5-2018

The Use of EHR data in Early Detection Systems: A Case in Sepsis and In-Hospital Mortality Prediction

Varisara Tansakul
University of Tennessee, vtansaku@vols.utk.edu

Recommended Citation

Tansakul, Varisara, "The Use of EHR data in Early Detection Systems: A Case in Sepsis and In-Hospital Mortality Prediction. " Master's Thesis, University of Tennessee, 2018.
https://trace.tennessee.edu/utk_gradthes/5022

To the Graduate Council:

I am submitting herewith a thesis written by Varisara Tansakul entitled "The Use of EHR data in Early Detection Systems: A Case in Sepsis and In-Hospital Mortality Prediction." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Industrial Engineering.

Anahita Khojandi, Xueping Li, Major Professor

We have read this thesis and recommend its acceptance:

John E. Kobza

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

The Use of EHR data in Early Detection Systems: A Case in Sepsis and In-Hospital Mortality Prediction

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Varisara Tansakul

May 2018

ABSTRACT

In this thesis, we aim to use electronic health records (EHRs) to predict sepsis and in-hospital mortality by using machine learning algorithms. We first explored EHRs dataset and performed data cleansing. Then, we extracted 57 features using data of vital signs and white blood cell (WBC) count. Two classification algorithms (i.e., random forest and neural network) were used to develop predictive models using the data from the first few hours after admission to predict sepsis and in-hospital mortality. In addition, we also used the data collected in the last few hours before sepsis developed to predict sepsis.

The results show promise in early prediction of sepsis and possibly providing an opportunity for directing early intervention efforts to prevent or treat sepsis.

TABLE OF CONTENTS

Chapter One Background and data 1

1.1 Introduction 1

1.2 Objectives 2

1.3 Literature Review 3

1.4 Dataset 5

1.5 Data Cleansing 6

Chapter Two Predictive models 10

2.1 Response Variables and Features 10

2.2 Experiments 13

2.3 Classification Algorithms 14

2.4 Metrics 18

2.5 Results 18

Chapter Three Conclusion and discussion 25

3.1 Discussions, Limitations and Future Work 25

3.2 Conclusions 31

References 32

Vita 38

LIST OF TABLES

Table 1: Number of missing values 7

Table 2: Units of heart rate and respiratory rate.. 8

Table 3: Summary of demographics 9

Table 4: Features used in predictive models.. 11

Table 5: Results of Experiment I with feature set (a) 21

Table 6: Results of Experiment I with feature set (b) 21

Table 7: Results of Experiment II with feature set (a) 23

Table 8: Results of Experiment II with feature set (b) 23

Table 9: Results of Experiment III with feature set (b) 23

Table 10: Top ten most important contribution factors to sepsis/mortality prediction for a subset of random forest models. 27

LIST OF FIGURES

Figure 1: SIRS and qSOFA criteria 2

Figure 2: Visualization of Experiment I using 12 hours after admission 14

Figure 3: Visualization of downsampling technique 16

Figure 4: Visualization of neural network 17

Figure 5: The breakdown of the dataset with respect to meeting SIRS criteria along with the discharged description 20

Figure 6: Importance of top ten most important features in Experiment I with feature subset (b) using the first 12 hours after admission 28

Figure 7: Importance of top ten most important features in Experiment II with feature subset (b) using the first 12 hours after admission 28

Figure 8: Importance of top ten most important features in Experiment III with feature subset (b) using the first 12 hours after admission 29

CHAPTER ONE
BACKGROUND AND DATA

1.1 Introduction

Sepsis is the systemic inflammatory response to severe infection, typically pneumonia, gastrointestinal or urinary tract infection [1], and can cause serious consequences for patients. The mortality rate following sepsis can reach up to 30%, with 50% and 80% for severe sepsis and septic shock, respectively [1]. Once a patient develops sepsis, the mortality rate goes up when left untreated. Therefore, detection of high-risk patients is necessary in order to decrease mortality through early intervention and optimal care.

Because sepsis is a system inflammatory response to infection, it is generally associated with elevated heart rate, temperature, and respiratory rate, as well as either low or high white blood cell (WBC) count. Accordingly, healthcare providers currently rely on patients' physiological symptoms to identify sepsis cases [2]. For instance, Systemic Inflammatory Response Syndrome (SIRS) criteria, which was introduced in 1992, categorizes a patient as septic from having two or more of the symptoms presented in Figure 1 [2]. In 2016, Sepsis-3 was introduced to replace the SIRS criteria with a new risk-stratification tool. In Sepsis-3, sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection [3]. Quick Sequential Organ Failure Assessment (qSOFA) was also introduced within Sepsis-3 to be used with patients who have suspected infection and are likely to have prolonged stay in Intensive Care Units

SIRS Criteria

Meet two or more of the followings:

- Temperature: $< 36^{\circ}\mathrm{C}$ or $> 38^{\circ}\mathrm{C}$
- Heart rate: > 90 beats per minute
- Respiratory rate: > 20 breaths per minute
- White blood cell count: $< 4{,}000$ cells per mm^3, $> 12{,}000$ cells per mm^3, or $> 10\%$ immature (band) forms

qSOFA Criteria

Meet two or more of the followings:

Respiratory rate: ≥ 22 breaths per minute
Altered mental status
Systolic blood pressure: ≤ 100 mm Hg

Figure 1: SIRS and qSOFA criteria

or to expire in the hospital [3]. The validation of Sepsis-3 and also qSOFA are subjects of ongoing research [4]. Identifying septic patients using these recent definitions and assessment tools is somewhat complex, which coupled with the lack of requisite data, may not be practical in our dataset [4]. Hence, in this study we opt to use the well-established SIRS criteria.

1.2 Objectives

The goal of this study was to retrospectively analyze historical electronic health records (EHRs) data to develop models that can predict sepsis and in-hospital mortality. Specifically, we used powerful machine learning techniques on physiological information collected shortly after admission to predict future incidence of sepsis and in-hospital mortality. In addition, we used these techniques on the physiological information collected shortly leading to incidence of sepsis to draw insights about the changes in patient symptoms. In general, these models can help healthcare practitioners in early detection of sepsis and provide patients with timely, personalized treatments before a sharp increase in the risk of developing sepsis or in-hospital mortality.

The time of sepsis is generally not recorded in EHRs. Hence, in this study, we categorized patients as septic as soon as they meet the well-accepted SIRS criteria. In addition, in this study we limited our attention to adult patients diagnosed with pneumonia, a group that is highly susceptible to sepsis.

1.3 Literature Review

There exists an extensive body of work on the use of data-driven models to predict sepsis or mortality. Most studies developed predictive models using machine learning algorithms with data collected from Intensive Care Unit (ICU) or emergency rooms (ERs). Awad et al. [5] used the MIMIC II [6] data of patients age 16 or older within a single ICU to predict in-hospital mortality using random forest, the predictive Decision Trees, the probabilistic Naive Bayes, and the rule-based Projective Adaptive Resonance Theory models. They conducted five experiments with different datasets (e.g., original dataset, modified datasets using the Synthetic Minority Oversampling Technique (SMOTE), replaced missing values by applying an algorithm). Random forest mostly outperformed other machine learning algorithms.

Jaimes et al. [7] used ERs data of patients age 15 or older with suspected or confirmed bacterial infection as admission diagnosis and having at least one of the symptoms in SIRS criteria. Data were collected from two hospitals located in Columbia. The goal of this study is to compare predictions of mortality within the first 28 days after admission to the ER using logistic regression and neural networks. Neural network outperformed logistic regression by having higher areas under the receiver operator characteristic (ROC) curves.

Taylor et al. [8] used emergency department (ED) visits data of patients age 18 or older and developed sepsis as meeting SIRS criteria with infectious admitting diagnosis to predict in-hospital mortality by using random forest, classification and regression tree (CART), logistic regression, and previously developed clinical decision rules (CDRs). Their results show that random forest outperformed other models and had the highest area-under-the-curve (AUC) under ROC. Gultepe et al. [9] used EHRs of adult patients who met a minimum of two on SIRS criteria and were admitted through ED using support vector machine (SVM) and Bayesian network (BN) to predict lactate level and mortality. These models were trained for sepsis patients, and all patients regardless of sepsis status, and achieved accuracies of up to 72.8% and 71.5% in predicting mortality, respectively.

Three studies below used data from EHR to develop models for early detection of sepsis. Giuliano et al. [10] used Project IMPACT dataset of adults with an admitting ICU diagnosis of sepsis to assess the predictive value of early detection of sepsis using physiological data recommended by the Surviving Sepsis

Campaign (SSC). They obtained an accuracy of approximately 62% in predicting sepsis using the logistic regression algorithm.

Giannini et al. [11] developed a real-time machine-learning algorithm by training random forest on EHR data to predict patients with risk of having severe sepsis and/or septic shock. They deployed the system in “silent mode” for two months and the results show that they achieved positive and negative predictive values of 29% and 97%, respectively. Another study [12] used data from the ICU to detect sepsis in real-time using decision trees (DT), SVM, and Naïve Bayes (NB) algorithms. All developed models successfully detected all patients experiencing severe sepsis and septic shock, except for the NB algorithm that misclassified only one septic shock patient as a severe sepsis patient, resulting in an accuracy of 99.82%.

1.4 Dataset

We used the data pulled from the Health Facts® (HF) dataset [13]. The de-identified dataset was provided by the Center for Health Systems Innovations (CHSI) at Oklahoma State University. The dataset contains EHRs from approximately 490 hospitals under Cerner Corporation, collected over approximately 14 years. The dataset includes the details of patients’ demographics (e.g., gender, age, marital status, race), patients’ information (e.g., admitted information, discharged information), clinical events (e.g., vital signs), lab procedure results (e.g., WBC count), medications administered (e.g., name of medication, order strength of medication), and diagnosis information (e.g., International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) code [14]). Data were extracted from HF dataset based on diagnosis code that started with 486 or 995.9 under ICD-9 code.

We used MySQL [15] to create a database that would allow us to store and extract high volume EHRs dataset for future use. We imported EHRs dataset to MySQL using Python language [16] along with mysql.connector library [17], and SQL. Data import process was done by importing one table at a time.

1.5 Data Cleansing

SQL was used to extract appropriate data for analyses. We extracted patient encounters' information (e.g., demographics, admission and discharged information), and diagnosis information. Vital signs (e.g., heart rate, respiratory rate, temperature), and White Blood Cell (WBC) count were extracted on a year-by-year basis due to the volume of data. We only focused on data that was collected from years 2008 to 2015.

As in most clinical datasets, our EHR dataset contains null values and duplicated observations, especially under patient encounters, clinical events, and lab procedure results tables. During data exploration, we found that the number of missing data of vital signs and WBC count were relatively low, compared to the number of missing data in temperature as shown in Table 1. Note that the numbers that are shown in this table are the total number of data points from 2008 to 2015 before data cleansing or limiting any patient encounters for analyses.

Table 1: Number of missing values

	Heart rate	Respiratory rate	Temperature	WBC count
Number of missing value	598,778	649,257	17,257,984	12,336
Total number of data points	40,839,761	48,099,417	39,818,969	4,260,964
Percentage (%)	1.47	1.35	43.34	0.29

For this analysis, null observations were removed and among duplicated observations with the same date and time, the one with the larger result value was kept. Because the data were collected from multiple institutions over a long time-span, there were major inconsistences across the units in which the data were reported, especially for vital signs and WBC count. Table 2 shows some units that were appeared under heart rate and respiratory rate in EHRs dataset. Hence, we converted the data when necessary. For instance, units in temperature (i.e. degree Fahrenheit) were converted to degree Celsius.

In addition, not all data was clinically meaningful after unit conversion. Hence, we removed the entries that fell outside of the following ranges: A respiratory rate between 4 and 60 breaths per minute, temperature between 32.2 and 41.1 degree Celsius, heart rate between 30 and 200 beats per minute [18], and WBC count between 500 and 50,000 cells/μL [19].

Table 2: Units of heart rate and respiratory rate

Type	Unit
Heart Rate	Beats per minute
	Milliseconds
	Not Mapped
	Pacing Rate (Pacing Beats per Minute)
	Second
	NA
Respiratory Rate	Beats per Minute
	Breaths per Minute
	Millimeters Mercury
	Minute
	Not Mapped
	per Minute
	NA

Specifically, here we only focused on data from 2008 to 2015 on adult patients (18 years or older) who were admitted due to either physician or clinical referral, and were diagnosed with pneumonia, captured by the ICD-9 code [14]. Table 3 summarizes the demographics of patients who were diagnosed with pneumonia and admitted with referrals.

Table 3: Summary of demographics

Total patient encounters (*n*)	332,006
Gender	
Female	52.73 %
Male	47.25 %
Race	
Asian or Pacific Islander	1.22 %
African American	14.66 %
White	79.36 %
Other/ Unknown	4.76 %
Marital Status	
Married	43.83 %
Widowed	18.84 %
Single	22.58 %
Divorced	11.88 %
Unknown	2.87 %
Payer Code	
Private/HMO	20.49 %
Medicaid	7.88 %
Medicare	46.32 %
Self-pay/uninsured	5.10 %
Other	20.21 %
Age (years)	63.56±18.41
Length of Stay (hours)	97.43±739.62

CHAPTER TWO
PREDICTIVE MODELS

2.1 Response Variables and Features

We used two response variables in this study, namely, in-hospital mortality and sepsis. For in-hospital mortality, we used the discharge description which was recorded under the patient encounters table. We eliminated observations corresponding to “not mapped” and “unknown”, as well as null values.

The EHRs did not contain the time of sepsis if it was developed. Hence, we used the well-established SIRS criteria [2] to estimate the time of sepsis if it occurred. This process was done by developing a for loop in Python [16]. We first converted result values of vital signs and WBC count to either zero or one. If a result value qualified as one of the symptoms under SIRS criteria shown in Figure 1, then we assigned the result value as one. Otherwise, we assigned the result value as zero. After the assignment of result values, we combined rows of data with new assigned result values of vital signs and WBC count. Then, we ordered rows of data by patient encounters’ ID, and event date and time. The for loop was developed to sum result values for each patient’s encounter ID. Patient encounters with summation of two on their result values would be marked as they developed sepsis. Specifically, we retrospectively examined each patient encounter to determine whether they acquired sepsis and if so, collect its initiation time.

We used a total of 57 features, including both categorical and continuous variables as shown in Table 4. Categorical variables include demographics

Table 4: Features used in predictive models

Demographics	Age groups: 18-44 years old, 45-64 years old, ≥65 years old
	Gender: Male, Female
	Race: Asian or Pacific Islander, African American, White, Other
	Marital status: Married, Widowed, Single, Divorced, Unknown
	Payer code: Private/HMO, Medicaid, Medicare, Self-pay/uninsured, Other
Features below were applied to heart rate, respiratory rate, temperature, and WBC count	
Basic statistics	Minimum, maximum, mean, standard deviation
Signal information	Shannon Entropy
Differences in consecutive values	Minimum, maximum, mean, standard deviation
Proportional differences	Minimum, maximum, mean, standard deviation

information, such as gender, race, payer code, and age groups [20]. Continuous variables included: information on vital signs, namely, heart rate, respiratory rate, and temperature, as well as WBC count. We calculated the basic statistics, such as minimum, maximum, and standard deviation, as well as information entropy, particularly Shannon entropy [21], for all of the four continuous variables. All continuous variables were calculated for each patient encounter by using Pandas library [22] in Python [16].

In addition, when possible, we generated features based on changes in consecutive clinical events. That is, we calculated the *differences in consecutive values,* as well as the *proportional differences*. Specifically, the differences in conservative values were calculated by finding the difference between consecutive observations for each vital sign and WBC count. The proportional differences were calculated when dividing the differences in conservative values from vital signs and WBC count by the difference of time between these consecutive observations. We then calculated the basic statistics of these features.

Note that to generate the differences in consecutive values and the proportional differences features, we need at least two values. Hence, due to the low frequency of data collection for some features, such as WBC count, differences in consecutive values and the proportional differences features may not be calculated. Therefore, we performed the analysis in two ways: (1) Kept the differences in consecutive values and the proportional differences features and removed patients from the dataset with fewer than two entries for vital signs or

WBC count, and (2) removed the differences in consecutive values and the proportional differences features and kept the patients for whom the parameter may not be calculated.

The datasets that were used for analyses contained patient encounter ID, features for each patient encounter, and a response variable depending on the goal of analysis. For instance, the prediction of sepsis would contain a response variable that indicated whether the patient encounter acquired sepsis.

2.2 Experiments

We performed three main experiments as follows:

Experiment I: Use the EHRs data from the first 12, 24, and 48 hours after admission to predict which patients would develop sepsis;

Experiment II: Use the EHRs data from the first 12, 24, and 48 hours after admission to predict which patients would expire;

Experiment III: Use the EHRs data from the 12, 24, and 48 hour-windows leading to sepsis to predict which patients would develop sepsis.

We used two feature subsets as follows:

(a) All features

(b) All features except differences in consecutive values and proportional differences

For each experiment, the dataset was refined to only include patients who had length of stay (LOS) longer than the number of hours used in the corresponding analysis. For instance, in Experiment I, when predicting which patients would acquire sepsis using the EHR data from the first 12 hours after admission, we excluded patients who acquired sepsis within the first 12 hours as shown in Figure 2. Note that LOS was calculated from the difference of admission time and discharge time.

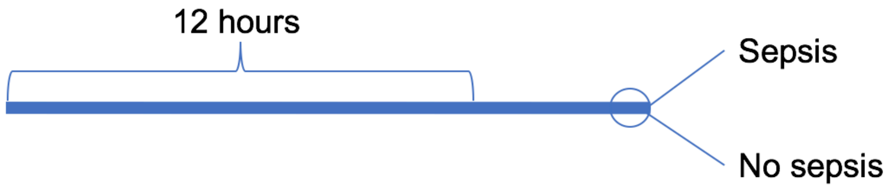

Figure 2: Visualization of Experiment I using 12 hours after admission

2.3 Classification Algorithms

In this study, we used two classification algorithms, namely random forest [23] and neural network [24]. Random forest is an ensemble learning method and can be used in classification and regression problems. Random forest relies on the aggregate results from a series of decision trees. We particularly used random forest in this study as the algorithm is very robust against overfitting due to randomly selecting subset of features at each split as it grows decision trees [23]. In addition, we replicated the analysis using the artificial neural network algorithm. Neural network has been widely applied in healthcare applications as it can deduce the non-linear relationship between independent and dependent variables, as well as the interactions between features [25]. We used Python 2.7 [16] for implementation. Specifically, we used Scikit-learn library [26] to develop random forest models and used Tensorflow library, developed by Google [27], to construct fully connected neural networks. We partitioned the dataset into 70%, 15%, and 15% for training, validation, and test sets. Based on the results of our preliminary experiments, we opted out of tuning hyper-parameters for random forest models, hence we combined training and validation datasets while training random forest models.

Our training datasets were, in general, highly unbalanced with respect to the response variables, e.g., there were approximately nine times more instances of expired patients than non-expired patients in the cleansed dataset under Experiment I with feature subset (a). To ensure that the developed models do not favor the more represented observations in the dataset, we used the downsampling technique to generate a series of balanced sub training datasets from the initial training dataset and exploited warm-starting to achieve higher accuracy. The visualization of the downsampling technique is shown in Figure 3 by assigning the gray color block to be the over-represented class and orange block to be the under-represented class. The right side of this figure is an example of a balanced dataset that was generated using the downsampling technique, which contained all data of under-represented class and the same amount of data from over-represented class.

Specifically, for random forest, we developed a 700-tree forest by building one tree at a time on a new sub-training set, while applying warm-starting

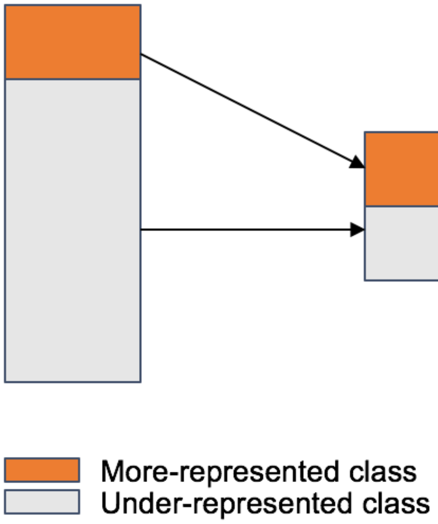

Figure 3: Visualization of downsampling technique

technique, and aggregating them into one model. Warm-starting allows us to reuse the solution to the previous call to fit function [28].

For neural network, we used *tf.contrib.learn* under Tensorflow library [27], which allowed us to create models while applying warm-starting. Specifically, we trained a model on a sub training set while applying a warm-starting method and moved on to the next sub training set when the accuracy of the validation set started to decrease. The procedure terminated when the accuracy of the validation set did not increase when transitioning to the next sub training set. We used two optimizers in neural network models. We used stochastic gradient descent (SGD) optimizer for Experiments I and II, and the Adaptive Moment Estimation (Adam) optimizer for Experiment III. Adam optimizer can deal with sparse gradients and non-stationary objectives [29], because it combines the advantages of AdaGrad

[30] and RMSProp [31] optimizers. Initially, we used SGD optimizer for Experiment III, however the validation accuracy reflected that the models did not performed well. Therefore, we changed optimizer from SGD to Adam. Note that each neural network model required parameter tuning (e.g., number of hidden layers, number of hidden nodes, learning rate) to optimize the performance for a model. The screenshot of partial architecture of neural network with three hidden layers that was constructed using Tensorflow library [27] is shown in Figure 4. The screenshot was captured from TensorBoard, which is a suite of web applications for inspecting and understanding Tensorflow runs and graphs [32]. Finally, the best trained models were applied on the corresponding, separate test sets to objectively evaluate the performance of the models.

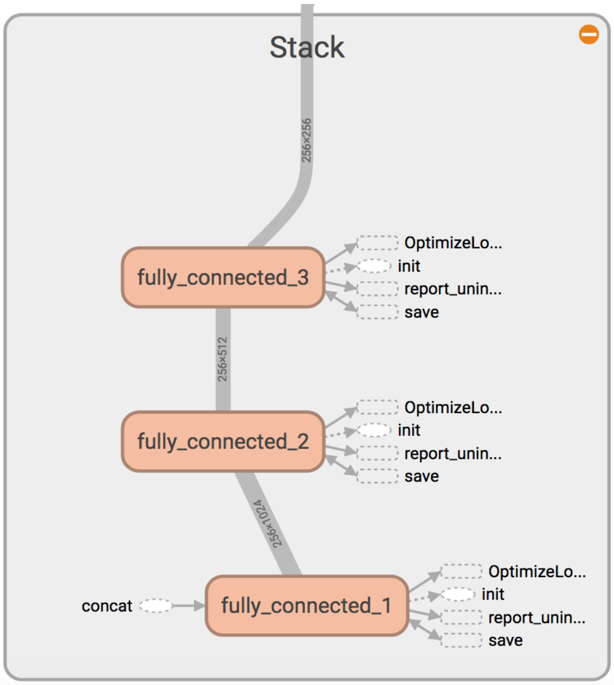

Figure 4: Visualization of neural network

2.4 Metrics

For all experiments, we report the accuracy and F1 score for the separate test sets to evaluate and compare the models. Confusion matrix can be produced based on prediction results using *ConfusionMatrix* function under pandas_ml library [33]. Accuracy gives the proportion of predicted values that match the true response value, and can be calculated by using *accuracy_score* function under Scikit-learn library [26]. F1 score is a weighted average of precision and recall, which can effectively evaluate the applicability of models in practice, especially when the dataset is unbalanced. *Classification_report* function under Scikit-learn library [26] was used to calculate F1 score. Formula for F1 score is shown below.

$$F-1\,score=\frac{2*(precision*recall)}{(precision+recall)}$$

where,

$$Precision=\frac{True\ positive}{True\ positive+False\ positive}$$

$$Recall=\frac{True\ positive}{True\ positive+False\ negative}$$

2.5 Results

In our dataset, the average LOS of patients who developed sepsis was 179.14 hours, compared to 53.96 hours for the average LOS of patients who did not develop sepsis. Figure 5 presents the breakdown of the dataset with respect to meeting SIRS criteria along with the discharged description. Consistent with our experiments, in the figure we stratify patients based on their LOS, i.e., LOS more than 12, 24 and 48 hours, as well as meeting SIRS criteria. We report the raw numbers and percentage of patients in each subcategory. For instance, out of the total of 332,006 patients remained in the dataset after cleansing, 261,258 (or approximately 79%) have a LOS that is greater than 12 hours, out of which 106,938 (41%) acquired sepsis at some point. Approximately 27% of patients with a LOS greater than 12 hours, acquired sepsis after 12 hours, i.e., 73% of patients acquired sepsis within the first 12 hours after admission. This highlights the importance of predicting/detecting sepsis immediately, or within only a few hours, after admission. However, this task is very difficult with current EHR systems that mostly require manual data entry. Hence, it is important to complement current EHR systems with automated data acquisition systems that can collect and store high frequency data without direct clinician intervention to be able to leverage the data in early detection/prevention of sepsis.

Table 5 and Table 6 summarize the results of Experiment **I**, i.e., predicting sepsis, using feature sets (a) and (b), respectively. As seen in Table 5, the best accuracy and F1 scores across the two models range from 61%-64% and 67%-75%, respectively. As seen in the table, the prediction accuracy and F1 score do not seem to be very sensitive with respect to the data collection window. This is

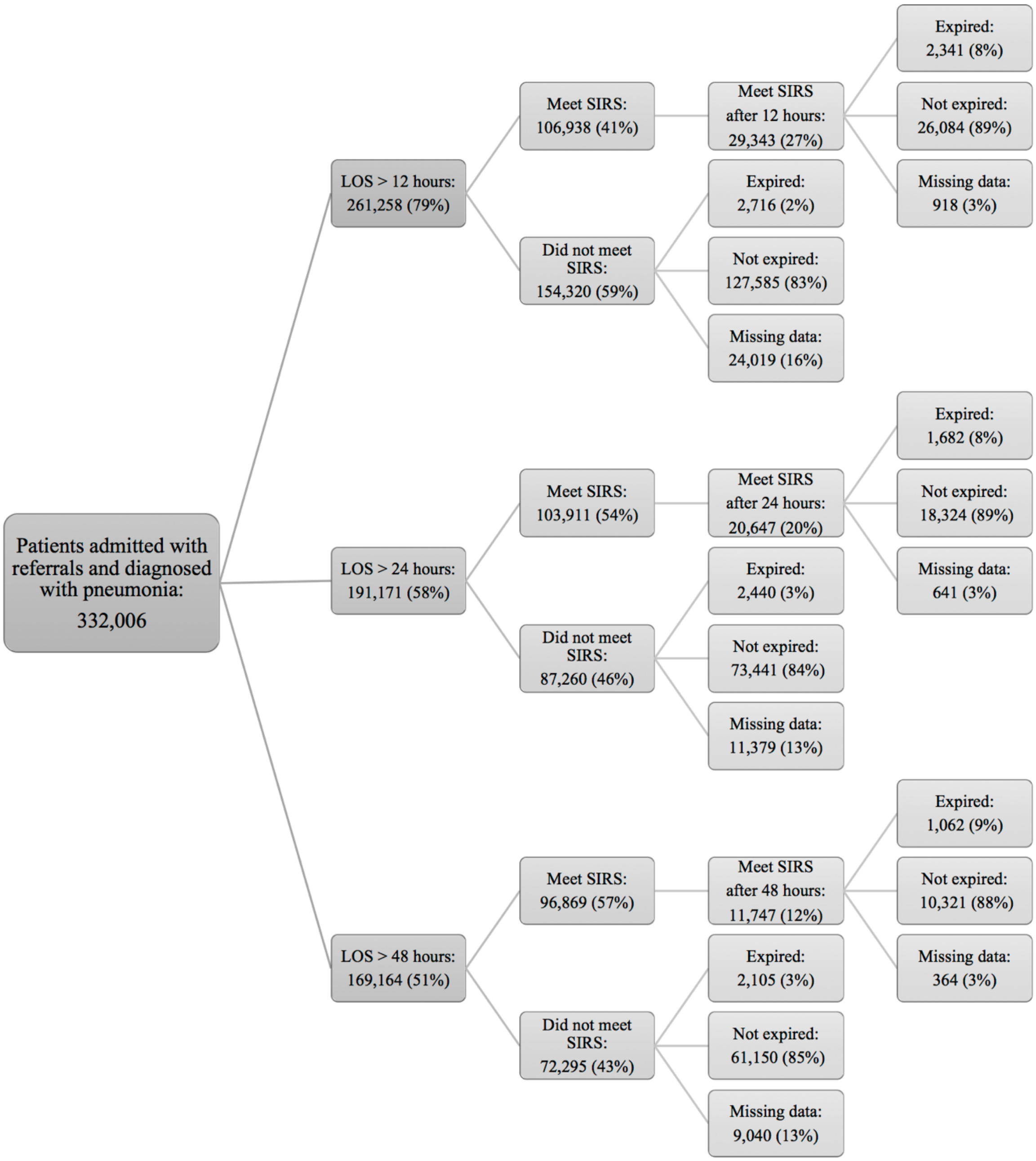
Patients admitted with referrals and diagnosed with pneumonia: 332,006
LOS > 12 hours: 261,258 (79%)
Meet SIRS: 106,938 (41%)
Meet SIRS after 12 hours: 29,343 (27%)
Expired: 2,341 (8%)
Not expired: 26,084 (89%)
Missing data: 918 (3%)
Did not meet SIRS: 154,320 (59%)
Expired: 2,716 (2%)
Not expired: 127,585 (83%)
Missing data: 24,019 (16%)
LOS > 24 hours: 191,171 (58%)
Meet SIRS: 103,911 (54%)
Meet SIRS after 24 hours: 20,647 (20%)
Expired: 1,682 (8%)
Not expired: 18,324 (89%)
Missing data: 641 (3%)
Did not meet SIRS: 87,260 (46%)
Expired: 2,440 (3%)
Not expired: 73,441 (84%)
Missing data: 11,379 (13%)
LOS > 48 hours: 169,164 (51%)
Meet SIRS: 96,869 (57%)
Meet SIRS after 48 hours: 11,747 (12%)
Expired: 1,062 (9%)
Not expired: 10,321 (88%)
Missing data: 364 (3%)
Did not meet SIRS: 72,295 (43%)
Expired: 2,105 (3%)
Not expired: 61,150 (85%)
Missing data: 9,040 (13%)

Figure 5: The breakdown of the dataset with respect to meeting SIRS criteria along with the discharged description

Table 5: Results of Experiment I with feature set (a)

Data Collection Window	Neural Networks		Random Forest	
	Accuracy	F1 score	Accuracy	F1 score
First 12 hours after admission	58%	64%	61%	60%
First 24 hours after admission	57%	70%	64%	67%
First 48 hours after admission	62%	75%	60%	66%

Table 6: Results of Experiment I with feature set (b)

Data Collection Window	Neural Networks		Random Forest	
	Accuracy	F1 score	Accuracy	F1 score
First 12 hours after admission	54%	69%	65%	66%
First 24 hours after admission	56%	70%	63%	66%
First 48 hours after admission	67%	80%	61%	67%

mainly because of the trade-off between having more information but fewer observations when training models using larger window sizes. As seen in Table 6, the best accuracy and F1 scores across the two models range from 63%-67% and 69%-80%, respectively. Hence, using feature set (b), in general, results in higher performances, which again may be attributed to the trade-off between having more information but fewer observations when training models. In general, neural network models seem to perform better in our study when less training data is available (i.e., with 48 hour windows).

Table 7 and Table 8 summarize the results of Experiment II, i.e., predicting mortality, using feature sets (a) and (b), respectively. As seen in Table 7, the best accuracy and F1 scores across the two models range from 85%-90% and 92%-94%, respectively. As seen in Table 8, the best accuracy across the two models ranges from 92%-93% and the best F1 scores equal 96%. Consistent with the observations from Experiment I, neural networks models generally outperform random forest models and the feature set (b) results in better performance compared to the feature set (a).

Lastly, Table 9 presents the results of Experiment III, i.e., predicting sepsis using the data collected in the time windows leading to sepsis. As expected, the accuracy is very high, i.e., up to 99%, in this case. Indeed the accuracy decreases as the window size increases as a longer window size introduces more uncertainty to the model. Granted, patients would most likely present symptoms in the 12-hour

Table 7: Results of Experiment II with feature set (a)

Data Collection Window	Neural Networks		Random Forest	
	Accuracy	F1 score	Accuracy	F1 score
First 12 hours after admission	85%	92%	65%	77%
First 24 hours after admission	87%	93%	68%	80%
First 48 hours after admission	90%	94%	69%	80%

Table 8: Results of Experiment II with feature set (b)

Data Collection Window	Neural Networks		Random Forest	
	Accuracy	F1 score	Accuracy	F1 score
First 12 hours after admission	92%	96%	69%	81%
First 24 hours after admission	93%	96%	69%	81%
First 48 hours after admission	93%	96%	70%	81%

Table 9: Results of Experiment III with feature set (b)

Data Collection Window	Neural Networks		Random Forest	
	Accuracy	F1 score	Accuracy	F1 score
12 hours leading to sepsis	88%	81%	99%	98%
24 hours leading to sepsis	84%	81%	97%	97%
48 hours leading to sepsis	78%	82%	92%	93%

window before meeting SIRS criteria, hence allowing clinicians to start treatment. However, when accounting for the information obtained from Experiments I and III, it is plausible to assume that the algorithms can help identify at-risk patients as early as 12 hours after admission and continue to increase in their accuracy if patients start to deteriorate or their risk of sepsis goes up in time.

CHAPTER THREE
CONCLUSION AND DISCUSSION

3.1 Discussions, Limitations and Future Work

In the future, it is foreseeable that clinicians will be able to rely on algorithms to predict sepsis/mortality using the data collected immediately after admission. Such algorithms would then enable clinicians to intervene in a timely manner to reduce patients' risks of acquiring sepsis or an untimely death. Our results suggest that such algorithms can be developed using the currently available EHRs data and would perform reasonably accurate to complement clinical care. Additionally, in cases where patients are facing a life-limiting illness or injury, predicting mortality can further empower patients and their caregivers with patient-centric pain management, emotional and spiritual support, and hospice care when appropriate.

In Experiments I and II, we developed models using two feature subsets and compared the model performances. Our results showed that the models generally became more accurate when more data were available for training. For instance, although generating features, such as differences in consecutive values and proportional differences in vital signs and WBC counts, make clinical sense when it comes to detecting sepsis/mortality, including them in the model reduced the number of observations and hence, reduced the model performance. We speculate that with adoption of automated high frequency data collection systems at bedside, which can store more data points for patients, features, such as differences in consecutive values and proportional differences, would contribute to higher accuracy models.

The developed models also allow for identifying the most important contributing factors to sepsis/mortality prediction. Table 10 presents the top ten most important features for some of the best performing random forest models across the three experiments. As seen in Table 10, the entropy of respiratory rate had the highest importance in discriminating sepsis/non-sepsis patients and expired/non-expired patients in both Experiments I and II when using the data from the first 12 hours after admission. We also obtained similar results when differences in consecutive values and proportional differences features were present when using feature set (a). Different from Experiments I and II, in Experiment III, i.e., predicting sepsis using the data collected in the time windows leading to sepsis, the maximum of the heart rates recorded was identified as the most important contributing factor.

Figure 6 to Figure 8 show the plots of importance for the top ten most important features of three experiments from Table 10. Figure 6 shows the plot of importance for the top ten features of Experiment I using data from the first 12 hours after admission. The importance of each feature was relatively close to each other after the top four features. Figure 7 shows the plot of feature importance from the top ten features of Experiment II using data from the first 12 hours after admission. The difference of importance for the top two features were somewhat large compared to differences between other consecutive features in this figure.

Table 10: Top ten most important contribution factors to sepsis/mortality prediction for a subset of random forest models.

Rank	**Experiment I with feature subset (b) using the first 12 hours after admission**	**Experiment II with feature subset (b) using the first 12 hours after admission**	**Experiment III with feature subset (b) using the 12 hours leading to sepsis**
1	Entropy of respiratory rate	Entropy of respiratory rate	Maximum of heart rate
2	Mean of heart rate	Mean of respiratory rate	Maximum of respiratory rate
3	Maximum of heart rate	Entropy of heart rate	Maximum of WBC count
4	Entropy of temperature	Mean of temperature	Mean of WBC count
5	Maximum of WBC count	Standard deviation of respiratory rate	Minimum of WBC count
6	Standard deviation of temperature	Maximum of respiratory rate	Minimum of temperature
7	Minimum of WBC count	Minimum of temperature	Mean of heart rate
8	Mean of WBC count	Entropy of temperature	Standard deviation of respiratory rate
9	Mean of temperature	Mean of heart rate	Standard deviation of temperature
10	Standard deviation of respiratory rate	Mean of WBC count	Maximum of temperature

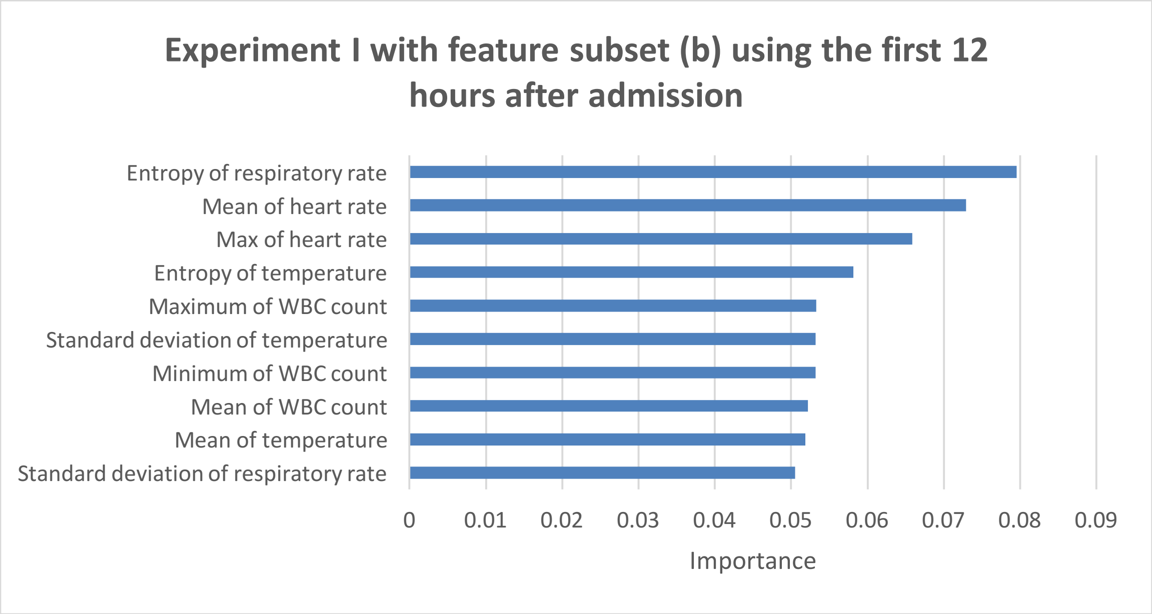

Figure 6: Importance of top ten most important features in Experiment I with feature subset (b) using the first 12 hours after admission

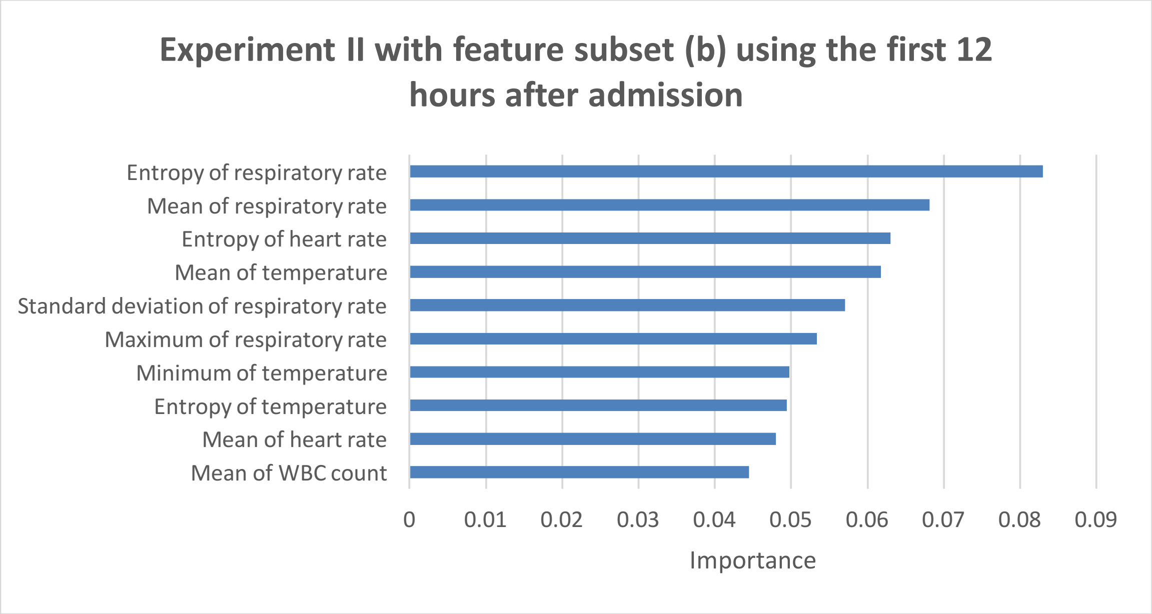

Figure 7: Importance of top ten most important features in Experiment II with feature subset (b) using the first 12 hours after admission

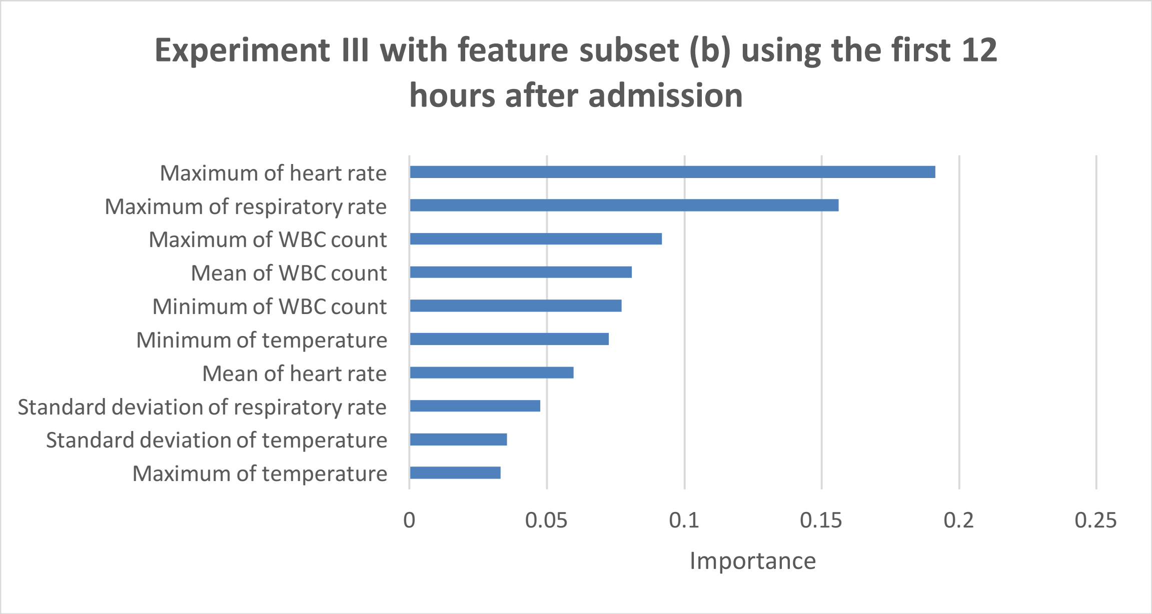

Figure 8: Importance of top ten most important features in Experiment III with feature subset (b) using the first 12 hours after admission

Figure 8 shows the plot of feature importance from the top ten features of Experiment III using data from the first 12 hours after admission. The top two features had very high importance compared to the rest of the features in the plot.

Sepsis is an important clinical event, the onset of which should be recorded in EHR systems. Under sepsis-1 and sepsis-2 definitions, patients who have infections and meet two or more symptoms under SIRS criteria [2] could be identified as septic. However, the true onset of sepsis for patients may only be identified by clinicians at bedside. Similar to most EHR systems, the system that had contributed to our dataset did not contain the diagnosis time of sepsis. Therefore, we used SIRS criteria to retrospectively approximate the time of sepsis in a given group of patients who had already been identified to have infection (i.e., patients with pneumonia). We believe recording the time of sepsis diagnosis by healthcare providers would prove very helpful in building more accurate predictive models in the future.

Note that sepsis definitions were published multiple times within 25 years, which indicated that the knowledge of sepsis is still limited. Sepsis-3 definition introduced new criteria, qSOFA and SOFA. We opted to use Sepsis-1 criteria in this study as Sepsis-3 would require keeping track of six parameters to determine whether patient encounters develop sepsis or not. Using the current dataset to mark sepsis patients with SOFA criteria would have resulted in a much smaller dataset with far fewer valid patient encounters.

We lost many patient encounters due to erroneous data or missing values. In our exploratory analysis, we encountered major inconsistences in units, many clinically non-meaningful values, missing data, as well as duplicated observations in patients' information and clinical events. It is most likely that these erroneous data or missing values were caused by data entry error, and hence, the observations were removed from the dataset. A more careful approach to form design and/or adopting automated data collection systems would reduce such errors and help with future algorithm developments.

We acknowledge that the demographics used in this study were not diverse. The summary of demographics is shown in Table 3. The predominate race of the majority of patient encounters was Caucasian. Further studies need to be performed to examine whether the risk factors or models translate well for other populations.

3.2 Conclusions

In this study, we developed models to predict sepsis and in-hospital mortality using EHR data. The developed models showed promise in early prediction of sepsis, possibly providing an opportunity for directing early intervention efforts to prevent/treat sepsis. We also examined the trade-off between the number of observations and the amount information extracted. Our results suggested that having more observations in general help increase the model performance. Lastly, based on our results, it is clear that the algorithms can help identify at-risk patients as early as 12 hours after admission. This accuracy increases dramatically when patients are at imminent risk of developing sepsis. Hence, it is plausible that continuous monitoring of patients using these algorithms can pave the way for a streamlined and improved care process.

REFERENCES

1. Jawad I, Lukšić I, Rafnsson SB. Assessing available information on the burden of sepsis: global estimates of incidence, prevalence and mortality. Journal of global health. 2012;2(1).

2. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, Schein RM, Sibbald WJ. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest. 1992;101(6):1644-55.

3. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM. The third international consensus definitions for sepsis and septic shock (sepsis-3). Jama. 2016;315(8):801-10.

4. Marik PE, Taeb AM. SIRS, qSOFA and new sepsis definition. Journal of thoracic disease. 2017;9(4):943.

5. Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. International Journal of Medical Informatics. 2017;108:185-95.

6. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. Physiobank, physiotoolkit, and physionet. Circulation. 2000;101(23):e215-e20.

7. Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. Critical care. 2005;9(2):R150.

8. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach. Academic Emergency Medicine. 2016;23(3):269-78.

9. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. Journal of the American Medical Informatics Association. 2014:315-25.

10. Giuliano KK. Physiological monitoring for critically ill patients: testing a predictive model for the early detection of sepsis. American Journal of Critical Care. 2007;16(2):122-30.

11. Giannini HM, Chivers C, Draugelis M, Hanish A, Fuchs B, Donnelly P, Lynch M, Meadows L, Parker SJ, Schweickert WD. Development And Implementation Of A Machine-Learning Algorithm For Early Identification Of Sepsis In A Multi-Hospital Academic Healthcare System. D15 CRITICAL CARE: DO WE HAVE A CRYSTAL BALL? PREDICTING CLINICAL DETERIORATION AND OUTCOME IN CRITICALLY ILL PATIENTS: Am Thoracic Soc; 2017. p. A7015-A.

12. Gonçalves JM, Portela F, Santos MF, Silva Á, Machado J, Abelha A. Predict sepsis level in intensive medicine–data mining approach. Advances in Information Systems and Technologies: Springer; 2013. p. 201-11.

13. DeShazo JP, Hoffman MA. A comparison of a multistate inpatient EHR database to the HCUP Nationwide Inpatient Sample. BMC health services research. 2015;15(1):384.

14. International Classification of Diseases, Ninth Revision (ICD-9) [Internet]. Centers for Disease Control and Prevention [updated September 1, 2009; cited 2017 October 14]. Available from: http://www.cdc.gov/nchs/icd/icd9.htm.

15. MySQL A. MySQL database server. Internet WWW page, at URL: http://www/ mysql com (last accessed/1/00). 2004.

16. Oliphant TE. Python for scientific computing. Computing in Science & Engineering. 2007;9(3).

17. MySQL A. Mysql connector/j. J http://dev/ mysql com/doc/refman/50/en/connectors html. 2004.

18. Bleyer AJ, Vidya S, Russell GB, Jones CM, Sujata L, Daeihagh P, Hire D. Longitudinal analysis of one million vital signs in patients in an academic medical center. Resuscitation. 2011;82(11):1387-92.

19. Barron HV, Harr SD, Radford MJ, Wang Y, Krumholz HM. The association between white blood cell count and acute myocardial infarction mortality in patients≥ 65 years of age: findings from the cooperative cardiovascular project. Journal of the American College of Cardiology. 2001;38(6):1654-61.

20. Hu G, Baker SP. An explanation for the recent increase in the fall death rate among older Americans: a subgroup analysis. Public health reports. 2012;127(3):275-81.

21. Shannon CE. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review. 2001;5(1):3-55.

22. McKinney W, editor. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference; 2010: SciPy Austin, TX.

23. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

24. Haykin S, Network N. A comprehensive foundation. Neural Networks. 2004;2(2004):41.

25. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology. 1996;49(11):1225-31.

26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825-30.

27. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, editors. TensorFlow: A System for Large-Scale Machine Learning. OSDI; 2016.

28. developers s-l. 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier [05 December 2017]. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

29. Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

30. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research. 2011;12(Jul):2121-59.

31. Tieleman T, Hinton G. Lecture 6.5-RMSProp, COURSERA: Neural networks for machine learning. University of Toronto, Tech Rep. 2012.

32. Visualisation with TensorBoard [06 December 2017]. Available from: https://learningtensorflow.com/Visualisation/.

33. Welcome to pandas-ml's documentation! [06 Dec 2017]. Available from: http://pandas-ml.readthedocs.io/en/stable/.

VITA

Varisara Tansakul was born on March 29th, 1995 in Bangkok, Thailand. She completed undergraduate studies in Industrial Engineering in May 2016. Her research focuses on data analytics in healthcare.