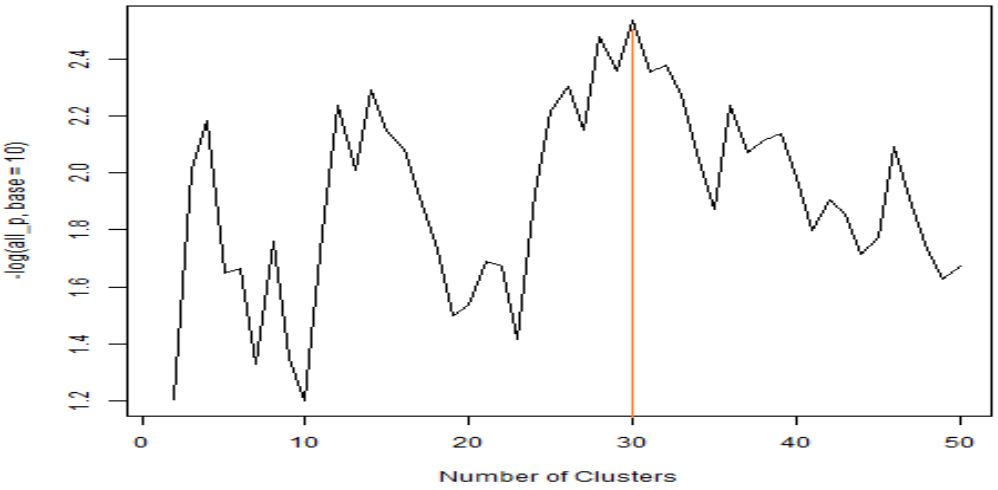


Figure 3.14: Negative log p.value for original time series HR and RR

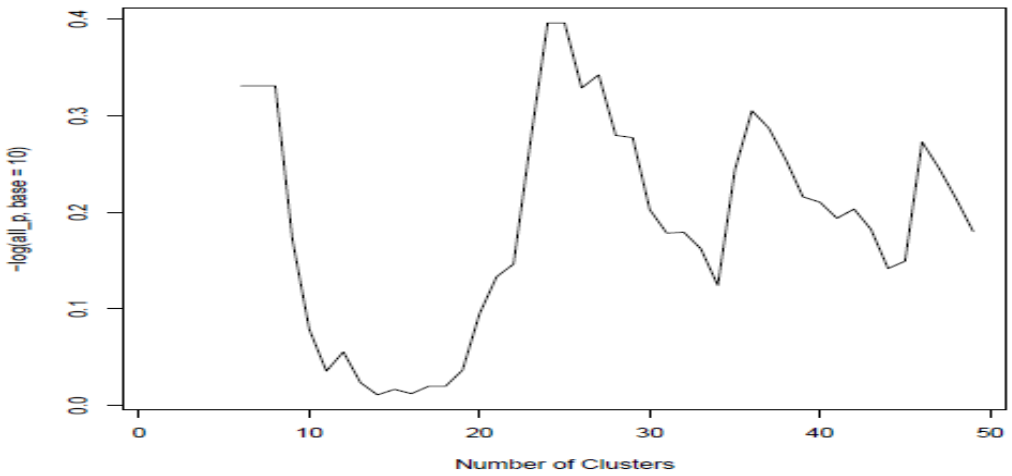


Figure 3.15: Negative log p.value for permuted time series (HR and RR)

### 3.3.10 Distribution of SIRS score in bivariate PDC clusters

The distribution of SIRS score in all clusters is shown in Figure 3.16 and and proportional to the number of object in one cluster is shown in figure 3.17. The clusters are of mixed SIRS score.

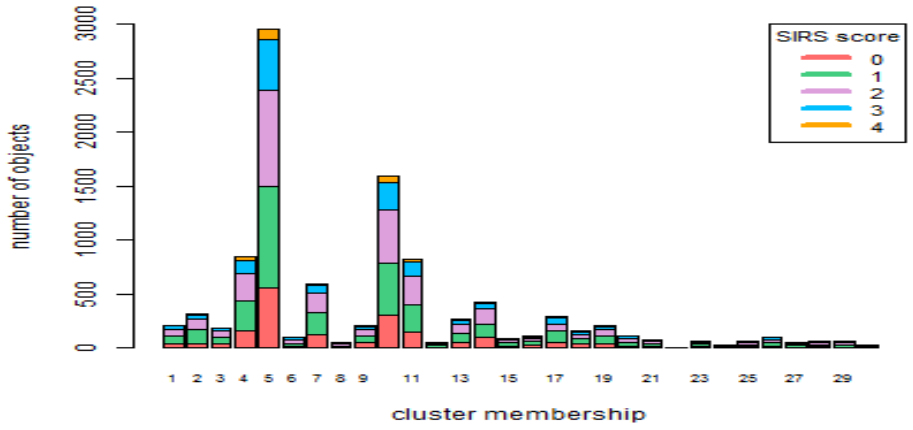


Figure 3.16: Distribution of SIRS score in clusters

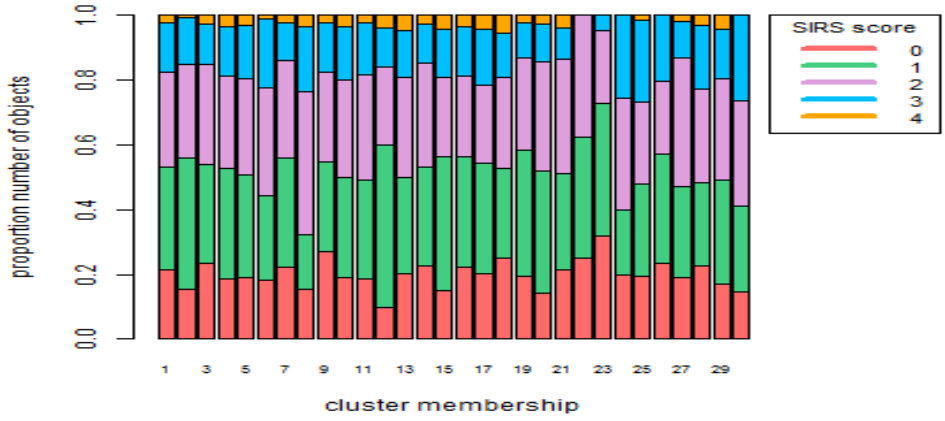


Figure 3.17: Distribution of SIRS score in clusters

# Chapter 4

# Implementing a Doctor-friendly Analytic Dashboard for Pediatric ICU

## 4.1 Introduction

The large data in health (big data) needs a platform where this data can be compile and visualize to be able to use. The ICU dashboard provides continuous monitoring to the status of patients condition. Studies in past shows continuous monitoring of patient's health record reduce mortality in the ICU, It helps in the decision making but increases antibiotic intake. [19] Philips healthcare launched graphical interface dashboard for its eICU program in 2013 to provide quality care and to monitor health status of a individual patients. It process the complex clinical data and visualize analytics in real time to its customers (the hospital staff i.e. clinicians, nurses etc). It helps clinician to take decision about a patient whether patient needs instant care or patient is ready for discharge. The study says that eICU program monitor more than 350,000 patients per year with no compromise of the individual patient care need. The purpose was to reduce mortality and stay days in ICU, improve quality of care. The eICU program results in improving access to care, individual attention even in the shortage of nurses and enable resources for a large population. It reduces mortality, stay days and readmission rate in hospital. [20] Author describes about **Decisio Health's** Clinical Platform, a dashboard for trauma and surgical intensive care. In another instance, it

was found that dashboard approaches reduce time, as on an average nurses took 15 minute to the compile the information. [21]. Moreover, it is expected that these can also help in decision making, risk prediction, performance improvement by analysing this big data.
Therefore, we have designed a doctor-friendly dashboard that carries out analytics upon the **Vital Monitoring** data from the **Pediatric Intensive Care Unit** (PICU) at the **All India Institute of Medical Sciences** (AIIMS), New Delhi and displays key patient indices calculated from the continuous data and updated in an hourly fashion.

## 4.2 Backgorund

### 4.2.1 Data Pipeline

Data warehouse is set up in the PICU, dashboard is build on top of its data pipeline. This pipeline consolidated the all information about the patient which can be visualized on the dashboard.
**Multi-parameter monitoring data at 15-second and 1-second resolution,** $Mindray^{TM}$ monitors for patients vitals monitoring with central monitoring station (CMS) are installed in PICU. The communication between the monitors and CMS was executed on a Raspberry Pi with a 1 Terabyte hard disk to store the upcoming data using python and c++ programming. According to device protocol, client-socket programming was used to query and receive the vital data at resolutions of 15 seconds (unsolicited data) and 1 second (real-time data). HL7 messaging was implemented to query all vitals signals including Heart Rate (HR), Respiratory Rate (RR), Oxygen saturations (SpO2), Pulse Rate (PR), Non-invasive Blood Pressure (NIBP), Arterial BP. System time (up to microsecond resolution) as time stamp was appended with every response HLT message. Deployment of these codes were done on an IBM Windows Server System X3650. **Treatment charts,** treatment notes are entered by resident doctors in a word document two-times in a day. Automated backup of word files was taken which were then parsed into a text file using Python docx2txt. It contains personal information of a patient, antibiotics and microbiotics details etc.

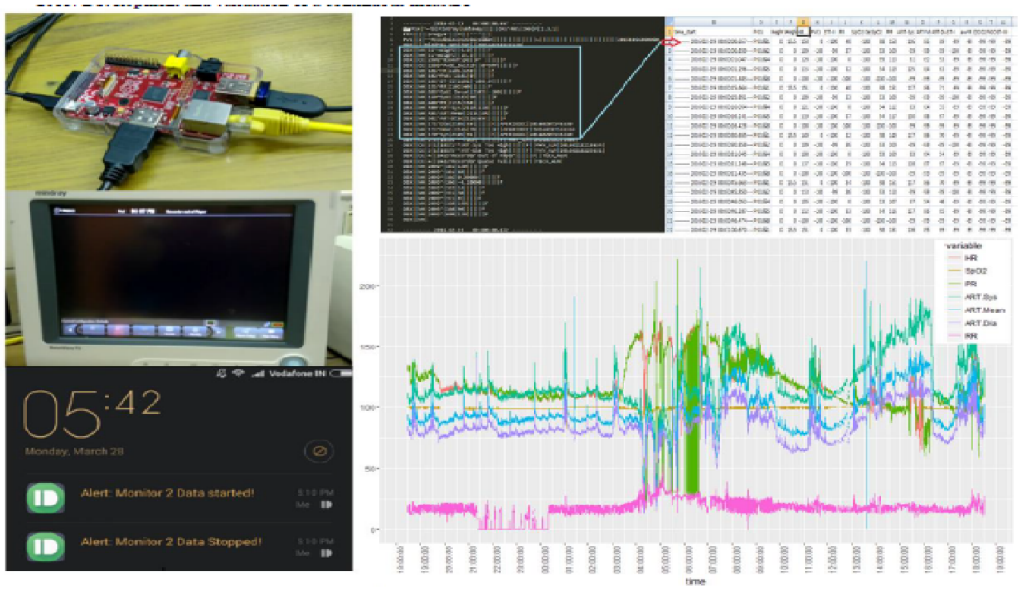


Figure 4.1: warehouse pipeline for collecting data in PICU [22]

## 4.3 Methodology

### 4.3.1 PostgreSQL database

PostgreSQL database was created on top of data pipeline in the icu for storing the monitor data and treatment chart data. Database contain three tables, one for storing the all monitor data with unique id of patient which is admission id given by hospital, second for storing which antibiotic is given to patient on a particular day, third for storing unique number to each antibiotic name to identify in second table. The **time-stamped** data, standardized using POSIXct functions is **automated** for **cleaning** and **preprocessing**, which are scheduled on a daily basis. These pipelines convert the raw data into structured tables stored on a local **postGreSQL** server into a relational database which is used for dashboard visualization. The data pipeline is shown in figure 4.1 (taken from [22])

### 4.3.2 Analytical Dashboard

The dashboard has currently deployed **Quality Check**, **Basic Analytics** and **Advanced Analytics** modalities. This dashboard is designed not only for Quality Improvement through display of basic Quality Check mechanisms such as patient information checks, technical anomalies and fidelity of sensor data, but also is designed to giving **care-related feedback** to Intensivists and Nurses. The latter are implemented as basic data-analytics such as **histograms** for hourly **Oxygen saturations** and the weekly use of **Antibiotics** displayed as **Radar plots**. The **advanced analytics** module is currently locked to be seen and assessed only by the analytic and the core clinical team and will be displayed to the resident doctors only after **validation** of their clinical potential. This is focused on deriving features for early detection of critical outcomes such as occurrence of **bursts** in physiological signals and **physiological anomalies** for learning **data-driven models** and **machine learning**. The dashboard is built using **shinydashboard** package in **R** and **PostGreSql** is used for maintaining the structured data updates. PostGreSql server is queried to receive data on hourly basis to analyze and visualize on the dashboard. Advanced Machine learning analytics and assessment of data quality in a pre-dashboard to a post-dashboard phase are currently being implemented in a clinical study at the PICU at AIIMS. The main page of dashboard visualized ICU level normalised antibiotic uses in last 7 days in the ICU, patient's oxygen saturation histogram (default range is 88 to 95, which can be change by clinician interactive input space on the dashboard) and data check (red color bar for incorrect pid, green color bar for correct pid) bar plot shown in figure 4.2.

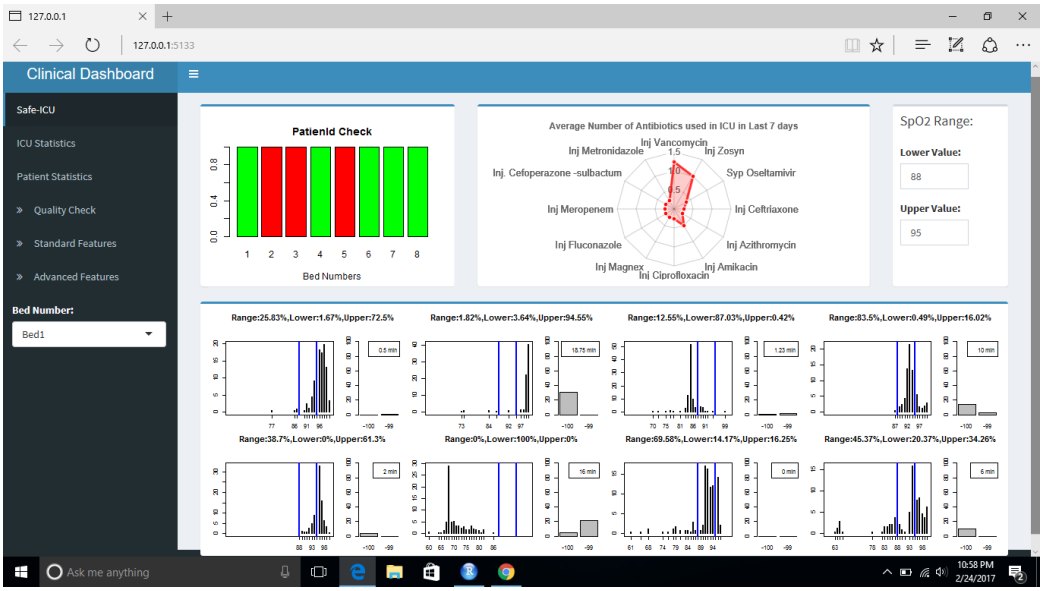


Figure 4.2: The dashboard front page, patient id check barplot, Radar plot of antibiotic used in last 7 days, oxygen saturation of each bed in last 1 hour with default rage of 88 to 95

# Chapter 5

# Conclusion and Future Directions

## 5.1 Conclusion

1. It was seen that the epochs of the same patient had similar PDC patterns, thus could be seen as a signature. These PDC patterns were non-random as assessed from the comparsison with shuffled time series. PDC patterns from shuffled time series displayed much less structure on the hierarchical dendrogram, which was expected. However, although there is a strong physiological hypothesis behind the use of PDC features in this thesis, the experiments so far have shown that these were not found to be optimal for predicting sepsis based upon SIRS score threshold in MIMIC II data. There can be multiple reasons for this. The scores we used were binary and based on expert driven thresholding derived from consensus committees. These may not be the most optimal scores for prediction of sepsis as other studies have also shown. Therefore, the PDC patterns also need to be tested against other available indices of sepsis and doctor's labels of sepsis (clinician judgment). Furthermore, since the complexity of physiological signals is known to be altered in many other conditions, these patterns are yet to be examined for other critical conditions such as hypertensive episodes, hypotensive episodes, acute kidney failure, pneumonia etc. Currently, this work is in progress but outside the scope of this thesis.

2. The dashboard deployed for the SAFE-ICU at AIIMS is a first of its

kind in Indian ICU settings. It is being used to devise further studies. Importantly, this dashboard displays important information with respect to antibiotic usage and patient patterns such as time spent in correct range of oxygen saturation. With these indices displayed on dashboard, we are devising a pre-dashboard to a post-dashboard comparison for evaluation of antibiotic use and critical outcomes. Since this is a quality study, it has been accepted for launch in PICU at AIIMS and the dashboard will form a major component of the pre-post dashboard phase comparisons. These indices derived automatically from data are not available at any tertiary care hospitals in India.

## 5.2 Future Directions

We would like to explore:

1. Usefulness of PDC clusters in other critical conditions such as hypertensive episodes, hypotensive episodes, acute kidney failure, pneumonia etc.

2. Assessment of PDC clusters for other standard scores such as SOFA and QSOFA.

3. Assessment of different norms other than the Euclidean norm upon the multivariate hellinger distances as some time series may be more important than the others.

4. Age-wise analysis of PDC patterns.

5. Optimization of PDC length patterns for supervised modeling and comparison with other features.

6. Supervised modeling using a battery of features including PDC. Currently, preliminary Random Forest models have been constructed, however these have high false negative rates. Optimization of features and models is work is in progress and outside the scope of this thesis.

7. Sepsis threshold may be changed from binary to ordinal and regression models are being built.

# Chapter 6

# References


1. https://www.cdc.gov/nchs/icd/icd10cm.htm

2. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Levin A, Acute Kidney Injury N (2007) Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 11:R31.

3. The Acute Respiratory Distress Syndrome Network (2000) Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 342:1301–1308.

4. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, Sevransky JE, Sprung CL, Douglas IS, Jaeschke R, Osborn TM, Nunnally ME, Townsend SR, Reinhart K, Kleinpell RM, Angus DC, Deutschman CS, Machado FR, Rubenfeld GD, Webb SA, Beale RJ, Vincent JL, Moreno R, Surviving Sepsis Campaign Guidelines Committee including the Pediatric S (2013) Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med 41:580–637.

5. Celi LA, Zimolzak AJ, Stone DJ (2014) Dynamic clinical data mining: search engine-based decision support. JMIR Med Informatics 2:e13.

6. Hicks J.F., Fairchild K. HeRO monitoring in the NICU: sepsis detection and beyond. Infant 2013; 9(6): 187-91.

7. Moorman J.R., Carlo W.A., Kattwinkel J. et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. J Pediatr 2011;159:900-06. E1.

8. Gang Y., Malik M. Heart rate variability in critical care medicine. Curr Opin Crit Care 2002;8:371-75.

9. Buchman TG, Stein PK, Goldstein B. Heart rate variability in critical illness and critical Care.

10. Griffin M.P., Moorman J.R. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. Pediatrics 2001;107: 97-104.

11. Griffin M.P., O'Shea T.M., Bissonette E.A. et al. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. Pediatr Res 2003;53:920-26.

12. Kovatchev B.P., Farhy L.S., Cao H. et al. Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome. Pediatr Res 2003;54:892-98.

13. Lake D.E., Richman J.S., Griffin M.P., Moorman J.R. Sample entropy analysis of neonatal heart rate variability. Am J Physiol Regul Integr Comp Physiol 2002;283:R789-97.

14. Chang K.L., Monahan K.J., Griffin M.P. et al. Comparison and clinical application of frequency domain methods in analysis of neonatal heart rate time series. Ann Biomed Eng 2001;29:764-74.

15. Fairchild K.D., Schelonka R.L., Kaufman D.A. et al. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. Pediatr Res 2013;doi:10.1038/pr.2013.136.

16. http://scidok.sulb.uni-saarland.de/volltexte/2012/4545/pdf/Andreas_M_Brandmaier_Dissertation.pdf

17. http://www.cs.ucr.edu/ eamonn/SIGKDD 2004.

18. https://about.citiprogram.org/en/homepage/

19. https://healthmanagement.org/c/icu/news/philips-introduces-graphical-dashboard-for-icu-management

20. https://www.healthdatamanagement.com/news/visual-dashboard-brings-together-key-clinical-data-in-icu?tag=00000151-16d0-def7-a1db-97f037710000

21. A. F. Simpao, L. M. Ahumada, M. A. Rehman, Big data and visual analytics in anaesthesia and health care.

22. Tavpritesh Sethi, Aditya Nagori, Ambika Bhatnagar, Priyanka Gupta, Richard Fletcher, Rakesh Lodha, Validating the Tele-diagnostic Potential of Affordable Thermography in a Big-data Data-enabled ICU

23. Brandmaier A (2012). Permutation Distribution Clustering and Structural Equation Model Trees. Dissertation, Saarland University, Saarbrücken.

24. Brandmaier A (2015). pdc: Permutation Distribution Clustering. R package version 1.0.3,URL http://CRAN.R-project.org/package=pdc.

25. M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L. Lehman, G.B. Moody, T. Heldt, T.H. Kyaw, B.E. Moody, R.G. Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access ICU database. Critical Care Medicine 39(5):952-960 (2011 May); doi: 10.1097/
CCM.0b013e31820a92c6.