APPROVAL SHEET

Title of Dissertation: Novel Machine Learning Tools to Enhance Mortality Prediction

Name of Candidate: Isaac Mativo
Doctor of Philosophy, 2021

Dissertation and Abstract Approved: Y. Yesha
Yelena Yesha
Professor
CSEE Department

Date Approved: June 21, 2021

Curriculum Vitae

Name: Isaac Mativo

Degree and Date to be Conferred: Ph.D., 2021

Collegiate institutions attended:

University	Dates	Degree
University of Maryland Baltimore County, Baltimore, MD, USA	8/2013 – 06/2021	Ph.D., Computer Science
University of Maryland Baltimore County, Baltimore, MD, USA	8/2007 – 12/2012	MS, Computer Science
University of Maryland Baltimore County, Baltimore, MD, USA	01/1998 – 12/2000	BS, Computer Science

Major: Computer Science

Research Areas: Machine Learning, Health Informatics, Big Data

Professional Publications:

- Validation, Consistency, and Conformance checking of Standard for Exchange of Nonclinical Data (SEND) Datasets. PHUSE Data Science Conference accepted. 2020.

- Hybrid Mortality Prediction using Multiple Source Systems. International Journal on Cybernetics & Informatics, published. 2019

- Smart Transformation of Clinical & Nonclinical Data for Insights. PHUSE Data Science Conference, published. 2019

- Patient Similarity in Clinical Outcome Prediction. The 4th Int'l Conf on Health Informatics and Medical Systems, accepted. 2018.

Talks and Research Presentations:

- "Using Patient Comorbidity to Improve ICU Mortality Prediction," Harvard University New England Science Symposium, March 2017

- "Improved Mortality Prediction for Diabetic Patients," 5th Annual Biomedical Informatics Symposium, Georgetown University, October 2016

- Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL), Johns Hopkins University, January 2015

- "Diabetes Comorbidities Analysis," PROMISE AGEP Research Symposium, February 2016

- "Predictive Modeling with Patient-Reported Data," PROMISE AGEP Research Symposium, February 2015

- "Electronic Health Records Impact on Safety and Efficacy," PROMISE AGEP Research Symposium, February 2014

- "Year-in-Review: Informatics in the Media 2013," American Medical Informatics Association (AMIA) Annual Symposium, November 2013

- "Personal Health Records," UMBC Computer Science Research Day, April 2012

Professional Positions Held:

2018 - 2021:	Adjunct Faculty, Health Information Technology University of Maryland Baltimore County, Baltimore, Maryland, USA
2018 - 2021:	Director, Development Projects PointCross Life Sciences, Silver Spring, Maryland, USA
	Project and Program Manager, Risk Modeling and Simulation Tool (RMST) Food and Drug Administration - Center for Tobacco Products (FDA/CTP), White Oak, Maryland, USA
2015 - 2017:	Program Manager, National Institutes of Health - National Cancer Institute Frederick National Lab for Cancer Research - Operated by Leidos Biomedical Research, Rockville, Maryland, USA
2013 - 2016	IT Project Lead (Part-Time), UMD Center for Integrative Medicine Baltimore, Maryland, USA
2009 - 2011	Programmer Analyst, CareFirst BlueCross Blue Shield Owings Mills, Maryland, USA
2001 - 2007	Software Engineer, Computer Sciences Corporation Rockville, Maryland, USA

ABSTRACT

Title of Dissertation: NOVEL MACHINE LEARNING TOOLS TO ENHANCE MORTALITY PREDICTION

Isaac Mativo, Doctor of Philosophy, 2021

Directed by: Professor Yelena Yesha
Department of Computer Science and Electrical Engineering

The tremendous growth of clinical data, both in healthcare institutions and personal devices, has led to the development of smart systems that can analyze this data to understand patterns and make predictions. The goal of understanding this data is to help make informed clinical decisions that improve health outcomes and save costs. Patient data has been leveraged by these intelligent systems to notice features that are important and actionable in improving clinical outcomes. One of the key clinical outcomes especially for intensive care unit (ICU) patients is mortality. Mortality prediction for ICU patients continues to be an important challenge in clinical care because of the complex nature of patients and data involved. With accurate mortality prediction, patients can be classified into different risk categories to ensure proper care is given therefore allowing for the best possible clinical outcomes and cost savings. Several mortality prediction models have been developed some of which use Machine Learning techniques. However, the existing mortality prediction tools are not generalizable to entire patient populations and are not designed to predict a diverse range of clinical endpoints. Additionally, these mortality prediction models are not personalized to the patients' unique and changing clinical profile.

In this dissertation, we present a mortality prediction approach that addresses these challenges thereby allowing for generalizable, extensible, and personalized mortality prediction modeling. We achieve this goal while improving prediction accuracy over existing state of the art tools. We present a novel methodology of combining clinical biomarker data, patient similarity features, and existing prediction tool output to predict ICU mortality. We examine patient comorbidities to discover their impact on mortality and use them as attributes in our modeling. We capture biomarker features as used in a severity scoring tool as in integral input in our modeling. In so doing, we have designed an approach to clinical prediction that uses matured machine learning techniques to combine patient similarity measures and patient biomarker information to improve prediction accuracy. Our approach builds on previous work that has been done in clinical outcome prediction and lays a strong foundation for continued innovation.

NOVEL MACHINE LEARNING TOOLS TO ENHANCE MORTALITY PREDICTION

By

Isaac K. Mativo

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:

Professor Yelena Yesha, Chair/Advisor
Professor Tim Oates
Professor Yaacob Yesha
Research Professor Milton Halem
Assistant Professor Michael Grasso

Dedication

Dedicated to my parents, David and Joyce Kivuva, who taught me the value of knowledge, my wife, Christine, for all the support, and our children, Kivuva, Mutanu, Mutheu, and Kavula.

Acknowledgments

I am eternally grateful to my Advisor, Dr. Yelena Yesha, who was patient with me and guided me through the process. Thanks also to my Ph.D. Committee Dr. Tim Oates, Dr. Michael Grasso, Dr. Yaacov Yesha, and Dr. Milton Halem. Without their guidance, insights, and support, this thesis would not be possible.

Thanks too to Dr. Quan Zhu for her mentorship early in my research. Thanks to Dr. Renetta Tull, Dr. Wendy Carter-Veale, and the entire PROMISE organization for the encouragement in the face of multiple competing priorities. I am also thankful to Dr. Braulio Cabral of the National Institutes of Health (NIH) National Cancer Institute (NCI) for helping me to stay focused on what is important. Thanks to Dr. Bill Rollow of the University of Maryland School of Medicine Center for Integrative Medicine for helping me understand the value of informatics in patient-centered care provision. Thanks to Dr. Ed Muchene for coaching me and believing in me.

Special thanks to the UMBC community, Computer Science and Electrical Engineering Department students, professors, and administrators. There are many others that were helpful in this dissertation that I have not mentioned, and for all of them I’m truly thankful.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

APACHE	Acute Physiology and Health Evaluation
APS	Acute Physiology Score
API	Application Programming Interface
AUC	Area under the Receiver Operating Characteristic Curves
ANN	Artificial Neural Network
CASUS	Cardiac Surgery Score
CCU	Coronary Care Unit
CDS	Clinical Decision Support
CFS	Correlation Feature Selection
CSRU	Cardiac Surgery Recovery Unity
DT	Decision Trees
ESN	Echo State Network
EHR	Electronic Health Record
EMR	Electronic Medical Record
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7
ICU	intensive care unit
ICD	International Classification of Diseases
LODS	Logistic Organ Dysfunction Score
LSTM	long short-term memory
MICU	Medical Intensive Care Unit
ML	Machine Learning
MIMIC	Medical Information Mart for Intensive Care
MPM	Mortality Probability Model
NV	Naive Bayes
NIH	National Institutes of Health
NLP	Natural Language Processing
OASIS	Oxford Acute Severity of Illness Score
PSM	patient similarity matrix
PART	Projective Adaptive Resonance Theory
RF	Random Forests
RNN	Recurrent Neural Networks
RRT	Renal Replacement Treatment
SICU	Surgical Intensive Care Unity
SOFA	Sequential Organ Failure Assessment
SAPS	Simplified Acute Physiology Score

SVM	Support Vector Machines
TRISS	Trauma and Injury Severity Score
TSICU	Trauma Surgical Intensive Care Unit

Chapter 1 Introduction

Intensive care unit (ICU) mortality prediction remains an imperative area of research. This is because ICU departments contribute to 22% of hospital costs while accounting for less than 10% of the hospital beds (Pirracchio et al., 2015). In addition, identifying the drivers of ICU mortality will aid in improving health outcomes by influencing healthcare policies and deciding on the use of innovative treatments. As a result, several ICU mortality prediction tools have been developed over the years. Although these prediction tools have measured levels of success, several areas of improvement remain yet concerning calibration and personalization. More recently, in precision medicine, patient similarity has been used to further improve clinical outcomes. The patient similarity is a concept that refers to the classification of patients based on their clinical characteristics such as laboratory tests, pharmacotherapy, comorbidities, disease history, treatments, hospitalizations, etc. (N. Wang et al., 2019). This sort of information is often accessible via electronic medical records (EMRs).

The analysis presented in this dissertation describes a novel approach to ICU mortality prediction that increases predictive accuracy by incorporating machine learning techniques to augment the SOFA severity scoring tool with both patient similarity based on International Classification of Diseases (ICD) -9 codes and a comprehensive view of patient comorbidity. This helps eliminate the conventional pitfalls of lack of personalization and inadequate prediction calibration, resulting in a more accurate and generalizable prediction of ICU mortality.

1.1 Machine Learning

In computer science, machine learning (ML) is a set of mathematical and statistical techniques implemented as a computer algorithm used to extract information from large sets of data to classify, predict, estimate, and regression (Campion, Carlsson, & Francis, 2017). Machine learning tools solve problems by inferring a solution from a dataset through different methods. ML has been around since around 1980. At this point, it is well-matured with numerous applications spanning multiple industries.

Machine learning has been widely applied in healthcare to develop solutions in disease diagnosis and detection, disease risk prediction, health monitoring, healthcare discovery, and epidemic prediction (Campion et al., 2017; Jayatilake & Ganegoda, 2021). Several studies have shown that the employment of ML has the potential to enhance the efficacy of the healthcare system, reduce the rising cost of healthcare, and advance personalized medicine (Panch, Szolovits, & Atun, 2018) (van der Schaar et al., 2020) (Ngiam & Khor, 2019). This study will employ Machine learning techniques to develop a predictive model for ICU mortality prediction.

1.1.1 History of Related Work

Predictive modeling uses various mathematical approaches to predict outcomes. These approaches include statistical and machine learning techniques. Generally, anticipating the result of a future event is possible through predictive modeling via machine learning approaches. The predicting model receives input and processes data (or, more precisely, attributes or features of the event) and then determines the probability of an outcome based on them. Models use at least one classifier to evaluate whether a set of data belongs to a particular class. A classifier is a mathematical algorithm that identifies which category an observation belongs to. A classifier, for example, may be used to determine whether an incoming email is spam or legitimate. Some of the most frequently used algorithms for prediction include the following:

- **Naïve Bayes**: Probabilistic classifiers that make a strong assumption about the independence of features. They are based on Bayes' theorem, associating current belief to prior belief, and manipulating conditional probabilities. Bayes' classifiers assume that all predictors contribute equally to the prediction of the outcome. According to Bayes' theorem, the likelihood of an event A given event B is computed below.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

- **Linear Regression**: A statistical technique used to model the relationship between a dependent variable and one or more independent variables. It is concerned with the conditional probability of the dependent variable given the independent variables. Linear regression can be used both to predict and quantify the strength of independent variables to dependent variables. The equation for linear regression below shows the dependent variable Y as a function of the independent variable X and other parameters.

$$Y_i = f(X_i, \beta) + e_i$$

- **Support Vector Machines (SVMs)**: Supervised models that are useful for classification and regression analysis. SVMs are associated with learning algorithms that analyze and recognize patterns. They define the optimal hyperplane for practices that are linearly separable. They help solve the problem of local minima.
- **Artificial Neural Networks (ANNs)**: These computational models are used to estimate functions that are highly dependent on a large number of inputs. Animals' central nervous systems serve as inspiration for ANNs. Due to their adaptive nature, they are capable of learning and recognizing a pattern. ANNs are helpful in various activities such as handwriting and speech recognition and computer vision. An example of ANN is the Echo State Networks (ESNs) that effectively predict chaotic time series. ESNs are examples of Recurrent Neural Networks (RNNs), where units are connected in a directed cycle.

ANNs have two disadvantages: they can become trapped in local minima and are susceptible to overfitting. The first occurs because of the iterative nature of moving in the direction of gradient descent. In contrast, the second can occur if training is continued iteratively until the

ANN considers noise as part of the pattern. A properly configured neural network can assure distinctive discrimination of noisy or missing data. The conceptual model of an artificial neural network is shown in figure 1 below, which includes an input layer, two hidden layers, and an output layer. The input layer received data, the hidden layer(s) performs the mathematical computations based on the input data. Whereas the output layer presents the results of the analyses. When there are many hidden layers, the word "deep learning" is used.

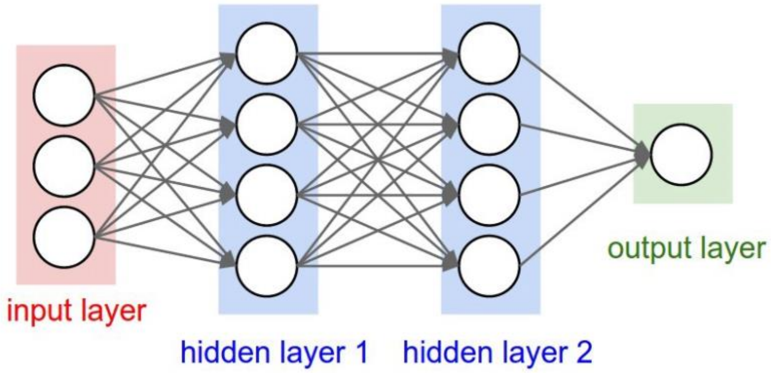


Figure 1- Artificial Neural Network

- **Logistic Regression**: This is a probabilistic classification model used to predict the outcome of a dependent variable based on one or more features. The number of dependent variables in logistic regression is binary (two available categories). The Trauma and Injury Severity Score (TRISS) was developed using logistic regression to predict mortality in injured patients. Logistic regression can be used to predict whether a patient has a disease based on observations (or attributes) such as age, gender, cholesterol level, etc., with each attribute being either continuous (e.g., level of cholesterol) or categorical (e.g., gender). Logistic regression models, also known as generalized linear models, have two operations. The first is a linear operation with a multiple linear regression obtained from the predictor variables. The second section is a sigmoid function for estimating the probability of belonging to a given class. In the following equation,

$$P(class|x_i, \ldots, x_n) = \frac{1}{1 + \ e^{-}(w_0 \sum_{k=1}^{N} w_k . x_k)}$$

N denotes the number of predictor variables, *w*$_k$ is the model parameters, and *x*$_k$ signifies the variables for a given patient.

- **Decision Trees (DT)**: These estimate a target value using a collection of binary rules. Decision trees (DT) can predict the value of a dependent variable using one or more independent variables. If the target variable is continuous, this is called a regression tree; otherwise, it is a classification tree. Each interior node of a DT represents an input variable; its leaves represent target variables based on the input variables defined by the path from the root to the leaf. A common problem with decision trees is overfitting, which can be mitigated by pruning. As a disadvantage, overfitting can occur if the training set is not well-distributed. This implies the training set must contain a sufficient number of entities for each performance to be predicted; otherwise, the prediction is either underfitted or overfitted. Two advantages of decision trees are that they are simple and non-parametric. Their simplicity allows them to be used as explanatory models. However, their non-parametric nature implies that they make no assumptions on parametric form for class densities and thus expand following the nature of the problem.
- **Random Forests (RF)**: This is an ensemble model, meaning that it computes a response by using results from different models. Several decision trees are generated, and the result is obtained based on the sum of the decision trees' outputs. In contrast to DT, RF is not

susceptible to overfitting problems. If used for classification, the result response is the decision trees' mode response (also called voting). In the case of regression, the outcome is the mean of the responses from all the DT. An issue in regression is that it has a tendency to overestimate lower values and underestimate higher values. This is because the response is computed from the mean. When used for regression, random forests cannot predict beyond the range of the training data.

1.2 Prediction Tools in Healthcare

Patient outcome prediction is at the heart of medical decision-making (Kattan & Gerds, 2020). Many studies have investigated various clinical attributes such as biomarkers, genetic features, medical history, and others to predict future clinical outcomes. Consequently, statistical prediction models and other prediction algorithms have been used to predict outcomes and help in decision-making. The sophistication of these prediction methods has evolved over time to include novel biomarkers, artificial intelligence (AI) approaches, and big data. Ultimately, they seek to answer the question of the most accurate prediction model to a disease. For this reason, the prediction model is based on a clear framework that defines the target patient population and the predicted outcome.

Prediction tools in healthcare come in different approaches of varying strengths and weaknesses. One approach involves identifying and listing the risk factors that contribute to disease. These factors are generated using univariate or multivariate analysis. They are tested for statistical significance in predicting disease using indicators such as p-values and hazard ratios. However, statistical significance does not necessarily translate to clinical significance. It may also be the case that some risk factors may be missed when using this approach. One significant drawback of this outcome prediction approach is that the risk of listed risk factors may not be actionable.

Another approach that has been used for predictive models in healthcare is counting the risk factors. A simplistic and less accurate way of achieving is simply using a sum of all the known risk factors of a disease for prediction. A problem with this approach is that all elements are given the same weight. In contrast, some continuous variables may not be easily counted. To overcome these weaknesses, points or weight coefficients can be added to each risk factor. This does not eliminate the challenge of categorizing continuous variables. An approach that uses monograms overcomes the challenge of using continuous variables. Nomograms are graphical representations of models and can be used mainly in two ways (Balachandran, Gonen, Smith, & DeMatteo, 2015).

Medical nomograms are used to illustrate a predictive prognostic model that predicts a clinical outcome based on clinical markers. Nomograms can represent linear and nonlinear continuous variables without requiring them to be categorized. For prediction, nomograms can be used in two ways. Firstly, a pictorial representation representing each variable with assigned points associated with its magnitude. The summation of the facts is then matched to a scale of the outcome. An example is shown in figure 33 below, where the recurrence-free survival (RFS) in a resected primary gastrointestinal stromal tumor (GIST) is estimated. Part B of the figure shows the corresponding calibration curves. Nomograms are not appropriate for all clinical questions. However, they are ideally suited for questions that are narrowly defined and require mathematical modeling.

The use of decision trees has been a popular approach in clinical outcome prediction. These are easy to construct and build. A patient is guided by a series of questions, with each question requiring a decision on how to proceed to the next. A question may be a risk factor or any other clinical data point. At the end of the series, the patient is classified in a category. One drawback of this approach is the possibility of variation within a category. For instance, a category labeled as ‘high risk’ may contain critically ill patients and others who barely made it. Consequently, decision trees are not ideal for individualized predictions but can suffice for group classifications.

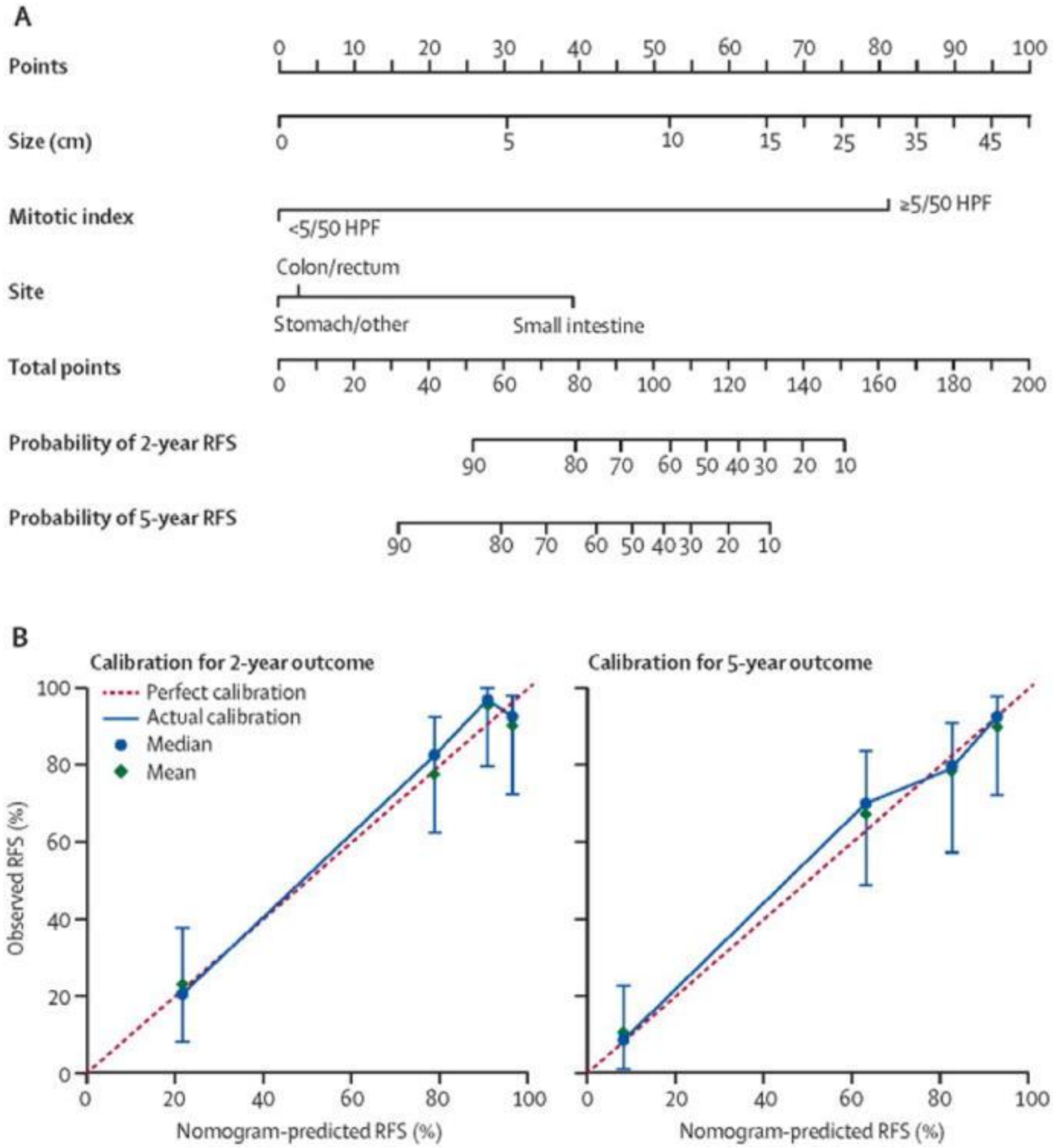


Figure 2 Nomogram Example (adopted from Balachandran et al.)

The use of online risk calculators is a more recent prediction method (Kattan & Gerds, 2020). These are available on several websites, such as Cleveland Clinic website which frequently provide the sources backing their approaches. Some of the online risk calculators are comprehensively defined and provide the requisite authoritative references. These are often beneficial for individual patients since they collect data directly from the patient. A drawback to this approach is the lack of standardized accreditation for these tools, making them susceptible to inferior or incorrect tools. These can then develop into a clinical risk if used by patients to make healthcare decisions.

Numerous pieces of patient clinical data are preserved in locations such as electronic medical records (EMRs). To use this data effectively in clinical practice, it becomes necessary to find effective ways to sift through the data and identify data that assist clinicians in making decisions. For this reason, predictive analytics are used to estimate outcomes and uncover associations that would otherwise be difficult to identify.

In healthcare, a prediction is most useful when it can be converted into practice. To maximize the leverage of prediction, the predictors should not be isolated from the decision-makers. Therefore, predictors should be available at the point of care, i.e., real-time, or as close to real-time as possible. Risk stratification models have been used to separate patients into high-risk, low-risk, and rising risk. Equipped with this knowledge, clinicians can concentrate their efforts on the most vulnerable patients.

An example of a stratification method is the Adjusted Clinical Groups that use inpatient and outpatient diagnoses to classify patients into one of 93 categories and is helpful for hospital utilization prediction. Another example is the Charlson comorbidity index, which estimates a patient’s ten-year mortality rate with comorbid conditions. Together with other variables such as age, gender, length of stay, etc., this index can be used as an input to a predictive algorithm to predict readmission.

Another way to look at how analytics is used in healthcare is the Healthcare Analytics Adoption Model (Sanders, Burton, & Protti, 2013). This defines the following 8 levels of maturity organizations go through, from level 0 to level 8, as detailed in Table 1.

Table 1- 8 levels of maturity organizations

Level	**Detail**
Level 8	Personalized Medicine & Prescriptive Analytics
Level 7	Clinical Risk Intervention & Predictive Analytics
Level 6	Population Health Management and Suggestive Analytics
Level 5	Waste & Care Variability Reduction
Level 4	Automated External Reporting
Level 3	Automated Internal Reporting
Level 2	Standardized Vocabulary & Patient Registries
Level 1	Enterprise Data Warehouse
Level 0	Fragmented Point Solutions

1.2.1 ICU Mortality Prediction

Prior to patients being admitted to the ICU, a decision must be made about their admission. Prominent factors considered in determining admission include age, the severity of illness, prognosis, diagnosis, etc. (Smith & Nielsen, 1999). In addition, when patients are admitted, additional data are captured, and procedures performed. All this creates a set of data represented in different datastore or conformed to various formats. As clinicians make decisions about care, they rely on this data to guide them. In critically ill patients, time becomes essential, and good decisions need to be made quickly. Therefore, access to tools that can aid decision-making is highlighted. One critical piece of information that is helpful is the risk of death. Many of these tools rely on data collected within the first 24 hours of ICU admission. A brief look at the state-of-the-art studies reveals that several tools that predict ICU mortality have been developed, which will be reviewed in the following section.

1.3 Patient Similarity

The patient similarity is receiving increased interest to better predict clinical outcomes (Chan, Chan, Cheng, & Mak, 2010); (Brown, 2016). Many healthcare data is generated and stored in various mediums such as electronic health records, mobile devices, and social media. Consequently, several approaches to data analytics have been employed to convert observations into actionable ideas. Specifically, various studies have explored the value of using patient similarity in predicting outcomes such as mortality (F. Wang, Hu, & Sun, 2012); (Gottlieb, Stein, Ruppin, Altman, & Sharan, 2013).

Patient similarity examines the closeness of various items of patient data to determine how patients can be grouped together based on their closeness, as illustrated in Figure 3. Patient similarity has been used as a tool to enable precision medicine (Parimbelli, Marini, Sacchi, & Bellazzi, 2018). Various other applications can be implied using patient similarity measures, such as patient cohort selection for treatment effectiveness research. Often, the data used for patient similarity is found in the electronic health records (EHR) and may include lab results, diagnostic codes, prescribed therapies, etc.

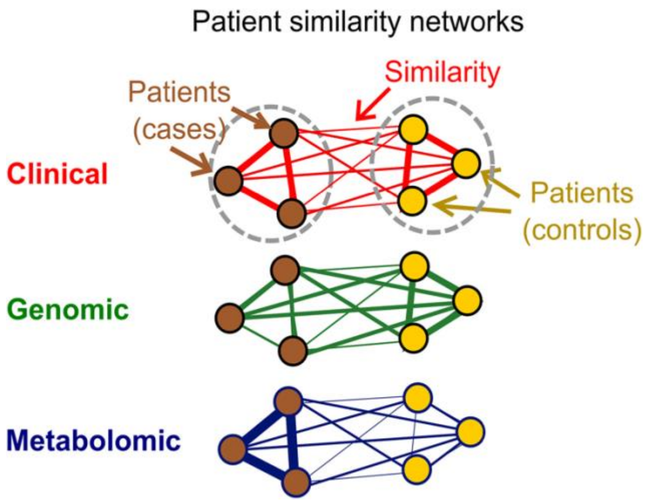


Figure 3- Areas of Patient Similarity (the image is adopted from (Pai & Bader, 2018))

Generally, patient similarity measures are challenging because of patient data's available high dimension space (Parimbelli et al., 2018). A balance must be maintained between having a cohort of similar patients that can produce generalizable clinical trial results. On the other hand, avoiding making each patient too unique. Three general approaches have been used to define patient similarity:

1. Patient similarity based on molecular data at the point of care. This may include measures based on genomics, transcriptomics, proteomics, and metabolomics.
2. Patient similarity based on different types of data to create novel meaningful subgroups. This data is taken from various sources such as electronic medical records.
3. Patient similarity based on clinical data collected when the patient is admitted to a care center.

1.3.1 A Brief Look to Patient Similarity Measurement

The patient similarity allows for establishing a framework that can enable predictive models that are accurate, generalizable, able to handle missing information, and able to integrate heterogeneous data (Pai & Bader, 2018). The patient similarity is based on comorbidities, medications, demographics, lifestyle, risk factors, etc. More recently, genome information such as mutations and Systems Biology has also been used to cluster patients (Breitling, 2010; Gottlieb et al., 2013). Systems Biology focuses on biological components such as cells, molecules, and systems. These biological components can then be considered through the clustering of patients to determine closeness in similarity. For genomics, actionable mutations of interest are identified in patients and can group patients together.

The success of these patient clusters depends in part on the correctness and completeness of patient data (Pedersen et al., 2017). Missing data can be classified into several categories, including missing entirely at random, missing at random, and missing not at random (Ibrahim, Chu, & Chen, 2012). Several techniques can be used to account for the missing data, such as single value imputation, complete case analysis, and worst-case and best-case scenarios (Hu et al., 2017).

One challenge in dealing with patient similarity is highlighting the exceptional cases that differentiate patients (Sharafoddini, Dubin, & Lee, 2017). This information is essential to capture because not all clinical events necessarily cause patients to be differentiated. Therefore, correctly identifying clinically significant events are an essential factor in creating accurate patient clusters.

Additionally, drugs have been used to cluster patients. Patients are grouped here according to the medications they are prescribed. The similarity between drugs can be determined by their chemical structure or target. The use of drug similarity coupled with patient similarity and prior patient-drug associations in creating clinical predictive models has been shown to yield favorable results w selecting new or rarely used drugs for patients (P. Zhang, Wang, Hu, & Sorrentino, 2014).

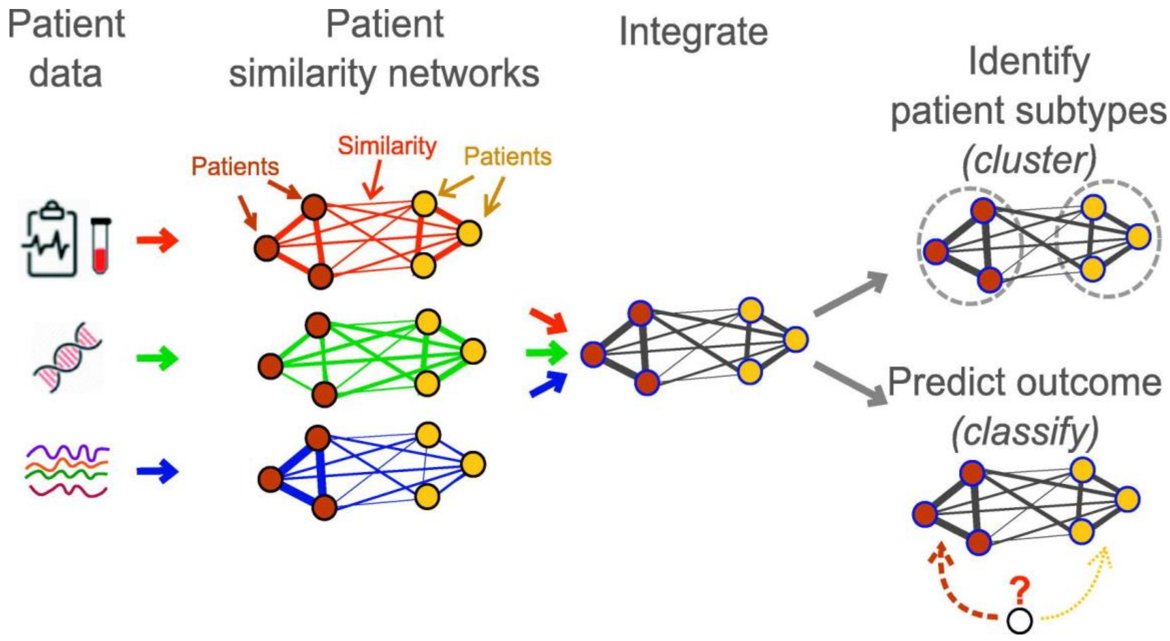

Patient data
Patient similarity networks
Patients
Similarity
Patients
Integrate
Identify patient subtypes (cluster)
Predict outcome (classify)
?

Figure 4- Patient Similarity Process (adopted from (Pai & Bader, 2018))

1.3.2 Patient Similarity Calculation

Similarity among patients can be computed based on several patient characteristics as shown below:

I. Based on age, as

$$FS_A(i,j) = \frac{min(Age_i, Age_j)}{max(Age_i, Age_j)}$$

Where Age_i and Age_j are the ages of patients *i* and *j*, respectively.

II. Based on sex, as

$$FS_A(i,j) = \begin{cases} 0 & if\ patients\ i\ and\ j\ have\ different\ sex \\ 1 & if\ patients\ i\ and\ j\ have\ the\ same\ sex \end{cases}$$

Where *i* and *j* are patients.

III. Based on laboratory test, as

$$d_{lab}(i,j) = \sqrt{(L_{i1} - L_{j1})^2 + (L_{i2} - L_{j2})^2 + \ldots + (L_{im} - L_{jm})^2}$$

$$d' = \frac{d_{lab}(i,j) - min(d_{lab})}{max(d_{lab}) - min(d_{lab})}$$

$$FS_L(i,j) = 1 - d'$$

Where *m* is the laboratory test normalized to L_{xy}~$N(0,1)$, and L_{xy} represents the test *y* for patient *x*.

IV. Based on disease diagnosis, as

Elixhauser comorbidity index is used to measure similarity based on the diagnosis. This index is based on conditions described by the ICD-9-CM (International Classification of Diseases, Ninth Edition, Clinical Modifications) discharge records.

At its core, similarity calculation evaluates the difference between patients using a predefined criterion. This is identical to the product recommendation systems used in e-commerce. In clinical prediction, a retrospective analysis can predict future events once a group of similar patients is found. To be successful, a big enough sample size of similar patients must be used to make clinically valid extrapolations. Therefore, the challenge becomes one of developing a patient similarity matrix that is sufficiently large to include enough patients while being sufficiently restrictive to provide clinically beneficial differentiation. This challenge has not been adequately discussed by many tools that utilize patient similarity as a clinical outcome prediction feature.

1.4 Classical Approaches to Clinical Prediction

Effectively assessing patient similarity is a non-trivial exercise. Although research has shown that prediction models based on patient similarity led to a more definite final decision, more research is required to better calibrate patient similarity features. On the one hand, particular patient clusters result in overfitting and inappropriate personalized medicine in general. On the other hand, loosely defined clusters will not adequately represent the unique and independent groups.

A network medicine approach may be a viable way to produce a patient similarity output. In Figure 5, various patient attributes considered essential for patient similarity, such as medication, comorbidity, age, etc., are presented as a network layer resulting in the patient similarity output.

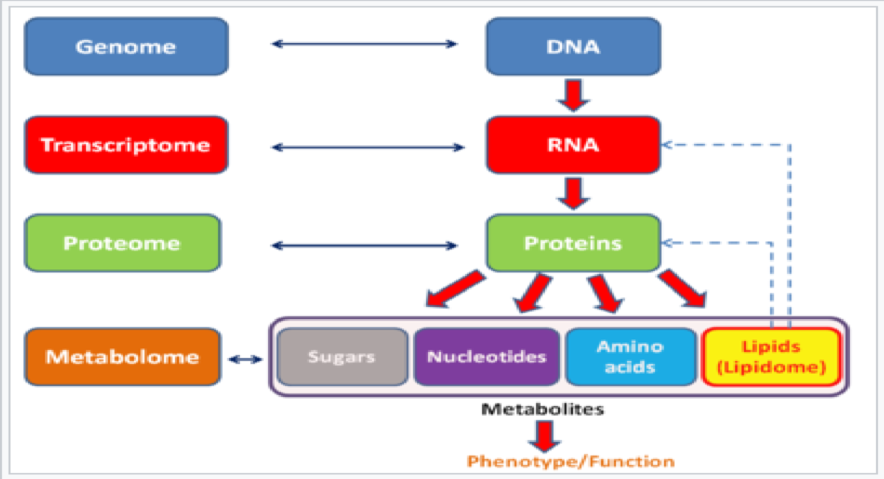


Figure 5- Omics Data in Patient Similarity

Determining patient similarity clusters provide an additional layer of complexity for critically ill patients. Essential aspects highlight the significance of attribute ranking in establishing accurate patient similarity groups since the priority for a diabetic patient hospitalized in an intensive care unit (ICU) with a higher ICU or diabetes ranking and an attribute to a predictive learning model. To provide a correct decision, attribute selection can be experimented upon to determine optimal attributes.

The practice of medicine is a series of questions directed toward a specific outcome, e.g., determining a diagnosis. A clinical decision support tool assists clinicians in identifying underlying patterns, uncovering hidden associations, processing large volumes of data, etc. However, the clinician ultimately makes the final decision. A hybrid framework that incorporates real-time clinician input is more flexible to the often evolving patient profile and suggests more suitable patient similarity classes.

Concerning the hybrid system, as mentioned earlier, a predictive tool that continually learns based on previous results and new data is vital since it allows for patient similarity to shift based on new data. For example, suppose a new drug is shown to produce adverse effects on some patients in a cluster. In that case, this observation, if clinically significant, should trigger a subdivision in the patient cluster.

Temporal information is vital in classifying similar patients. For example, for patients with comorbidities, the order of diagnosis and treatment is significant. Several electronic medical record chart events contain a timestamp. This temporal information can be used to further refine patient similarity clusters. The processing of temporal information in medical narratives remains a significant challenge. However, time representation in medical natural language processing is not well developed (Zhou & Hripcsak, 2007).

An extendable predictive tool would adapt to changes in data. This is particularly important since a patient’s condition often changes in the ICU, and new readings are recorded almost constantly. In addition, procedures may be performed, which must be accounted for when making new predictions about clinical outcomes. Therefore, it is necessary to have a feedback loop that regularly updates the training models to obtain results based on the most current available information.

One goal of patient similarity research is to develop data integration strategies into clinical decision support systems (Parimbelli et al., 2018). In the quest to enhance precision medicine through the discrimination of genetic, biomarker, phenotypic, and other patient attributes, patient similarity approaches highlight the difference among patients. This can potentially undermine the goal of obtaining sufficiently similar patient cohorts for clinical trials to produce generalizable outcomes. Therefore, the task of defining the appropriate granularity of patient cohorts becomes a challenging task. Defining a similarity measure between patients then becomes a crucial step in allocating patients into clinically meaningful cohorts.

1.5 Measuring Quality and Improving Clinical Outcomes

The measure of quality is inextricably linked to clinical prediction. This is because clinical outcomes are impacted by the quality of care. Clinical Quality Measures (CQL) are a mechanism for evaluating patient care observations, treatment, processes, experience, and/or outcomes (Mosko, Leiman, Ketwaroo, Gupta, & Quality Measures Committee of the American Gastroenterological Association, 2020). Therefore, the prediction of clinical outcomes must either presume an acceptable quality or incorporate quality measures.

Over the last few decades, the methodology that has been used to apply quality measures has evolved through the work and mandates by several organizations. These include the National Committee for Quality Assurance (NCQA), the National Quality Forum (NQF), and the Physician Consortium for Performance Improvement (PCPI) convened by the American Medical Association (AMA), among others. These organizations are dedicated to advancing clinical practice excellence, primary care, and specialty care (McCormick & Gugerty, 2013). Additionally, they address performance measures, best practices, reporting guidelines, and frameworks. Consequently, assessing patient care outcomes then becomes a shared goal between quality measurement and clinical outcome prediction.

The Agency for Healthcare Research and Quality (AHRQ) has listed 28 Inpatient Quality Indicators (IQIs) that are used to provide a perspective quality (Mull, Borzecki, Chen, Shin, & Rosen, 2014). These are categorized into four groups: volume indicators, mortality indicators for inpatient procedures, inpatient conditions, and utilization indicators. Volume indicators consider the number of complex procedures performed in the hospital to understand that the more the procedures, the better the outcomes. Mortality indicators for inpatient procedures show the quality of care concerns based on higher mortality in certain hospitals than others for similar procedures. Mortality indicators for inpatient conditions show higher mortality rates in certain hospitals than others, indicating a more inferior quality of care. Finally, utilization indicators focus on the overuse, underuse, or misuse of procedures noted to vary significantly across hospitals.

Another important aspect of quality in mortality prediction is the health-related quality of life (HRQoL). One analysis measuring responses from Short Form 12 (SF-12) survey forms from seniors showed that the SF-12 scores were predictors of hospitalization and mortality (Dorr et al., 2006). This is important because it is now easier to collect self-reported or survey data from patients. Although this data collection is done before patients are hospitalized, this information can augment existing EMR data and add value to mortality prediction models.

1.6 Use of Biomarkers

According to the National Institutes of Health (NIH), a biomarker is a quantifiable biological parameter measured and evaluated as an indicator of normal biological, pathogenic, or pharmacologic responses to a therapeutic intervention (FDA-NIH Biomarker Working Group, 2021). Biomarkers have played an essential role in the diagnosis, detection, and treatment of many diseases. This is especially true in cancer, where gene data has been used to identify the association of the gene ALK with lung cancer. Appendix D shows some predictive and diagnostic biomarkers used in the cancer treatment decision.

Biomarkers may potentially be used for personalized, preventive, and predictive medicine. This is part of the Human Genome Project’s pledge, and significant progress has been made to that end. However, some roadblocks use biomarkers for predictive purposes, as the focus is primarily on their prognostic value. There is a lack of awareness of the full potential of the biomarkers (Simon, 2011). The types of biomarker applications are depicted in figure 4 below.

Genetic biomarkers have been developed for predicting the risk of diseases in individuals. However, their use in prediction modeling has been limited by weak associations between the biomarkers and risk of contracting diseases. Part of this observed weak association is the heterogeneity of some conditions, such as cancer (Simon, 2011). For example, in breast cancer research, estrogen receptor-negative and estrogen receptor-positive are different in somatic mutations and responsiveness to treatment. This observation is present in many other chronic diseases that are phenotypically and molecularly heterogeneous. This makes it a challenge to perform research based on broad genome-wide associations.

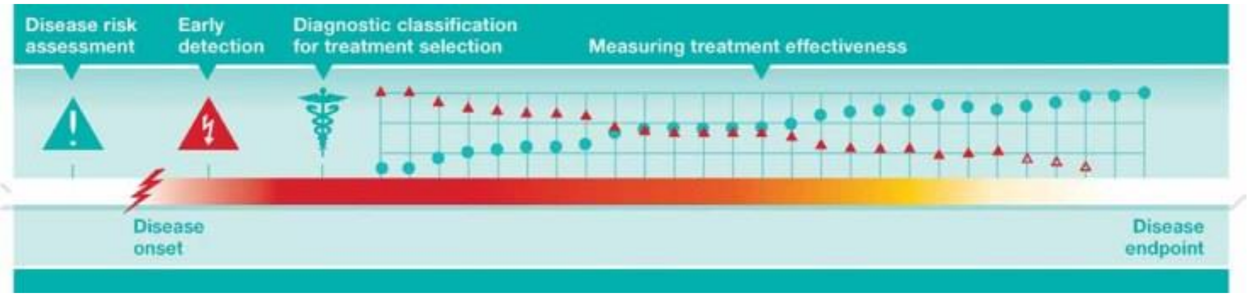

Disease risk assessment
Early detection
Diagnostic classification for treatment selection
Measuring treatment effectiveness
Disease onset
Disease endpoint

Figure 6- Categories of the intended use of biomarkers (adapted from Simon, 2011)

Another challenge posed in predictive modeling for genetic biomarkers is that chronic diseases are most likely due to multiple genetic polymorphisms caused by genetic and environmental factors. This then becomes more difficult to find and associate distinct genetic properties with diseases. Also, as in the example of most cancers, a complex series of somatic mutations interact with each other to affect the tumor progression (Maxwell et al., 2008).

1.7 The Problem and Approach

While considerable research has been conducted on ICU mortality predictive models, to our knowledge, no work has been done on creating a mortality predictive tool that is personalized and extendable, and sufficiently generic to be used by the general ICU patient. Currently, the available resources are either too rigid or customized for a particular patient profile set, making them unsuitable for general deployment.

1.7.1 Problems

The analysis of state-of-the-art methods reveals that the current approaches of ICU mortality prediction generally share three shortfalls as follows:

1.7.1.1 Problem 1: Lack of a generalizable ICU mortality predictive model

Most predictive models are either too specific to a patient cohort or too general in scope. Traditional predictive models use all the available data on a patient, typically derived from electronic health records. This data contains some information that may be irrelevant for predicting mortality, thus increasing the amount of noise present in model creation. It has been a challenge to create pragmatic models that can be generalized over a broad enough patient category.

1.7.1.2 Problem 2: Challenge in incorporating patient similarity in clinical predictive models

Several techniques have been used to compute patient similarity. However, the data used for calculating the similarity between two patients is often the case either not available in electronic health records or is at such a granular level that it cannot be used for more substantive patient populations.

1.7.1.3 Problem 3: Lack of a flexible hybrid framework for patient mortality prediction

Although several mortality prediction models exist and a few more have been proposed, there is no standard framework of patient mortality prediction customized for selected patient cohorts. Such a framework would allow prediction over a broad patient population, with relatively minimal customizations made to specific cohorts.

1.7.2 Approach

Several investigations have been done on mortality prediction in intensive care patients. To the best of our knowledge, no study takes comorbidity into account for mortality prediction, which is our primary focus for this present study. The presented research incorporates a hybrid approach to predicting patient mortality by combining machine learning algorithms, pre-existing mortality prediction scores, and patient comorbidity.

Dependent variable: Mortality

Independent variables: Patient similarity measure, Sequential Organ Failure Assessment (SOFA) score, and Elixhauser comorbidity scores

The SOFA score is chosen because it is developed from a large sample of ICU patients and is comparable to our patient cohort. The SOFA score is validated in this study using our patient cohort. Additionally, to identify the comorbidities of the patients, the Elixhauser Comorbidity Index is used (Elixhauser, Steiner, Harris, & Coffey, 1998). Also, the SOFA score is combined with the comorbidities and to create a hybrid prediction.

1.8 Contribution to Knowledge

Our work is inspired by and builds upon prior bodies of research done on healthcare prediction. This includes work on the Super ICU Learner Algorithm (SICULA) and other work on mortality prediction comparing SAPS II, SOFA, and APACHE II (Pirracchio et al., 2015) (Xia et al., 2019). We have realized that it would be of immense value to propose an approach to ICU mortality prediction that uses existing prediction tools while also focusing on the uniqueness of various patient cohorts. In our approach, we employ Machine Learning approaches that have evolved over time and provide robust algorithms for optimal results. At the core of our solution is a novel approach that combines Machine Learning tools with patient biomarkers as contained in the SOFA severity score, and similarities based on similar patient disease diagnosis. We use comorbidities to as additional attributes to improve on the accuracy of mortality prediction.

What defines our breakthrough is that it shows a system for predicting mortality using readily available data in the patient EHR. Our primary objective is to predict the mortality of patients hospitalized in ICU using patient similarity based on ICD codes and other traditional clinical biomarkers based on SOFA scores. Among these biomarkers are the fraction of inspired oxygen, partial pressure of oxygen, platelet count, bilirubin, and creatinine. Incorporated in these measures is the Glasgow coma score that measures the patient's level of consciousness.

We also recognize the challenge of converting biomarker exploration to practical clinical use (Selleck, Senthil, & Wall, 2017). The National Institutes of Health defines a biomarker as a quantified biological parameter measured and evaluated as an indicator of normal biological, pathogenic, or pharmacologic responses to a therapeutic intervention (FDA-NIH Biomarker Working Group, 2021). One beneficial clinical use of biomarkers is increasing diagnostic accuracy and outcome prediction with increased sensitivity and specificity.

Contributions of this research include the following three items:

1. **Patient Similarity**

   We present a method to account for patient similarity by using the ICD codes that identify the disease diagnosis of the patients.

2. **Clinical biomarker analysis**

   We repurpose the SOFA score quantifying the magnitude and frequency of organ failure to utilize measured parameters across six organs in our proposed ML mortality prediction model.

3. **Mortality prediction framework**

   We propose a framework for predicting mortality by considering a hybrid approach that incorporates readily available patient information utilizing known machine learning techniques. This flexibility enables the system to be tailored to individual patient cohorts without significant modification.

1.9 Thesis Organization

The remainder of this thesis will be structured as follows. Chapter 2 summarizes previous research on ICU mortality prediction. It begins by describing diabetes, the inclusion criteria for this study, and highlights the critical nature of ICU mortality prediction tools. Then it compares the latest ICU mortality prediction tools. Following that, it describes ensemble prediction models and the used techniques. It then explores the ICU risk factors that influence mortality and are often used as independent variables in mortality prediction. The chapter also describes the work that has been done on personalized ICU mortality prediction as well as comorbidities using the Elixhauser index. It then turns to present the mortality measures used in this research and the justification for the choice.

Chapter 3 details the methodology employed in this research. This section includes an introduction to the database the cohort used. Additionally, it also describes the methodology and techniques used to identify the comorbidities for the prediction model. It concludes with the mortality prediction results for both our prediction model and traditional prediction models.

Chapter 4 encapsulates the results obtained in our work, comprising the patient cohort selection, the mortality and comorbidity measure calculations, and the ranking of the most predictive comorbidities for mortality prediction. It concludes with the review of mortality prediction results and discussion of the findings. Finally, Chapter 5 includes the conclusion of this research and discusses the limitations of this work. Future study is also monitored to set the stage for further investigation.

Chapter 2

Survey of previous work

Clinical outcome assessment is not a novel concept. In 1863, Florence Nightingale started (Aravind & Chung, 2010) addressing the issues of evidence-based medicine. They improved care through the evaluation of treatments, procedures, and outcomes. Initially, outcome predictions were primarily based on the subjective opinions of the clinicians. Following the Copenhagen polio outbreak in 1952, the ICU concept arose (Kelly, Fong, Hirsch, & Nolan, 2014). Since then, the rapid growth of ICUs and subsequent compilation of qualitative and quantitative outcome measures led to better medical interventions.

Consequently, there have been several ongoing research efforts on prediction modeling in various domains, including within healthcare. For our work, we survey previous work focused on ICU mortality prediction. Much of the data used to predict mortality of ICU patients is based on the first 24, 48, or 72 hours of a patient's stay in the ICU. However, much of the data used to perform outcome prediction is missing or not yet available. Additionally, multiple innervations are performed in the ICU during the first few hours and days of a patient stay.

Making accurate decisions early following an ICU admission is crucial because of the clinical and financial implications. A good prediction early on allows both the patient and their families to make the best choices possible. For the patient, this can imply more patient-centered care that incorporates the treatment plan options. For the families, this involves financial obligations or even end-of-life conversations with loved ones. Given the high cost of ICU hospitalization, any prudent decision made early on translates to significant cost savings. Lastly, predicting outcomes can help inform clinicians on the kinds of conversations to have with the patients and their families.

This chapter discusses ICU data challenges and diabetes as a chronic disease to highlight the importance of mortality prediction for hospitalized diabetic patients. Later, material including the scores used in this study will be detailed. Finally, the conducted state-of-the-art research will be studied.

2.1 Challenges in ICU Data

The ubiquitous tracking and monitoring of patients in the ICU generate any data that can be quite valuable for analytics. This also provides data that clinicians can use to diagnose, treat, discharge, and follow-up with patient care. However, some challenges arise when dealing with data typically available in the ICU (Johnson, Ghassemi, et al., 2016a). We will address a few of these issues in this subsection.

2.1.1 Heterogenous Data

In the ICU, multiple measurements are done to determine a patient's status and track their progress. The measurements are done using various devices, capturing different sorts and formats of data. These include lab measurements from body samples such as blood and urine, real-time monitoring data such as the electrocardiogram (EKG) feeds, routinely collected data such as blood pressure, imaging data, service procedure codes for billing, and many more. The collected data is not integrated into a single system for several reasons, including privacy and technical challenges of harmonizing it. Therefore, it becomes challenging to integrate and curate this data for analysis.

ICU data heterogeneity contains the following attributes (Yun Chen & Hui Yang, 2014):

- Variable heterogeneity: ICU patient data contains many variables. These variables are necessary to capture the patient's complete profile, including any recorded trends. Quantitative data may be continuous or discrete, and qualitative data may be categorial or contain additional variables. Examples of quantitative data include heart rate, blood pressure, laboratory results, etc. Qualitative data examples include survey questions about one's level of pain and care, etc. Consequently, the commonly used univariate analysis cannot be used in ICU clinical data. Instead, it becomes important to understand the interaction of the heterogenous variables and their impact on clinical outcomes.

- Patient heterogeneity: Patients admitted to the ICU exhibit a diverse combination of characteristics. The patient population may be categorized by age, gender, type of surgery, disease, etc. Other subcategories may exist within these categories. This presents both a challenge but also an opportunity to do research based on various patient characteristics.
- Time asynchronization: Data is collected according to the clinician's discretion and may contain missing data or data that do not represent a uniform sampling of patients. For example, there is not a universally accepted standard on when data collection should occur in ICU. In addition, certain items recorded in the first 48 hours of ICU admission may be recorded at a non-uniform sampling rate.

2.1.2 Corrupt Data

The primary reason for collecting patient data is to enhance patient care, not to facilitate data analysis. As a result, when patient data is combined, it may contain gaps that challenge analysis. This could include missing data, erroneous data, or inaccurate data. The distinction between erroneous and imprecise data is that erroneous data does not reflect the truth. In contrast, inaccurate data, while accurate, does not capture the concept of interest. Erroneous data is problematic in the creation of clinical predictive models because it may cause false alarms. Such mistakes can be detrimental to clinical outcomes and financially costly to patient and hospital, not to mention possible litigation.

Missing data can be categorized into three types (Johnson, Ghassemi, et al., 2016b). These are (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR). MCAR occurs when the event causing the missing data is completely random, such as a medical equipment failure. In this case, bias is avoided in the imputation of the missing data. MAR occurs when the event causing the missing data is unrelated to the value of the data. MNAR occurs when the missing data depends on the value of the measurement. For example, a clinician may choose not to collect certain data if they suspect that the data is normal and would not add value to their decisions. MNAR is the most problematic for clinical analytics and prediction modeling.

Ambiguous data can complicate supervised learning. This is because the training of a model would require known labels. Therefore, if labels are fuzzy, prediction models can perform poorly. Several approaches have been proposed to mitigate this challenge. These include the provision of manual labels by experts or the use of anchors to define a feature in place of labels (Halpern, Choi, Horng, & Sontag, 2014).

2.1.3 Complex Data

Clinical data is complex because of several factors (Hunter et al., 2008). To begin, it is subject to rapid change over time. The data would often consist of:

- Data that is constantly monitored for physiological values. For instance, the heart rate is constantly captured and sampled every few seconds.
- Data that contains discrete events such as results of blood tests and other laboratory findings.

This dynamic nature is much more prevalent in an ICU setting, where interventions can dramatically change registered biomarkers. Where the data is recorded in narrative forms, such as clinical notes, natural language processing (NLP) techniques have been used to generate data better used in clinical decision making.

Second, the data is multimodal, originating from a variety of sources with varying standards. This requires expensive and often time-consuming data curation before performing clinically relevant data analysis. Lastly, certain clinical parameters cannot be measured explicitly in the ICU and are therefore simply estimated. An example is a cardiac output where no validated devices are available in the ICU, and estimate thus suffice.

2.1.4 Data Harmonization

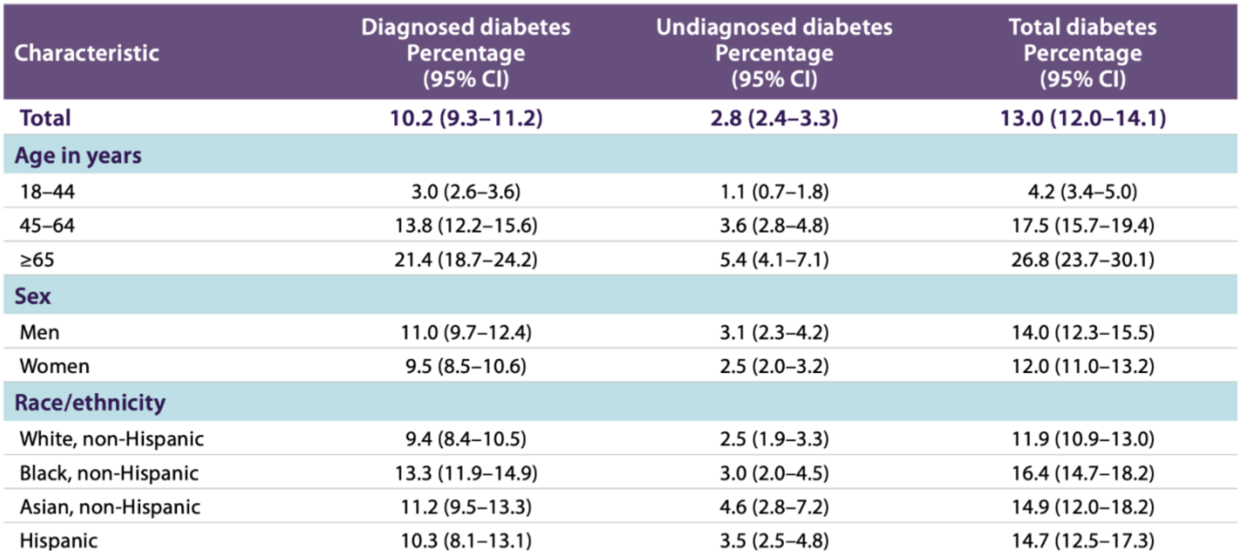
Because of the nature of ICU data, it becomes critical to design ways of harmonizing the data for clinical research and predictive modeling. As previously stated, patient clinical data is sourced and exists in several places. It is also in heterogeneous format and may use alternative or sometimes non-standard terminology. All of this adds to the challenge of conducting clinical research. Researchers would be able to find trends within the data and work with familiar words if the data were harmonized. Additionally, such data would also improve search and retrieval and aid in predictive modeling. In summary, harmonized data would improve the data quality, reusability, and interoperability (Geneviève, Martani, Mallet, Wangmo, & Elger, 2019).

Ideally, ICU clinical data should be circulated across national borders. This would entail the existence of an internationally recognized data standard or model. Currently, no such standard exists that is widely accepted. Should such a standard be developed, there would be a need for data harmonization. In addition to already noted challenges, medical centers have different ways to deliver care. Differences would also exist in how clinical documentation is done due to cultural considerations and how variables are accessed due to technological differences. These challenges would make it quite challenging to harmonize data for international modelling applications (Haendel, Chute, & Robinson, 2018).

2.2 Diabetes

In our study, we use a cohort of diabetic patients to showcase our approach. Diabetes is a disease that occurs when the glucose level in the blood is unusually high. An individual with diabetes has a higher risk of many other long-term diseases such as neuropathy, kidney disease, and cardiovascular diseases. Diabetes affects over 45% of ICU patients aged 65 years or older (Anand et al., 2018). Diabetes has also been associated with ICU bloodstream infections, impacting the wellbeing of critical care patients (Michalia et al., 2009).

Diabetes remains a significant public health challenge both in the United States and globally. According to the CDC’s 2020 National Diabetes Statistics Report, 34.2 million people or 10.5% of the US population, had diabetes. Of these, 31.4 million adults, or 13% of all adults, had diabetes. In 2020, the prevalence of diabetes increased with age, with 26.8% of those aged 65 or older being diabetic. Figure 7 below shows some of these CDC statistics.



Characteristic	Diagnosed diabetes Percentage (95% CI)	Undiagnosed diabetes Percentage (95% CI)	Total diabetes Percentage (95% CI)
Total	**10.2 (9.3–11.2)**	**2.8 (2.4–3.3)**	**13.0 (12.0–14.1)**
Age in years			
18–44	3.0 (2.6–3.6)	1.1 (0.7–1.8)	4.2 (3.4–5.0)
45–64	13.8 (12.2–15.6)	3.6 (2.8–4.8)	17.5 (15.7–19.4)
≥65	21.4 (18.7–24.2)	5.4 (4.1–7.1)	26.8 (23.7–30.1)
Sex			
Men	11.0 (9.7–12.4)	3.1 (2.3–4.2)	14.0 (12.3–15.5)
Women	9.5 (8.5–10.6)	2.5 (2.0–3.2)	12.0 (11.0–13.2)
Race/ethnicity			
White, non-Hispanic	9.4 (8.4–10.5)	2.5 (1.9–3.3)	11.9 (10.9–13.0)
Black, non-Hispanic	13.3 (11.9–14.9)	3.0 (2.0–4.5)	16.4 (14.7–18.2)
Asian, non-Hispanic	11.2 (9.5–13.3)	4.6 (2.8–7.2)	14.9 (12.0–18.2)
Hispanic	10.3 (8.1–13.1)	3.5 (2.5–4.8)	14.7 (12.5–17.3)

Figure 7- Diabetes statistics 2013-2016 (image from cdc.gov)

In 2017, the direct and indirect cost of diagnosed diabetes in the United States totaled $327 billion. A person with diabetes incurred additional medical costs of $9,601. In the same year, diabetes was the 7th leading cause of death, killing over 364,000 people directly or indirectly.

When a diabetic person is diagnosed with other health conditions, their risk of developing complications and having poorer health outcomes increases. When diabetic patients have other comorbid conditions, their risk of developing health complications and worse health outcomes increases. Adults with diabetes have a 50% greater risk of death than adults without diabetes (*[PDF]National Diabetes Statistics Report, 2014 - CDC*, n.d.), (Diabetes, 2001). Several diseases have been associated with diabetes, such as hypertension, stroke, and kidney disease.

According to the Medical Expenditure Panel Survey, most people with diabetes have at least one comorbid chronic disease, and about 40% have at least three (Diabetes, 2001; Piette & Kerr, 2006). These comorbidities have implications on both the management of care and the availability of care. Conditions such as rheumatoid arthritis and depression pose unique challenges such as lifestyle changes and regimen adherence (Ciechanowski, Katon, & Russo, 2000; Diabetes, 2001; Piette & Kerr, 2006). Diabetic patients with comorbidities may experience a financial burden in paying for care for all their conditions, resulting in cost-related medication underuse (Dubois, Chawla, Neslusan, Smith, & Wade, 2000). Clinicians treating diabetic patients may not have enough time to deal with all comorbidities within the office visit time allocated (Gorin, 2014). Therefore, it is important to consider comorbidities in diabetic patients as this affects their health outcomes.

Accurate mortality prediction can assist healthcare administrators, insurance companies, and physicians in making more informed decisions. Mortality prediction can also assist in clinical trial design for new treatments by ensuring that patients with similar baseline are being compared (Gorin, 2014; Kollef & Schuster, 1994). Moreover, mortality prediction can be used to manage healthcare costs by allocating lower-risk patients to less expensive settings (Kollef & Schuster, 1994). Diabetes has a high mortality rate, especially in patients with a high incidence rate of comorbidities, increasing the mortality rate. Thus, effectively predicting the mortality risk of hospitalized diabetic patients and identifying clinical factors that affect mortality risk in diabetic patients is critical for early interventions for diabetes patients. We sought to predict the mortality of diabetes patients with diverse types of comorbidities.

According to the conducted researches targeting the prediction of mortality in diabetes, Diabetes patients have a higher than average mortality rate than non-diabetics(Kollef & Schuster, 1994; Koskinen, Reunanen, Martelin, & Valkonen, 1998). Braun et al. reported that diabetic and stroke patients share the same mortality rate with no significant difference (Braun, Otter, Sandor, Standl, & Schnell, 2012). This could be because improved treatment and early intervention offset a worse prognosis. Diabetic patients with certain conditions such as depression and congestive heart failure have been noted to have higher mortality rates (Katon et al., 2005).

Several studies have been done to predict mortality among type 2 diabetic patients. A Cox proportional hazards regression model utilizes about 20 attributes associated with mortality and medication class in one study. The Cox model coefficients were then used to create a prediction model (Wells et al., 2009). The predictor values included ingestion of aspirin, history of coronary heart disease, diastolic blood pressure, HDL and LDL cholesterol, sex, race, and age, among other predictors. The results of this study were validated by repeated resampling of the model's validation dataset.

In the last few decades, the data collected in ICUs has increased with EMRs and has been subjected to data mining (Ramon et al., 2007/7). Using diverse prediction models, studies have been performed to use EMR data to forecast clinical outcomes such as mortality. Despite the availability of this data, improved mortality prediction for patients in ICU continues to be a challenge. There are variables used in prediction weighted by experts based on their perceived relevance to mortality prediction. Also, some scores are recorded or measured subjectively in a variety of format.

An example may be the categories of admitting diagnosis or recording of pre-existing chronic conditions. Furthermore, some attributes require constant updates at regular intervals. Lastly, some of the variables used in prediction may be extracted directly from the medical records (e.g., age and gender). In contrast, other variables may need to be constructed from the recorded data (e.g., number of existing chronic conditions in a patient).

One study looked at using machine learning tools combined with clinical notes to predict mortality in critically ill diabetic patients (Ye, Yao, Shen, Janarthanam, & Luo, 2020). Unified Medical Language System (UMLS) resources were used together with natural language processing (NLP) and machine learning in this study. The study noted that the use of NLP on clinical notes improved mortality prediction using a convolutional neural network (CNN) model to learn hidden features.

In this chapter, the materials, including clinical data and different mortality and comorbidity measures, are described, followed by introducing the study's novelty.

2.3 Clinical Data, Mortality, Comorbidity Measures

The materials are organized and described in this section, including clinical data and different mortality and comorbidity measures. Detailing the material is critical. In the next chapter, the proposed methodology will be constructed on the concepts presented in the following.

2.3.1 Medical Information Mart for Intensive Care (MIMIC) III

MIMIC-III (Johnson, Pollard, et al., 2016) is a large and single-center database containing de-identified health-related information containing 46,520 patients. Of these, 20,399 are female, and 26,121 are male. Between 2001 and 2012, these patients were admitted to the Beth Israel Deaconess Medical Center's critical care units. The database includes demographics, vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of the hospital). MIMIC-III is accessible to researchers under a user agreement, and it includes de-identified data of the patients. MIMIC-III is a relational database composed of 26 tables. Each patient is uniquely identified with a “subject_id”, and each hospital admission is identified with a unique “hadm_id”. Therefore, a single patient could have several hospital admission ids if they had multiple admissions.

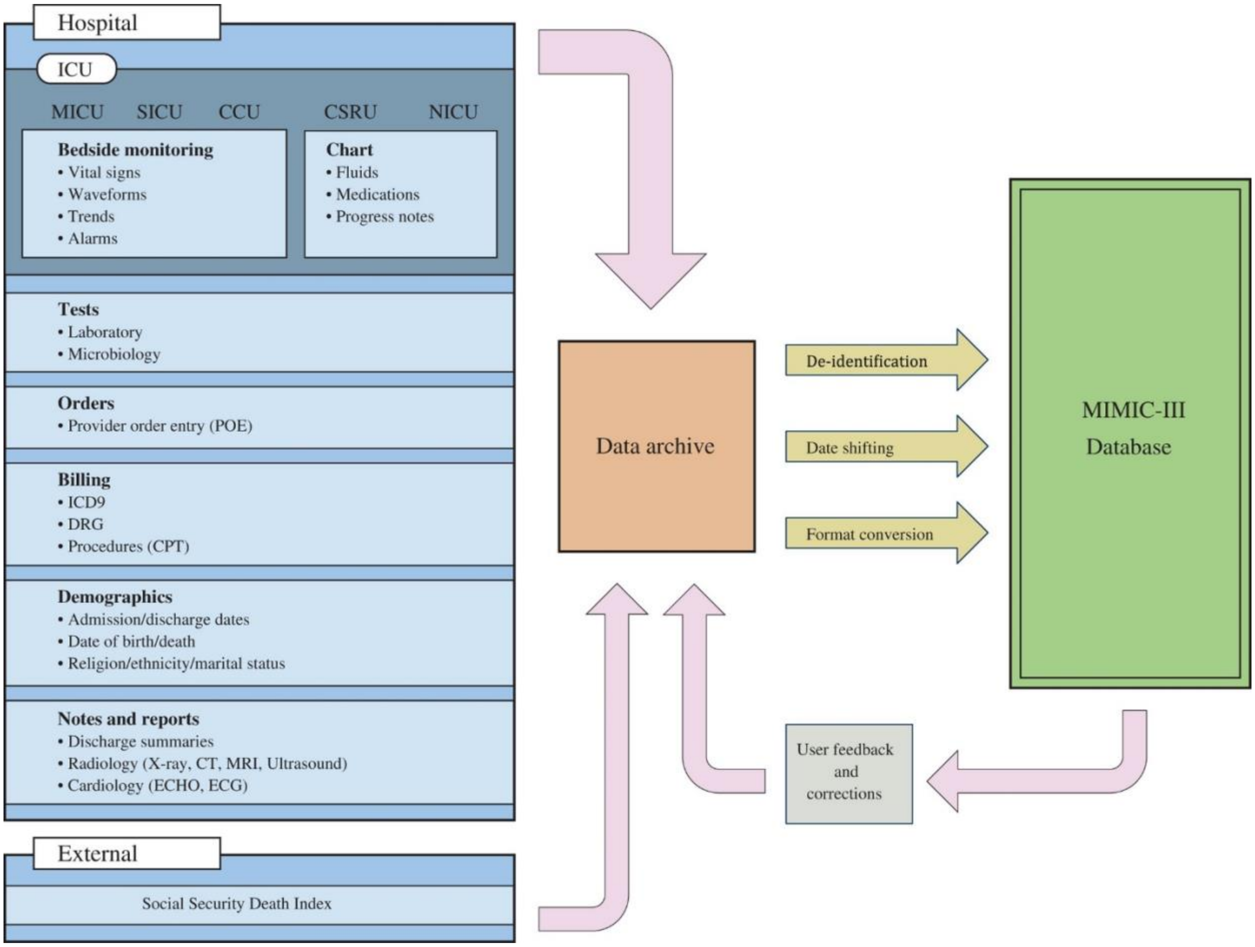
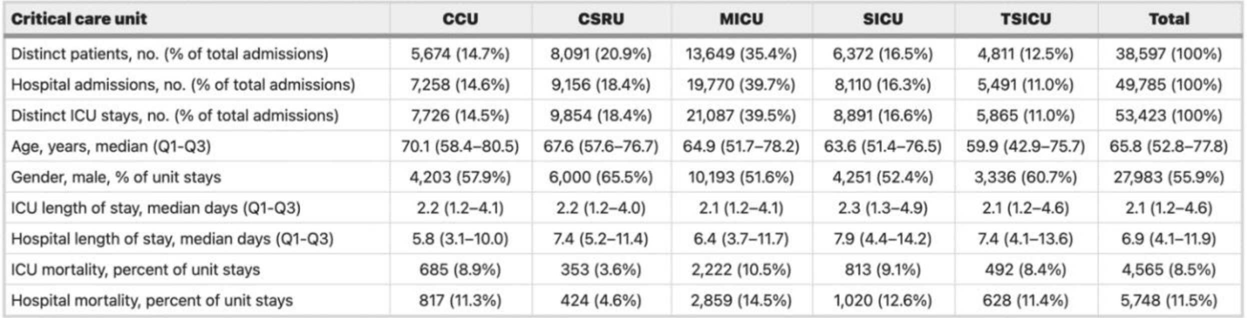


Figure 8- MIMIC-III Database (adopted from Alistair E et al., 2016)

Table 2 below shows the MIMIC-III patient population details for patients aged 16 years and older broken down by the critical care unit. CCU is abbreviated as Coronary Care Unit. CSRU is an abbreviation for Cardiac Surgery Recovery Unit. The MICU is Medical Intensive Care Unit, SICU is a Surgical Intensive Care Unit, and TSICU is Trauma Surgical Intensive Care Unit. The table further breaks down the patient population for each critical care unit by distinct patients, hospital admissions, different ICU stays, and ICU mortality, among other important measures. This analysis provides a broad outline of the ICU patients, facilitating comparisons with the diabetic patient cohort used in this study.

Table 2 – MIMIC-III Patients Aged 16 Years and Older



Critical care unit	CCU	CSRU	MICU	SICU	TSICU	Total
Distinct patients, no. (% of total admissions)	5,674 (14.7%)	8,091 (20.9%)	13,649 (35.4%)	6,372 (16.5%)	4,811 (12.5%)	38,597 (100%)
Hospital admissions, no. (% of total admissions)	7,258 (14.6%)	9,156 (18.4%)	19,770 (39.7%)	8,110 (16.3%)	5,491 (11.0%)	49,785 (100%)
Distinct ICU stays, no. (% of total admissions)	7,726 (14.5%)	9,854 (18.4%)	21,087 (39.5%)	8,891 (16.6%)	5,865 (11.0%)	53,423 (100%)
Age, years, median (Q1-Q3)	70.1 (58.4–80.5)	67.6 (57.6–76.7)	64.9 (51.7–78.2)	63.6 (51.4–76.5)	59.9 (42.9–75.7)	65.8 (52.8–77.8)
Gender, male, % of unit stays	4,203 (57.9%)	6,000 (65.5%)	10,193 (51.6%)	4,251 (52.4%)	3,336 (60.7%)	27,983 (55.9%)
ICU length of stay, median days (Q1-Q3)	2.2 (1.2–4.1)	2.2 (1.2–4.0)	2.1 (1.2–4.1)	2.3 (1.3–4.9)	2.1 (1.2–4.6)	2.1 (1.2–4.6)
Hospital length of stay, median days (Q1-Q3)	5.8 (3.1–10.0)	7.4 (5.2–11.4)	6.4 (3.7–11.7)	7.9 (4.4–14.2)	7.4 (4.1–13.6)	6.9 (4.1–11.9)
ICU mortality, percent of unit stays	685 (8.9%)	353 (3.6%)	2,222 (10.5%)	813 (9.1%)	492 (8.4%)	4,565 (8.5%)
Hospital mortality, percent of unit stays	817 (11.3%)	424 (4.6%)	2,859 (14.5%)	1,020 (12.6%)	628 (11.4%)	5,748 (11.5%)

Table 3 summarizes the MMIC-III data tables. As can be seen, MIMIC-III data is rich with patient data and granular enough to limit data transformation. Most of the data is a close reflection of the raw patient data collected in the hospital (Johnson, Pollard, et al., 2016). The data set is composed of many tables that contain information about various subjects. The tables defining the patients and containing patient stays include the Admission, Patients, ICU Stays, Services, and Transfer tables. Another set of tables contain services relevant to patient care, such as billing and physiological measurements. Lastly, tables are dictionary tables that map codes to their definitions, such as procedure codes and diagnoses codes. A combination of all these tables forms a rich source of data analysis and predictive modeling.

Table 3- MIMIC-III Data Tables

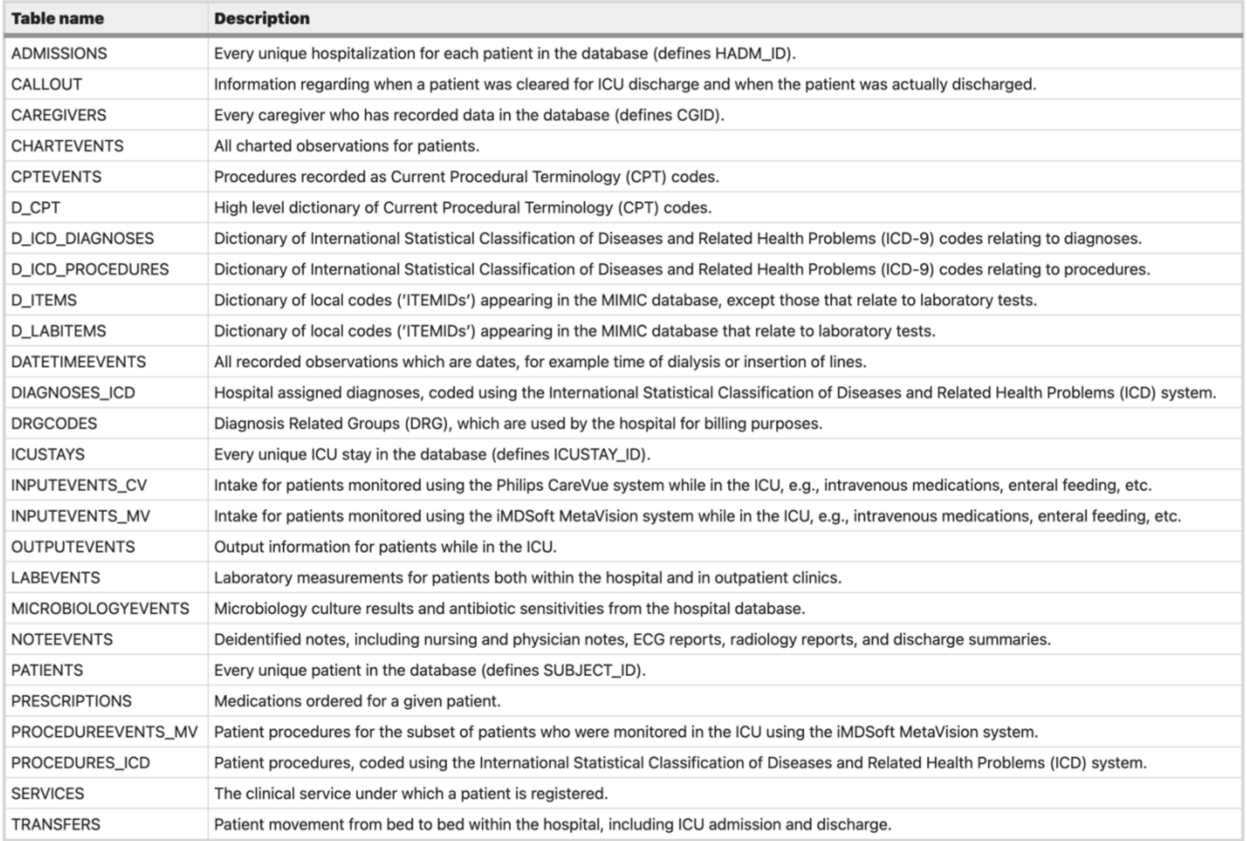


Table name	Description
ADMISSIONS	Every unique hospitalization for each patient in the database (defines HADM_ID).
CALLOUT	Information regarding when a patient was cleared for ICU discharge and when the patient was actually discharged.
CAREGIVERS	Every caregiver who has recorded data in the database (defines CGID).
CHARTEVENTS	All charted observations for patients.
CPTEVENTS	Procedures recorded as Current Procedural Terminology (CPT) codes.
D_CPT	High level dictionary of Current Procedural Terminology (CPT) codes.
D_ICD_DIAGNOSES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes relating to diagnoses.
D_ICD_PROCEDURES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes relating to procedures.
D_ITEMS	Dictionary of local codes ('ITEMIDs') appearing in the MIMIC database, except those that relate to laboratory tests.
D_LABITEMS	Dictionary of local codes ('ITEMIDs') appearing in the MIMIC database that relate to laboratory tests.
DATETIMEEVENTS	All recorded observations which are dates, for example time of dialysis or insertion of lines.
DIAGNOSES_ICD	Hospital assigned diagnoses, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system.
DRGCODES	Diagnosis Related Groups (DRG), which are used by the hospital for billing purposes.
ICUSTAYS	Every unique ICU stay in the database (defines ICUSTAY_ID).
INPUTEVENTS_CV	Intake for patients monitored using the Philips CareVue system while in the ICU, e.g., intravenous medications, enteral feeding, etc.
INPUTEVENTS_MV	Intake for patients monitored using the iMDSoft MetaVision system while in the ICU, e.g., intravenous medications, enteral feeding, etc.
OUTPUTEVENTS	Output information for patients while in the ICU.
LABEVENTS	Laboratory measurements for patients both within the hospital and in outpatient clinics.
MICROBIOLOGYEVENTS	Microbiology culture results and antibiotic sensitivities from the hospital database.
NOTEEVENTS	Deidentified notes, including nursing and physician notes, ECG reports, radiology reports, and discharge summaries.
PATIENTS	Every unique patient in the database (defines SUBJECT_ID).
PRESCRIPTIONS	Medications ordered for a given patient.
PROCEDUREEVENTS_MV	Patient procedures for the subset of patients who were monitored in the ICU using the iMDSoft MetaVision system.
PROCEDURES_ICD	Patient procedures, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system.
SERVICES	The clinical service under which a patient is registered.
TRANSFERS	Patient movement from bed to bed within the hospital, including ICU admission and discharge.

Focusing on the objective of this study, MIMIC-III is assessed to extract diabetes patients for mortality prediction.

2.3.2 Elixhauser Comorbidity Measure

The Elixhauser comorbidity index is used to categorize patients' comorbidities based on conditions described by the ICD-9-CM (International Classification of Diseases, Ninth Edition, Clinical Modifications) discharge records. It consists of 30 comorbidity measures that have been shown to impact patient hospital length of stay, cost, and mortality (Elixhauser et al., 1998). The ICD coding is dichotomous – therefore, the disease is either present or absent. It has been successfully applied to predict patient mortality and hospitalization (Elixhauser et al., 1998; Quail, Lix, Osman, & Teare, 2011).

The Elixhauser index assigns a value of 1 for each of 30 predefined conditions and computes the sum. Therefore, a score of 10 indicates a patient currently has 10 conditions. Several experiments have been conducted to determine the Elixhauser index's ability to accurately assess severity compared to other approaches such as the Charlson index. . The Elixhauser index made better mortality predictions for hospitalized patients (Menendez, Valentin, van Dijk, & David, 2014). Thus, in this study, the Elixhauser index is applied for mortality prediction.

2.3.3 Mortality Scoring Systems

Several ICU scoring systems have been developed. Three of the more common ones include the Acute Physiology and Chronic Health Evaluation (APACHE) (Knaus, Draper, Wagner, & Zimmerman, 1985), the Simplified Acute Physiology Score (SAPS) (Le Gall, Lemeshow, & Saulnier, 1993), and Sequential Organ Failure Assessment (SOFA) (Jones, Trzeciak, & Kline, 2009). Both SAPS and APACHE consider patient clinical variables obtained within the first 24 hours of ICU admission to determine disease severity. Models have been tailored for specific geographic areas such as the Mediterranean countries, Western Europe, and Southern Europe. Also, the scoring systems have now evolved and been improved over time.

From a study of 13,152 patients, a new Simplified Acute Physiology Score (SAPS II) was created to assess patient severity and transform the score to a probability of an in-hospital mortality prediction (Le Gall et al., 1993). SAPS II was developed using critical care adult (18 years and over) patients in 12 countries. It includes 17 variables, of which 12 are physiological. The others are age, type of admission, and three underlying disease variables. Appendix D shows the variables and score assignments used in SAP II.

APACHE II is another tool that provides a classification system for the severity of the disease. It is determined using 12 physiological measurements, the patient's age, and previous health status. It is a point-based system with a score from 0 to 71 depending on the computed measurements. Higher scores indicate a higher indicates a greater risk of death. Variables not measured are given a score of zero. Patient mortality prediction is then computed using a logistic regression equation based on the calculated APACHE II score, the patient diagnostic category, and presence of emergency surgery (Knaus et al., 1985).

Our research concentrates on the SOFA scoring system, which incorporates six organ systems in its scoring. SOFA is a tool used for estimating and scoring morbidity (Jones et al., 2009). It is developed from a large sample of ICU patients worldwide and has been successfully applied in various studies (Ceriani et al., 2003; Jones et al., 2009). The SOFA score is computed using six variables, with each variable corresponding to an organ system. The organ systems included in SOFA are respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems. Each organ system is assigned a score ranging from 1 to 4, depending on its severity. The SOFA score is calculated as the sum of the values for each organ system. Table 4 reports how the SOFA score is computed. For example, for respiration, the Carrico index (Vincent et al., 1996) assigns points. The Carrico index compares the oxygen (O_2) level in the blood, and the O_2 concentration breathed in. The purpose of the Carrico index (calculated as PaO_2/FiO_2) is to determine the extent of any possible problems in how the lungs transfer O_2 to the blood. PaO_2 is the partial pressure of O_2 in arterial blood, and FiO_2 is an indicator of hypoxemia. Suppose the Carrico index is 150 and the patient is under respiratory support. In that case, the respiratory point score for that patient is 3. Each organ system has a formula that describes how the points are computed, as shown in Table 4 below.

Table 4- SOFA organ areas and points thresholds

Organ/Points	1	2	3	4
Respiratory PaO_2/FiO_2 (mmHg)	< 400	< 300	< 200 and respiratory support	< 100 and respiratory support
Cardiological Mean arterial pressure	MAP < 70 mm/Hg	dop ≤ 5 or dob	dop > 5 OR epi ≤ 0.1 OR nor ≤ 0.1	dop > 15 OR epi > 0.1 OR nor > 0.1
Renal Creatinine (mg/dl) [µmol/L] (or urine output)	1.2 – 1.9 [110 – 170]	2.0 – 3.4 [171 – 299]	3.5 – 4.9 [300 – 440] (or < 500 ml/d)	> 5.0 [> 440] (or < 200 ml/d)
Hepatic Bilirubin (mg/dl) [µmol/L]	1.2–1.9 [> 20 – 32]	2.0–5.9 [33 – 101]	6.0–11.9 [102 – 204]	> 12.0 [> 204]

Neurological Glasgow coma scale	13 – 14	10 – 12	6 – 9	< 6
Coagulation Platelets×103/µl	< 150	< 100	< 50	< 20

SOFA is sufficiently reliable. Studies have shown that SOFA is comparable to other severity scoring methods in mortality prediction (Minne, Abu-Hanna, & de Jonge, 2008). Although SOFA is primarily a morbidity score, some studies have shown that SOFA scores can be used to predict mortality (Ferreira, Bota, Bross, Mélot, & Vincent, 2001; Minne et al., 2008).

Table 5 Section presents SOFA ratings and their associated anticipated mortality rates.

Table 5- SOFA score mortality

SOFA Score	**Mortality**
0 to 6	< 10%
7 to 9	15 - 20%
10 to 12	40 - 50%
13 to 14	50 - 60%
15	> 80%
16 to 24	> 90%

In healthcare, survival analysis has been used to predict the time interval between two events, such as death or disease diagnosis. Several methods can be used to conduct this analysis. One of the frequently used is the Kaplan-Meier method, which uses observed survival times to estimate the survival probability of an event. The Kaplan-Meier curve represents the probability of survival and is useful in showing changes in disease-related mortality. This probability of survival $P(t_j)$ at time t_j is computed with the equation below.

$$P(t_j) = P(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$$

In this equation, n_j is the number of patients alive before time t_j, and d_j is the number of deaths at time t_j.

2.4 Comparison of Existing ICU Prediction Models

Over the last three decades, considerable efforts have been made to model the probability of death in ICU patients, and numerous severity scores have been developed. Concerning the studies assessing the validity of ICU mortality prediction, some of the more popular tools include Acute Physiology and Health Evaluation (APACHE) (Knaus et al., 1985), Simplified Acute Physiology Score (SAPS) (Le Gall et al., 1993), Mortality Probability Model (MPM) (Lemeshow et al., 1993), and SOFA (Vincent et al., 1998); (Arabi, Shirawi, Memish, Venkatesh, & Al-Shimemeri, 2003).

These tools compute severity scores obtained from baseline patient characteristics. Generally, these are the measurements recorded within 24 hours of admission for ICU patients. Among the used scores, APACHE and SAPS are based on variables assigned subjectively by a panel of experts. Both APACHE and SAPS are the most commonly used mortality prediction tools, but they are not adequately calibrated (Le Gall et al., 1993; Pirracchio et al., 2015). Most of these prediction tools use parametric methods, thus imposing assumptions on the relationship between mortality and the mortality predictor variables. APACHE and SAPS rely on subjective methods of variable selection. Still, subsequent models such as SAPS II rely on statistical modeling techniques for variable selection.

Several studies have been done to evaluate ICU predictive systems (Kollef & Schuster, 1994; Kramer, Higgins, & Zimmerman, 2014; Kuzniewicz et al., 2008; Minne et al., 2008; Nassar et al., 2012). However, major randomized studies demonstrating the superiority of one system compared to another are lacking (Juneja, Singh, Nasa, & Dang, 2012; Keegan, Gajic, & Afessa, 2011). Nevertheless, due to the intricate clinical interactions that influence ICU mortality, these assumptions can be interpreted differently depending on the patient's severity.

A comparative study of APACHE II, SAPS, SOFA, and Cardiac Surgery Score (CASUS) revealed that these scoring systems were not well suited for predicting mortality in patients undergoing cardiac surgery (Doerr et al., 2011). However, this study concluded that SOFA was a reliable ICU mortality risk stratification model for cardiac surgery. Also, both SAP and APACHE had undesirable calibration and discrimination statistics.

2.5 Ensemble ICU Prediction Models

Extensive research has been conducted to enhance ICU mortality prediction by using an ensemble of prediction models. The advantage is that prediction outcomes are determined using the best member of the ensemble (in terms of the confidence score) (Pirracchio et al., 2015). This is important because the diversity of patients and medical conditions make it challenging for a single model to perform optimally throughout. Awad et al. has reported improved ICU mortality prediction using a combination of ensemble learning RF, DT, NB, and the rule-based Projective Adaptive Resonance Theory (PART) models (Awad, Bader-El-Den, McNicholas, & Briggs, 2017). Another research demonstrated that an ensemble of long-short term memory (LSTM) networks outperform a random forest classifier and a single LSTM network (Xia et al., 2019).

Ensemble modeling contributes to the reduction of generalization error. Generalization error quantifies the prediction accuracy of a model with unseen data. If the base models used in ensemble modeling are independent, the prediction error is reduced. In regression analysis, a low prediction error indicates that the response variables are correctly predicted. A low prediction error indicates a more accurate placement of samples in the correct category in classification problems. The clinical prediction uses both regression and classification type models.

2.6 Focus on ICU Risk Factors

Numerous research efforts have been directed toward developing ways of characterizing risk factors to improve ICU mortality prediction. The objective is to determine which ICU risk factors are associated with mortality (Begum, 2017). Recently, with the increase in biomarkers, it has become more important to study how biomarkers influence clinical outcomes, including mortality. Yet another research was conducted to identify risk mortality risk factors that could be modified for ICU patients (Chen et al., 2001). Another study aimed to identify ICU mortality risk factors independent of the APACHE score showed no substantial improvement in the prediction model's discriminative prediction ability (Li et al., 2016).

Mukhopadhyay et al. discovered that age is the largest risk factor for mortality in ICU patients, followed by comorbidities, malignancy, and renal replacement treatment (RRT) before hospital admission (Mukhopadhyay et al., 2014). To better understand risk factors and improve prediction model features, Lehman et al. used topic models to ICU patient risk stratification in combination with SAPS-1 (Lehman, Saeed, Long, Lee, & Mark, 2012). They recorded improved prediction accuracy.

2.7 Personalized ICU Mortality Prediction

To improve the accuracy of personalized outcome predictions for individual patients, studies have been carried out to develop prediction models based on the similarity of patients. This research has incorporated similar patient cohorts in training mortality prediction models that can then be applied to similar clusters of patients. Lee et al. developed mortality prediction models using a cosine-similarity-based patient similarity matrix (PSM) (Lee, Maslove, & Dubin, 2015).

This study discovered that smaller training sets of similar patients predicted mortality more accurately than using all available data from the EMR. Another analysis on patient similarity used an index patient (on whom the prediction was being made) to create models based on patients like the index patient (N. Wang et al., 2019). Thus, while using fewer data to train the models, outcome prediction was improved.

The increased collection of patient data enables the development of multiple cohorts of patients from a population using specified biomarkers and other attributes. This then allows for detecting signals indicating correlations or clusters, which can then be separated and analyzed further. For instance, patients who share a known identified mutation and a defined demographic profile can be filtered out and used to construct a training set for such as cohort.

Additionally, intra- and inter-cohort research were delineated as part of a customized mortality prediction application. The intra-cohort analysis examines variations and other characteristics within a defined patient cohort that has been labeled as similar based on identified features. At This Point, a common similarity baseline will be equally applied to all patients while focusing on calibrating the differences among patients within this baseline. The inter-cohort analysis involves the differentiators among distinct patient cohorts. Through inter-cohort analysis, the similarity functions are exposed and validated.

2.8 Summary of different methodologies

As shown by our investigations, several methods have been used to enhance ICU mortality prediction, indicating that this is an active research area. Our descriptions are not exhaustive, but they illustrate the landscape of mortality prediction. Our descriptions are not exhaustive, but they illustrate the landscape of mortality prediction.

A commonly used approach uses several published prediction models to determine which one performs best in each clinical scenario. This is analogous to a brute-force approach where the information is sought on existing prediction tools and decisions are taken in accordance with the tool's intended objective. In this case, several tools would exist targeting various patient populations such as cardiology or cancer patients. In some instances, the calibration would be fine enough to align with more granular patient populations, such as lung cancer patients fitting specified demography. This is not an optimal solution since it is neither versatile nor generalizable.

Another common approach uses a model that encompasses an ensemble of various classifiers to get the best result. This approach advantageous since it avoids prediction and generalizable errors. Depending on the base models used in ensemble modeling, good prediction accuracies can be achieved. However, an ensemble modeling approach does not account for patient similarity and is limited in personalized predictions.

Another approach relies heavily on defining and augmenting features to a model to improve its outcome prediction accuracy. While it is common to continuously interrogate features as domain awareness is extended, this approach emphasized using any available clinical data to enhance prediction accuracy. This takes advantage of the ability to collect and store many clinical attributes. As technology advances, more biomarkers are being discovered and recorded. These can be used as attributes in the prediction models. A drawback to this approach is the lack of personalized prediction and scale flexibility.

Using new techniques such as machine learning (ML) to learn and optimize prediction, several approaches have embraced techniques such as patient similarity to improve cohort selection. As we have noted, the data in electronic medical records can classify patients into groups based on similarity. These groups then form the training data that is used to train prediction models. Therefore, the results are more individualized for the specific groups instead of training the models using all the available clinical data.

Natural language processing (NLP) and topic modeling improve on existing mortality prediction models. This is still an active and promising research area with the advantage of augmenting existing models with additional predictive attributes. NLP and topic modelling can be used to identify themes in some clinical data, such as clinical notes. These themes can then be a basis for stratifying patients into various categories based on similarity. A disadvantage of this method is that not all patients will have sufficient narrative data to benefit from natural language processing and topic modeling. In addition, diverse and technical clinical corpus presents a challenge of using such attributes in machine learning models.

Chapter 3

Research Approach and Novel Methodology

Having a flexible approach that allows for ICU mortality prediction is important. Since critical care patients present with a diverse set of conditions, a generalizable prediction framework can cater to a wider scope of patients and endpoints. We have noted in this study that most of the existing ICU mortality prediction solutions are not well suited for diverse patient populations or calibrated for different endpoints. This limitation renders them not scalable or transferable across the clinical landscape. In this chapter, the methodology of the current research is described. This work's main idea incorporates ML's employment to propose a hybrid approach of predicting patient mortality based on existing mortality prediction scores and patient comorbidity. This includes performing a retrospective study on a cohort of critical care diabetic patients. As explained in this chapter, our work fills the gap between existing ICU prediction models and a generalizable flexible prediction model.

3.1 Proposed Approach

We provide a method of mortality prediction that utilizes existing patient information, domain knowledge, and machine learning technics. Our novel method provides a foundation framework for various clinical scenarios to predict an outcome. We believe it is important to demonstrate our approach by selecting diabetes as a common clinical condition and showing our ability to predict mortality using a cohort analysis of diabetic patients admitted to the ICU for any reason including other non-diabetic conditions. Diabetes was selected because it is a well-known and understood clinical condition. Matured treatment plans are available for diabetic patients. The symptoms are well known, and the comorbidities associated with them have been thoroughly studied.

Our approach lays the groundwork for the prediction tool and enables clinical informaticians to customize it as appropriate. The following are the big pieces of our approach, and these subsequent subsections will address the major components of our strategy in detail:

- Identification of patient population to perform mortality prediction
- Collection of clinical data attributes from the patients to train the prediction model
- Identification of patient comorbidities for selected patient cohort

- Use of ranking algorithms to assign predictor weights and selection of comorbidity model attributes
- Computation of SOFA scores based on patient data
- Use of machine learning algorithms to build a prediction model

Our approach works on various supervised machine learning algorithms such as neural networks, Support Vector Machines, decision trees, and others. Due to the complexities and nature of clinical data, we adhere to the No Free Lunch theorem while utilizing learning algorithms to optimize our prediction results. Therefore, although we used two supervised learning algorithms in this analysis, any other learning algorithm may be used. We based our choice of machine learning algorithms on experimentation with our dataset and attributes. The flexibility to use existing machine learning techniques allows our approach to be versatile and customizable as needed.

Our system input is a labeled dataset of diabetic patients, a machine learning model, computed SOFA scores, and a list of ranked comorbidities. Following that, our system uses these inputs to predict mortality and output the results. As such, we implement the data curation process, model selection, SOFA scores computation, and weighted comorbidity list generation. These steps are one-time measures for the dataset. They are not required to be replicated when using a different machine learning prediction model.

These steps are manual but can be automated with scripts. The novelty of our method is not the automation, but automation would save on time while preserving the prediction accuracy. Automation can allow notifications to happen as well. The subsections below guide this process.

3.2 Patient Cohort Retrieval

In this study, The MIMIC III database was queried for a diabetes patient cohort using the International Classification of Diseases, ninth revision, Clinical Modification (ICD-9) codes for diabetes. The MIMIC III database contains ICU patient data from Israel Deaconess Medical Center. Since the database contains various patients, we use an SQL query to retrieve the diabetic patients. We then use this result to generate a database view that we can use for our continued processing. The ICD-9 codes we used in this study are shown in Table 6

Table 6- ICD-9 codes applied for diabetic patient retrieval

ICD-9 codes	**Description**
250.00 – 250.33	Diabetes without chronic complications
648.00 – 648.04	Diabetes without chronic complications
249.00 – 249.31	Diabetes without chronic complications
250.40 – 250.93	Diabetes with chronic complications
249.40 – 249.91	Diabetes with chronic complications

To further predict mortality in diabetes patients with multiple comorbidities, we created an SQL view that includes the entire diagnosis history after diagnosing diabetes. Vital signs, laboratory tests and demographics, among others, are some of the information queried include.

We have described above the selection of a predefined group of patients – in this case, diabetic patients from our dataset. Another possibility is clubbed groups, in which we combine cohorts from multiple predefined groups. This approach is advantageous for zooming in or out of certain cohorts. For instance, we can have a clubbed group comprising of all cancer patients. This clubbed cohort would be a superset of individual cohorts of specific cancer diagnosis such as breast, lung, prostate, etc. This approach of clubbed groups adds an aspect of patient similarity because similar groups can contribute to the versatility of patient similarity. An extension of patient retrieval is that ad-hoc groups can compare prediction performance to clubbed or predefined groups. Ad-hoc groups can be obtained from the dataset using any user-defined criteria. They can be deliberately chosen from the dataset or generated using a search query or query mask that contains a set of queries. Ad-hoc groups are analogous to a control group in that they can be used for various user-defined purposes.

Regardless of the criteria used to retrieve the patients, it is helpful to define a method of storing the selected cohort. In our study, data is stored in an Oracle Database, which is acceptable given our data's nature. As patient data becomes more complex or increases in several orders of magnitude, other storage approaches can be considered. A good option would be the use of distributed storage system such as the Apache Hadoop File Distributed System (HFDS).

3.3 Comorbidity Measure Calculation

Comorbidity is a significant predictor of our system. It has been used to predict clinical outcomes in different ways. Our analysis aimed to efficiently capture the predictive abilities of the comorbidities in our patient cohort. Diabetic patients often have multiple co-morbid conditions, and these impact their clinical outcomes. Not all these conditions are equally weighted in their impact on mortality. Therefore, we needed to design a method that would allow our prediction model to account for the impact of comorbidities while considering the relative weights on mortality.

A critical aspect of capturing comorbidity is defining its representation. Several studies have been done exploring various ways in which comorbidities are represented in clinical research. Individual comorbidities are counted and given either similar or different weights in some approaches. While others have used weighted summarized measures such as the Charlson Comorbidity index. Using summarized measures combined with individual comorbidities is another hybrid system. These approaches have been shown to have a negligible effect on test results. For our study, we chose a summarized measure, the Elixhauser index, based on extensive literature review and in consideration of study goals. However, we note that the Charlson Comorbidity index is an outpatient predictor and The Elixhauser index looks at inpatient mortality, and therefore n

The approach we implemented included computing the comorbidity scores for all the patients in our cohort. The Elixhauser index includes 30 comorbidity indicators. These are based on
We used ICD-9 codes to implement a coding scheme to identify the comorbidities (Quan et al., 2005). Several studies have shown that the Elixhauser index has been adapted for specific patient populations (Yande, Gohil, & Johnson, 2020) (Mehta et al., 2018). Our approach sought to make our system as generic as possible to allow extensibility across various conditions and endpoints. We, therefore, used the unmodified Elixhauser index without modification. Previous research has

shown that the Elixhauser index outperforms the Charlson index in several scenarios (Menendez, Neuhaus, van Dijk, & Ring, 2014).

One consideration we had was to avoid using all 29 comorbidity indicators in our model, given efficiency considerations. In addition, studies have shown that different diseases impact patient mortality differently. We were, therefore, careful to capture the impact of the comorbidities while being efficient on the modeling considerations. We opted not to use a composite score of all comorbidities since they each have different mortality predictive values. Instead, we generated the comorbidities for each patient and ranked them depending on their mortality predictive weights.

As noted, the Elixhauser index was used to measure the severity of comorbidity. This is as opposed to other approaches such as counting the individual comorbidities. Studies have shown that using comorbidity summary measures such as Elixhauser is appropriate for prognostic or adjustments in mortality analysis (Austin, Wong, Uzzo, Beck, & Egleston, 2015). 29 different diagnoses were extracted and quantified for each diabetic patient using the Elixhauser index. Finally, all Elixhauser index values were added to account for comorbidities prevalent in the diabetic population.

3.4 Most Relevant Comorbidity Identification

With the relatively large number of comorbidities represented in the Elixhauser index, it is critical to define a process for selecting the most relevant comorbidities for our selected patient cohort. Numerous approaches have been proposed and evaluated for feature selection in problems such as ours. Some of them include using machine learning techniques such as artificial neuro networks (ANN) to identify predictive modeling features (Steinmeyer & Wiese, 2020). While these are rational approaches, we favored a simpler technique that produced satisfactory results while allowing our system to remain generalizable. In addition, we wanted to avoid the potential for overfitting in our model, which can occur when many features are used, some with minimal predictive value.

We discovered that we could have a subset of the Elixhauser index conditions through experimentation and literature review and still obtain robust model features. We then needed to determine the right number of comorbidities to achieve an optimal balance of comorbidities and mortality predictive weights. We obtained this number (five) by repeatedly running our model with a range of different values and observing t the predictive outcomes. The next step was to develop a system that would allow us to select the five comorbidities.

To improve the efficiency and generalizability of our classifier by reducing the number of features, we identified the top five comorbidities most predictive of mortality across the identified diabetic patients. To accomplish this, an ensemble approach of five different feature selection algorithms was experimented, which are Gain Ratio (Karegowda, Manjunath, & Jayaram, 2010), Correlation (M. A. Hall & Smith, 1999), Symmetrical Uncertainty (Yu & Liu, 2003), Information Gain (Jia, Dai, Pan, & Zhu, 2006), and Correlation Feature Selection (CFS) Subset Evaluator (Mark A. Hall & Smith, 1998). Figure 9 d the conceptual view of this process.

We chose five ranking algorithms to avoid ranking bias and account for the diversity of our data. We chose five ranking algorithms to avoid ranking bias and account for the dataset we use and the algorithmic techniques we employ. Consequently, the ranking algorithms we decided on can be added or replaced based on user preferences and the nature of patient data. The specific algorithms we chose to emphasize different techniques based on different aspects of the data. This way, bias is avoided by taking advantage of the "wisdom of the crowd" heuristic.

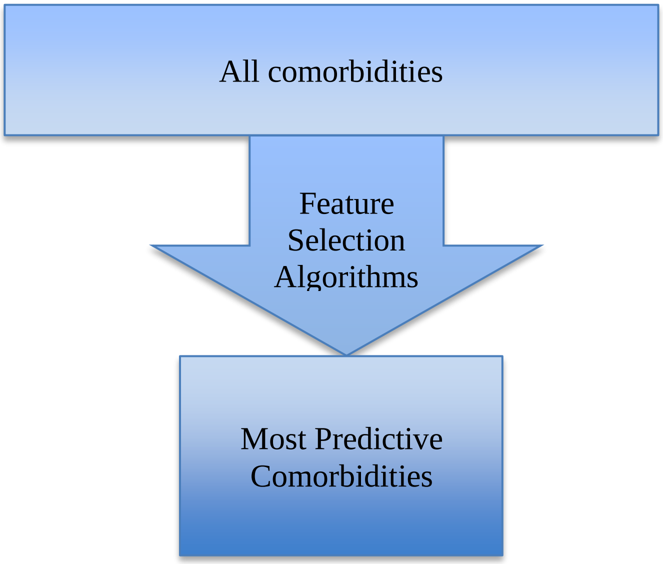


Figure 9 – Most predictive comorbidities

Our comorbidity ranking was determined using the retrieved diabetic cohort. To extend this approach, it is possible to rank the comorbidities based on the entire dataset in the MIMIC-III. Various resampling techniques may be employed to determine which training dataset to use. This approach would yield more generalized results on how comorbidities rank for all patients in critical care. This result would serve as a useful counterpoint to a ranking focused solely on the category we are making our prediction. In our experimentation, we noticed better results by using a ranked list from our selected cohort.

An extension to this approach would be to use different groups (ad-hoc, clubbed, or predefined) to generate graded lists of ranked comorbidities for each group. This information could then be used to better understand the effect of comorbidities on mortality or other clinical endpoints. This may also be useful in getting modeling features for a family of associated conditions or very narrowed-down diseases.

For each algorithm, the comorbidities were ranked based on the resulting predictive mortality. Our final top five most predictive comorbidities were: cardiac arrhythmias, coagulopathy, metastatic cancer, congestive heart failure, and fluid electrolyte. Our final ranked results were then used as an input to our system due to the fact that the final list was generated from an averaged longer list; we gave equal weights to the five comorbidities.

3.5 SOFA Score Computation

We computed the SOFA scores for each of our patients in the patient cohort to determine their SOFA score. This score was then used as input into the mortality prediction model. The SOFA score is a severity indicator based on six major systems in the human body. These systems encompass a diverse range of biomarkers that are indicators of patient mortality. We computed the SOFA scores based on the clinical data in MIMIC-III. We obtained the results to input into the model.

Biomarkers can be used meaningfully in clinical practice as they are used to predict clinical outcomes. Although biomarker discovery has continued, there remains a significant gap between their discovery and use. Our study noticed that the data included in SOFA scoring includes important markers collected early in ICU admission (Selleck et al., 2017). Our hypothesis was that using this information as predictors in a machine learning model would capture the biomarker data in our patient cohort. SOFA captures biomarkers spanning the respiratory system, coagulation, liver, cardiovascular system, kidney, and central nervous system. We were able to show that this information impacts the accuracy of mortality prediction is used in machine learning models.

Incorporating additional biomarkers not currently used in severity scoring or mortality prediction tools will expand our system's biomarker usage. This would include genome data that could have implications on data storage and analysis because of the enormous amount of data captured. A challenge would be identifying actional variants among all the variants discovered. This is an effort that goes together with domain knowledge and continues technological breakthroughs. We anticipate that such an extension in our model would improve its performance.

A challenge that exists in the use of biomarkers is the availability of the data. Decisions regarding critical care patients must be made quickly and early after admission. This may not be possible for many biomarkers, especially those that need lab processing for additional verification. However, better tools are being developed to improve diagnostics and measurements. Another important question is whether there is a need that the biomarkers can meet in terms of predicting outcome. This is not usually an easy question because clinical knowledge may not always be available for new biomarkers. Machine learning techniques using unsupervised can then be used to discover patterns or clusters.

3.6 Summary of Predictor Variables

The predictor variables we used an input into our system are summarized below:

- Patient cohort dataset with all the accompanying clinical data for each patient
- A list of the top five comorbidities based on their mortality prediction weights
- SOFA scores for all patients in our cohort dataset for six organ systems

We captured what we needed to ensure that our model exhibited the three goals we had set to achieve with these inputs. These include being generalizable, personal, and accurate.

- Personal: Cohort selection and comorbidity information allowed for more personalized predictions
- Accurate: Improved on an existing tool by adding important predictors
- Generalizable: Our method uses existing patient data and machine learning tools, and its constituent parts can be user-customized.

3.7 Mortality prediction

Our study's primary outcome measure is in-hospital mortality prediction diabetic patients in the ICU. The term "in-patient mortality" refers to any death during hospitalization, regardless of the cause. The variable collected was the patient identifier which we could use to get all the other patient attributes in the patient record, such as laboratory test data, demographics, and comorbidities.

It is important to make sure that we use the appropriate resources for prediction. In our study, we experimented with multiple machine learning algorithms and selected two to present. We believe that our prediction framework holds and is not dependent on any one algorithm. By the time we come to train and test the model, we have already interrogated all the attributes and selected the ones to use in the model. However, having several algorithms to train the model is a best practice consistent with the “no free lunch” theorem.

We chose Random Forest (RF) and Naive Bayes (NB) . RF is an ensemble method that builds decision trees in parallel, avoiding prediction errors using a single decision tree. This is accomplished by testing decision trees on samples of the dataset and then averaging the results to improve accuracy and avoid overfitting. Naïve Bayes works on the Bayes theorem. Therefore, it assumes independence between the features. The features we selected for our model are rational from a clinical standpoint. One advantage of NB for clinical applications is that it can be used for multi-class classification problems.

To train our mortality prediction model, NBs and RF were chosen with 10-fold cross-validation. Three different classifier models were generated by using distinct predictive features. First, six SOFA features representing the six organ systems were applied, followed by the second step, encompassing the top five ranked comorbidities. Third, the comorbidities and the SOFA features (11 total features) were combined, as shown in Table 7. TheAUC for both the NB and RF was obtained for both of these models to measure their ability to distinguish the patients who died against those still alive.

Table 7 – Prediction Models Used

Model	**Number of Attributes**	**Source**
SOFA	6	SOFA captured biomarkers
Comorbidities	5	Ranked Comorbidities
SOFA and Comorbidities	11	

3.8 Challenges Addressed

The following challenges were addressed in our study

- A hybrid approach of mortality prediction incorporates both existing patient severity scoring and novel attributes. This allows for the repurposing of existing tools to take advantage of machine learning and other advancements in technology.
- Expanding the use of biomarkers in machine learning prediction modeling. As more biomarkers are being discovered, the gap continues to expand between the number of

biomarkers and their use in clinical analysis. Our study allows for more rapid incorporation of biomarkers in clinical prediction modeling.

- Capturing patient similarity in mortality prediction modeling and thus making it more individualized. This we can achieve by training models based on patient cohort like those we want to predict outcomes on. The patient similarity can be defined as broadly or narrowly as needed to achieve this.
- Overall, developing a framework for mortality prediction that is generalizable and flexible. Many of the existing ICU mortality prediction tools are stringently defined. This inhibits their ability to be modified for use in applications other than those they have been designed for.

We have come a long in terms of mortality prediction. Our work continues to push the envelope when it comes to using available clinical data to predict outcomes. This is particularly valid as we face a period where the volume and velocity of clinical data are increasing exponentially. The general field of clinical informatics is catching up to technical innovations.

3.9 Validation and Testing

It was important to test our approach to make sure that the framework functions properly. This is a challenging task since there are several components of our methodology to test. Our goal of testing is to make sure that our system is generic enough to be tailored as needed and extensible. Ideally, an ICU mortality prediction tool should be flexible enough for different diseases or groups of diseases and different endpoints. This is a challenging gap in current prediction tools and a challenge we addressed in our approach. Throughout our process and decision making, we made a point of challenging all assumptions and interrogating any decisions we made. We did part of what we did to collaborate various claims or findings either by experimentation with our data or citing published articles. The narrative below details some of the validations conducted during this report.

Our dataset was validated against publicly available SOFA severity scores. This was done to verify that SOFA scoring functioned correctly with our data. This task was accomplished by running the SOFA tool against our patient cohort and verifying that our mortality count was consistent with SOFA's prediction. We acknowledged that the findings were largely consistent with expectations. . We then hypothesized about the minimal variations we observed. To augment our experimental findings, we reviewed the literature on work that was done on SOFA scoring validation (Huerta et al., 2018) (Y. Zhang, Luo, Wang, Zheng, & Ooi, 2020).

We validated the use of our comorbidity attributes for modeling by running repeated experiments using various numbers of comorbidities. We settled on our number by striking a balance between efficiency in modeling and attribute predictor weights. To increase the versatility of our approach, we executed an ensemble feature selection method that avoids algorithmic bias or dataset-based variance. Through this approach, we validated our selections by noticing a general overlap of similar comorbidities across the selected algorithms inside the ensemble. Again, we also reviewed literature that showed studies done on diabetes comorbidities and their impact on mortality (Dworzynski et al., 2020). We verified that our selected comorbidities were consistent with the reviewed literature.

We validated our prediction results by building several predictive models and comparing the resulting predictions for accuracy. We used 10-fold cross-validation to train the model and tested it. The use of cross-validation inherently lessens the possibility of selection bias. To determine the stability of our results, we ran 5 repeats of our 10-fold cross-validation to observe the standard deviation of the results.

We retrieved data and selected cohort based on previously published ICD-9 codes. We stored all our MIMIC-III data in PostgreSQL. We used scrips made available by the data owners (Johnson, Pollard, et al., 2016) and verified a clean load. We validated the findings of our retrieval data by verification scrips and consistency with other published sources.

Chapter 4
Results and Discussion

In this study, we developed and implemented a hybrid machine learning model to predict ICU mortality for a cohort of diabetic patients. We developed machine learning tools to create a prediction model based on patient clinical data and domain awareness. We demonstrated that we can implement a hybrid mortality prediction system using readily available resources and data while enhancing accuracy on current systems.

We obtained comorbidity and mortality scores for a patient using the SOFA severity score to train ML models for mortality prediction. Patient comorbidities were rated according to their mortality predictor likelihood. This ranking was used to select the model features used. The prediction model based on a combination of SOFA scores and patient comorbidity was more accurate than using SOFA scores alone. The subsections below illustrate and explain our results in greater detail.

4.1 Results for Diabetic Patient Cohort Retrieval

We extracted a total of 10,403 diabetes patients contained in the MIMIC III database. Statistical numbers of retrieved diabetic patients are summarized in Table 8. This number shows about 22% of the total ICU population in our dataset. This is a significant portion of the critical care population and, therefore, ideal for our research. The diabetic patients were selected based on the ICD codes for diabetes shown in Table 6.

We found 28 different comorbidities. 10,403 diabetes patients, comorbidities were based on the Elixhauser Comorbidity Index. The comorbidity information was based on the information recorded in the MIMIC III database for each of the patients during their hospital stay.

Table 9 summarizes the ranked list of comorbidities in diabetic patients based on the occurrence rate. The diseases contained in the comorbidity list were filtered using ICD codes. Since our cohort was only diabetic patients, we excluded diabetes as part of the comorbidities explored. 1,513 diabetic patients died during their hospital stay, which is about 15% of all diabetic patients. However, the cause of death was not always diabetes.

Table 8- Statistics of diabetes patients on MIMIC III

Age group	Gender	Count
adult	F	4,189
adult	M	5,784
>89	F	264
>89	M	165
neonate	F	1
Total		10,403

Out of our diabetic patient cohort, around 43% of the patients are female while 57% are male. For this research, no distinction was made based on gender in creating our predictive model. This gender proportion distribution in our patient cohort is consistent with the reviewed literature on intensive care admissions.

Table 9- List of comorbidities in MIMIC-III diabetic patients

Condition	Frequency
fluid electrolyte	4,306
renal failure	3,582
deficiency anemias	3,213
congestive heart failure	3,116
hypertension	3,005
chronic pulmonary	2,854
cardiac arrhythmias	2,778
peripheral vascular	1,760
hypothyroidism	1,659
coagulopathy	1,468
obesity	1,398
depression	1,276
other neurological	1,000
liver disease	990
valvular disease	836

pulmonary circulation	558
psychoses	540
alcohol abuse	523
weight loss	435
metastatic cancer	405
paralysis	393
solid tumor	369
rheumatoid arthritis	350
blood loss anemia	313
drug abuse	249
lymphoma	142
AIDS	35
peptic ulcer	8

According to the comorbidity results, some disorders are more prevalent in diabetic patients than others. This finding is consistent with the literature we reviewed.

4.2 Results for Mortality Measure Calculation

The SOFA score was determined for each patient, and the results are summarized in Table 10. The table shows that three estimated mortalities are within the range of predicted mortality based on established SOFA statistics (Ferreira et al., 2001). However, the computed mortality is lower than the expected one for SOFA scores = 15 and score > 16. More discussion will be conducted in Section 0 concerning the observed disparity of mortality between the published ones and the calculated values.

Table 10- Expected vs. Computed sofa mortality prediction

SOFA Score	**Expected Mortality**	**Computed Mortality**
0 to 6	< 10%	7%
7 to 9	15 - 20%	20%
10 to 12	40 - 50%	39%
13 to 14	*50 - 60%*	*59%*
15	*> 80%*	*64%*
16 to 24	*> 90%*	*72%*

4.3 Results for Comorbidity Measure Calculation

In Table 11, we have the results of comorbidity calculations reported for diabetic patients using the Elixhauser Comorbidity Index, which indicated that the majority of diabetic patients suffering from multiple comorbidities. This result was consistent with our study in the literature concerning diabetic patients. Numerous studies have shown that comorbidities are a mortality risk factor to diabetic patients, with some diseases being more predictive of death than others. With the help of this result on patient cohort, we show that comorbidities are prevalent within diabetic patients. Then we can drill down on them for further analysis. Such a high prevalence of comorbidities demonstrates the importance to predict mortality by taking into consideration of comorbidity.

Table 11- Elixhauser scores

Elixhauser Score	Explanation	# Patients
1	Associated with one comorbidity	1838
2	Associated with two comorbidities	2495
3	Associated with three comorbidities	2467
4	Associated with four comorbidities	2299
5	Associated with five comorbidities	1846
6	Associated with six comorbidities	1170
7	Associated with seven comorbidities	750
8	Associated with eight comorbidities	411
9	Associated with night comorbidities	181
10	Associated with ten comorbidities	85
11	Associated with eleven comorbidities	30
12	Associated with twelve comorbidities	9
13	Associated with thirteen comorbidities	1

4.4 Results for the Most Relevant Mortality Predictive Comorbidities

use comorbidity information had to be included in our model as a part of our input attributes. To accomplish this, we conducted several experiments to assess the respective weights of the comorbidities concerning mortality prediction. Since we had 29 different diseases, we wanted to find a way to reduce the number of diseases we used while maintaining the predictive power of our model. To accomplish this, we used five different feature extraction algorithms. We weighed their results to get a final ranking to use for our model.

Tables 12-16 illustrates the results of the five feature extraction algorithms used to evaluate comorbidities to be applied for prediction classifiers. Five algorithms were selected for a better estimate of the diseases' weight in mortality prediction. Each disease was ranked from 1 (most predictive) to 6 (6th most predictive). The final ranking was calculated by taking the rank of each disease averaged over the ranking algorithms, taking only the top 5. Table 17 summarizes the comorbidities based on their average ranking.

Feature selection methods generally fall into four types: filter methods, wrapper methods, embedded methods and hybrid methods. Regardless of the type of feature selection method used, the idea is to have the most predictive feature of the target variable identified and used.

Feature selection filter methods use univariate statistics instead of cross-validation to express the intrinsic properties of the features. In comparison to wrapper methods, filter methods are quicker and more efficient in terms of computation. They are, thus, suitable for dealing with multidimensional data, which is common in clinical data. Examples of filter approaches include information gain, which evaluates the gain of information by each variable in the context of impacting the target variable. This is accomplished by computing the reduction in entropy associated with a change in a dataset. Other examples are Fisher's score and chi-square test, which is used for categorical features in a dataset. The chi-square test for feature selection is used to test the independence of two events. From the data of two variables, we compute the anticipated count E and get the observed count O. Chi-square then is used to measure how E and O deviate from each other using the Tab below:

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where c is the degrees of freedom.

One method we experimented with and tabulated the results is the correlation coefficient. This metric determines the linear relationship between two or more variables. We can predict one variable from the other(s). The fundamental principle here is that the desired variables are highly correlated with the target but not with one another. Several correlation measures can be used, such as the Pearson correlation or the Spearman's correlation coefficient. The equation below shows the correlation feature selection (CFS) given by the maximum from a feature subset *S* consisting of *k* features.

$$CFS = \underset{S_i}{max}\left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{k + 2(r_{f_1 f_2} + \cdots + r_{f_i f_j} + \cdots + r_{f_k f_{k-1}})}\right]$$

Where r_cf is feature-classification correlations and *rff* is the feature-feature correlations.

Feature selection wrapper methods use a greedy approach to evaluate all possible combinations of features to assess their performance by evaluating them on a classifier with that feature subset. This feature selection process is based on an ML algorithm to fit into a specified dataset. Generally, wrapper methods produce more reliable results than filter methods, but they are more computationally expensive. Some of the techniques used on wrapper methods are forward feature selection, backward feature elimination, complete feature selection, and recursive feature elimination.

The forward feature selection technique begins with the variable that is most target's predictive component. The next iteration selects a second variable, giving the best performance when combined with the first variable. This process continues until the defined criteria are met. The backward feature wrapper method of elimination works similarly to the forward method. Yet, it begins with all the available features, systematically eliminating one until the criteria are met. The exhaustive feature selection method uses a brute force approach to evaluate every possible combination of the variables. As such, this is the most robust method but also the most computationally expensive.

Embedded feature selection approaches incorporate the advantages of wrapper and filter methods. They accomplish this by iteratively balancing computational cost with the inclusion of interactions of features. Some examples of embedded methods include LASSO regularization and Random Forest importance methods. To prevent overfitting, LASSO regularization applies a penalty to different parameters of a machine learning model. It can vary the coefficients of the features and shrink some to zero, effectively removing them from the model. Hybrid feature selection techniques combine approaches from any of the embedded, wrapper, or filter techniques.

The Symmetrical Uncertainty method is used to measure relevance between two random variables. We use this technique as one of our five techniques to ensure that we capture all and any significance between comorbidities that may be independent of each other. Although comorbidities are often associated with diabetes, the cause of death for a diabetic patient can be caused by any condition they present within the ICU. As observed, clinical data is complex, and so are the interactions between various diseases.

Table 12- Correlational feature selection

Comorbidity	Rank
Cardiac arrhythmias	1
Peptic ulcer	2
Metastatic cancer	3
Coagulopathy	4
Obesity	5
Fluid electrolyte	6

Table 13- Correlational attribute

Comorbidity	Rank
Coagulopathy	1
Fluid electrolyte	2
Cardiac arrhythmias	3
Congestive heart failure	4
Metastatic cancer	5
Renal failure	6

Table 14- Gain ratio

Comorbidity	**Rank**
Coagulopathy	1
Metastatic cancer	2
Peptic ulcer	3
Fluid electrolyte	4
Cardiac arrhythmias	5
Congestive heart failure	6

Table 15- Information gain

Comorbidity	**Rank**
Coagulopathy	1
Fluid electrolyte	2
Cardiac arrhythmias	3
Congestive heart failure	4
Metastatic cancer	5
Renal Failure	6

Table 16- Symmetrical uncertain

Comorbidity	**Rank**
Coagulopathy	1
Fluid electrolyte	2
Cardiac arrhythmias	3
Metastatic cancer	4
Congestive heart failure	5
Liver disease	6

Table 17- Average ranking

Comorbidity	**Average Rank**
Coagulopathy	1.6
Cardiac arrhythmias	3
Fluid electrolyte	3.2
Metastatic cancer	3.8
Congestive heart failure	5.2

4.5 Results for Mortality Prediction

The results of the NB and RF classifiers are shown in Table 18. When the two classifiers are compared, the results are comparable. The receiver operating characteristic (ROC) curve is lowest when using only the 5 comorbidity features in Table 17, and better when using the 6 SOFA scores described by Table 4, and best when both the 5 comorbidities and 6 SOFA scores are combined. We use NB and RF classifiers as examples of ML tools, but any appropriate classifier can replace these. One crucial takeaway is recognizing that what we present here is a framework that is not dependent on the actual ML algorithms or feature selection techniques. From an architectural standpoint, our approach is sufficiently adaptable to accommodate a variety of techniques and tools. Care, however, should be taken exercised regarding the selection criteria to avoid bias. We have demonstrated approaches that minimize bias based on our patient cohort, target variable (mortality), and attributes.

Table 18- Mortality prediction results

Features (count)	Random Forest ROC Area	Naive Bayes ROC Area
Comorbidities (5)	0.667	0.672
SOFA (6)	0.731	0.743
Combined (11)	0.763	0.772

4.6 Discussion of numerical results

We have introduced a novel approach to predict mortality by applying different mortality measures and comorbidity measures to improve the predictions. We have shown that combining comorbidity information with features in the SOFA severity measure can predict mortality with higher accuracy than using the SOFA features by themselves. This is important and powerful because it enables the repurposing of existing mortality prediction tools to utilize machine learning techniques. To accomplish this, an understanding of the drivers of the existing tool is needed to translate them to attributes in a prediction model. The ability to add additional attributes then augments the model to improve its performance.

Our study recognized that SOFA severity scoring is a good candidate for our approach because it captures numerous clinical biomarkers covering six body systems. This understanding helped us know that the constituent SOFA scores for each body system were different enough to use Naïve Bayes as our machine learning algorithms. The precise weighing of our attributes was based on an understanding of the relative importance of the selected comorbidities and the SOFA variables. This is important because other mortality prediction tools or scores may not necessarily have the properties to allow for equal attribute weights. We have mitigated this by using multiple feature selection techniques to account for the relevancy and redundancy of predictive attributes.

As shown in Table 10, we validated the SOFA against MIMIC-III diabetes patients. This was a critical exercise since a substantial portion of our work is predicated on the published claims of SOFA's ability to predict mortality. We, therefore, needed to make sure that our patient cohort aligns with the expected SOFA mortality results. For our selected cohort, 4 out of the 6 SOFA categories shown in Table 10 corresponded to the expected results. However, diabetic patients with a SOFA value of 15 or higher experienced lower mortality than predicted by SOFA. We hypothesize that this disparity is due to improved treatment and management of diabetes. Over the years, further research has been conducted on the treatment of diabetes in both medical interventions and dietary improvements. There are more accurate methods available for monitoring and controlling blood sugar levels. Some automated devices can now be used to administer insulin. We, therefore, accepted SOFA as validated by our patient cohort.

We showed that in mortality prediction modeling, we can use feature selection techniques to identify relevant comorbidities. One critical observation made was that the most predictive comorbidities were not necessarily the most prevalent, as shown in Table 11 vs Table 13. While there are clinical explanations for this discovery, our findings demonstrate that our methods can detect these indicators. This becomes crucial when dealing with features that are not well known or researched. Having automated tools that ingest the clinical data and provide this ranking can help advance outcome prediction. An example is when using genome data where the clinical impact of such biomarker is obvious.

Extracting the most predictive comorbidities is also important in reducing the number of features in our classifier, making it more efficient without losing valuable information. Having an excessive set of features can result in overfitting and rendering the models ungeneralizable. We validated that reducing the number of featured items had no adverse effect on our test results. As a result, there is a balance between reducing the features while still retaining the predictive ability. We experimented with our comorbidity features to arrive at the number of comorbidities to use in our modeling.

We demonstrated that patient similarity is an essential component in ICU mortality prediction and can be combined with other techniques to improve prediction. We unconventionally approached patient similarity. We looked at patent similarity from the lens of patient cohort selection and shared comorbidities. We showed that we can define patient cohorts in various ways within our framework using inpatient electronic medical records. We used clinical conditions based on ICD-9 codes for this, but other published ways of defining patient similarity.

Chapter 5
Conclusion and Future Work

This dissertation has demonstrated a novel approach that has improved diabetic patient mortality prediction by using a combination of SOFA scoring and ranked comorbidity data over using SOFA scoring alone. We have shown that we can better predict mortality outcomes by combining machine learning algorithms with data from electronic health records and comorbidity information. We believe that the approach we have presented lays a solid foundation for the future development of Artificial Intelligence in clinical informatics.

In our research, we have presented several strengths and perspectives in ICU mortality prediction. Extensive literature research has uncovered the inherent potential of patient similarity measures in predicting health outcomes. We summarize previous research in this area, with an emphasis on ICU mortality prediction. We note the gap that exists in having a generalized tool that can be used for mortality prediction. Although Machine Learning is well developed and matured, its application in the optimal utilization of clinical data has remained a challenge. Our work demonstrates a step in realizing the full potential of machine learning applications within clinical data.

We have taken a resource-constrained approach to clinical prediction, making it more accessible to ICU providers. The Machine Learning tools used are standardized and readily available. The clinical data requires SOFA scoring, which can be obtained from patient records. This also provides details about comorbidities, which is documented in the patient medical record. This arrangement enables smaller ICU providers the opportunity to utilize our novel prediction tools while being sensitive to other competing priorities for hospital resources.

The successful use of biomarkers in prediction modeling is one of the most significant challenges in general clinical prediction. We observed that as more biomarkers are being developed, there is minimal uptake of them in modeling. Our framework tackles this issue by demonstrating the efficacy of biomarkers through the use of SOFA outcomes. With existing matured machine learning tools capable of handling big data, new biomarkers, including genomic data, can be harnessed to predict various clinical outcomes.

To our knowledge, this is the first investigation that combines a pre-existing mortality score with patient comorbidity to predict ICU mortality of diabetic patients. This is vital because diabetes is a prevalent chronic disease and the 7th leading cause of death in the United States. Diabetes, as a condition that affects the sugar level in the blood, adversely affects other body systems. Diabetic patients, especially those hospitalized in critical care, often have comorbidities. As previously mentioned, comorbidities play a significant role in mortality prediction modeling.

By ranking comorbidities based on their mortality predictor weights, the machine learning models could train with fewer features while maintaining predictive accuracy. With 11 attributes for training our model, we have demonstrated that successful prediction models do not rely on many attributes. Instead, having an informed set of predictors that encompasses the clinically relevant variables would suffice.

This study demonstrated that machine learning that ML algorithms can be used to improve the accuracy of predicting the mortality of a cohort of diabetic patients using standardized electronic medical records. It has been demonstrated that combining SOFA scores with carefully chosen comorbidities within a diabetic patient cohort can improve mortality prediction. These results indicate that a hybrid approach of mortality prediction of ICU diabetic patients provides a potential for integration into critical care clinical workflows.

We have gone further to show how to maximize the use of comorbidity, ranking the predictive weights of the comorbidities. The rationale to rank the comorbidities rather than using all available comorbidity information is twofold. First, we observed that the top five weighted comorbidities adequately predicted our clinical endpoint. This observation is consistent with the principle of parsimony (Occam’s razor). Second, we noted that the predictive model was more time and resource-efficient while preserving its accuracy. Ranking of attributes for clinical outcomes is important because clinical data can consist of numerous variables, which may not be of clinical significance to the identified outcome.

In our study, we have used diabetic patients in ICU. It is important to note that there are several co-occurrences of diabetes including hypertension, coronary artery disease, chronic kidney disease, and peripheral vascular disease. Therefore, in studying our patient cohort, it is also possible that we are studying patients with other co-occurrences of diabetes which are significant diabetes comorbidities. Consequently, our results may have unintended confounders.

Our report identifies many potential future opportunities. First, our research can be expanded to showcase patient cohorts with various other diseases and endpoints. In our work, we have shown that our methodology applies to diabetic patients. However, nothing about these patients precludes our results from being beneficial to any other medical condition. This is because the machine learning approaches, and comorbidity data can still be successfully utilized in outcome prediction. Likewise, we used mortality as a measure of success in our research. We observe that with minimal modifications to our methodology, the endpoint can be modified to any meaningful clinical endpoint. An example would be important measures such as hospital length of stay, central line infections, etc.

We intend to strengthen our topics beyond ICU patients in the future. This will ensure the appropriate calibration for other patient populations while adhering to our core solution approach. Indeed, we would like to expand our solution by investigating prediction models encompassing multiple conditions. For instance, we could study patients with co-occurring conditions such as both obesity and diabetes. This would allow us to experiment with the combination of attributes that collectively drive specified endpoints. Such an approach would utilize Machine Learning algorithms to measure the similarity of diseases.

One approach could be to use an ontological metric to measure semantic similarity between diseases (Mathur & Dinakarpandian, 2012). Another approach may quantify the pairwise similarity in diseases by investigating the associations at the molecular, phenotypic, and taxonomic levels (Cheng et al., 2019).

Another future advancement of our approach would be experimenting with several similar endpoints. Often patients with chronic diseases are concerned about several endpoints, including overall survival, duration of intervention response, time to disease progression, etc. Our approach can be modified to fit specific endpoints. This is beneficial in treating diseases like cancer, where the clinical team and patient seek to obtain knowledge about patient survival and cancer progression.

In the future, we intend to incorporate other cutting-edge mortality prediction tools in our method. This would establish an ideal situation in which our method would compete for the best results based on the available mortality prediction scores. Certain mortality prediction tools may be best suited for specific patient cohorts. In our method, we demonstrated how SOFA scoring can be used to improve mortality prediction. Given that this is a hybrid approach where we combine several predictors, selecting from a stack of available scoring tools would increase the versatility of our method.

The discovery of several additional biomarkers is a more recent trend in personalized medicine. The National Institutes of Health (NIH) defines a biomarker as a quantifiable biological parameter measured and evaluated as an indicator of normal biological, pathogenic, or pharmacologic responses to a therapeutic intervention. Technological and scientific advancements have aided in the comprehension of genome data. Certain elements of this data are prosecutable and have been known to influence clinical outcomes. The research in this document does not include genome data, which could be an important extension in ICU mortality prediction.

An important and futuristic opportunity for clinical predictive modeling is having a common standard for storing clinical data that can be used by researchers across different organizations and specialties. This is akin to a unified data model containing patient data that can be anonymized and available to the research community. This would allow different electronic medical record systems to have their data exported to this data model standard or separately processed to conform to the standard. Semantic enrichment of the data could then happen to ensure standard terminologies are used and a dictionary of these terminologies exists. This dictionary would evolve as new terminologies are added. With this approach, we could have a time-invariant data model that can benefit clinical analytics. It would be possible to use large sets of data sourced from a wide range of patients, making predictions more generalizable.

Lastly, our approach has the potential to be used within cohorts in pharmacological clinical trials. These clinical trials are divided into four phases. The first phase is concerned with safety, while the second phase deals with both safety and efficacy. The third phase focuses on effectiveness, efficacy, and safety. In contrast, the fourth phase is involved with post-marketing approval and long-term effects. Patients participating in clinical trials have their response to interventions recorded, tracked, and monitored. This trial data can predict various outcomes based on the characteristics of the patients and observed responses. This information can then help inform the study directors and others about potential protocol amendments or future study designs.

Appendices

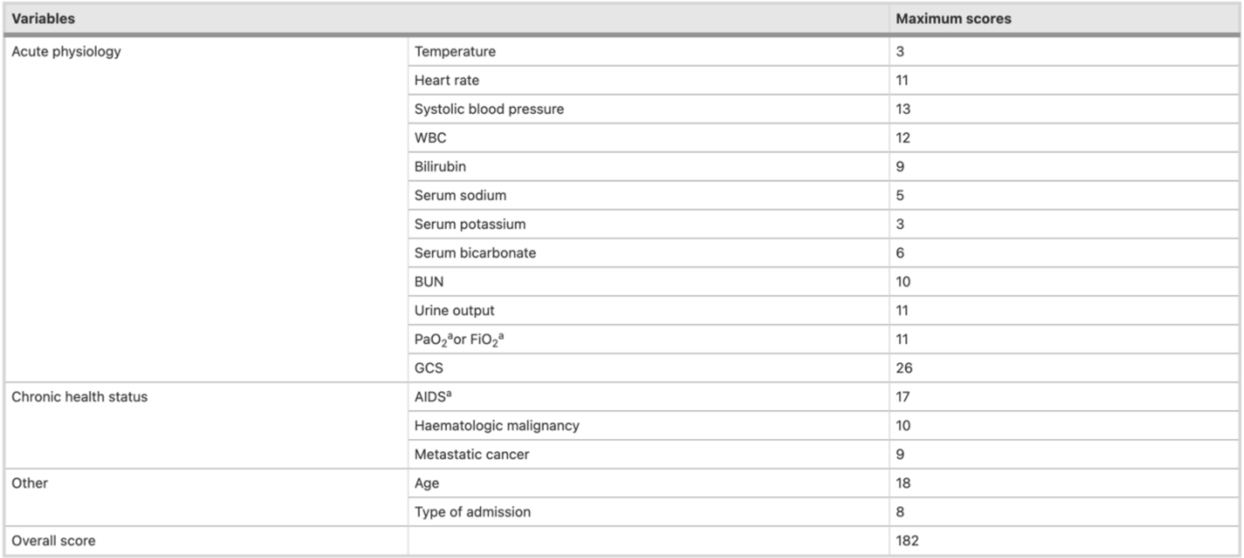
Appendix A: Elixhauser Comorbidities

Source: (Quan et al., 2005)

Comorbidities	**ICD-9 Codes**
Congestive heart failure	398.91, 402.11, 402.91, 404.11, 404.13, 404.91, 404.93, 428.x
Cardiac arrhythmias	426.10, 426.11, 426.13, 426.2-426.53, 426.6-426.8, 427.0, 427.2, 427.31, 427.60, 427.9, 785.0, V45.0, V53.3
Valvular disease	093.2, 394.0-397.1, 424.0-424.91, 746.3-746.6, V42.2, V43.3
Pulmonary circulation Disorders	416.x, 417.9
Peripheral vascular disorders	440.x, 441.2, 441.4, 441.7, 441.9, 443.1- 443.9, 447.1, 557.1, 557.9, V43.4
Hypertension, uncomplicated	401.1, 401.9

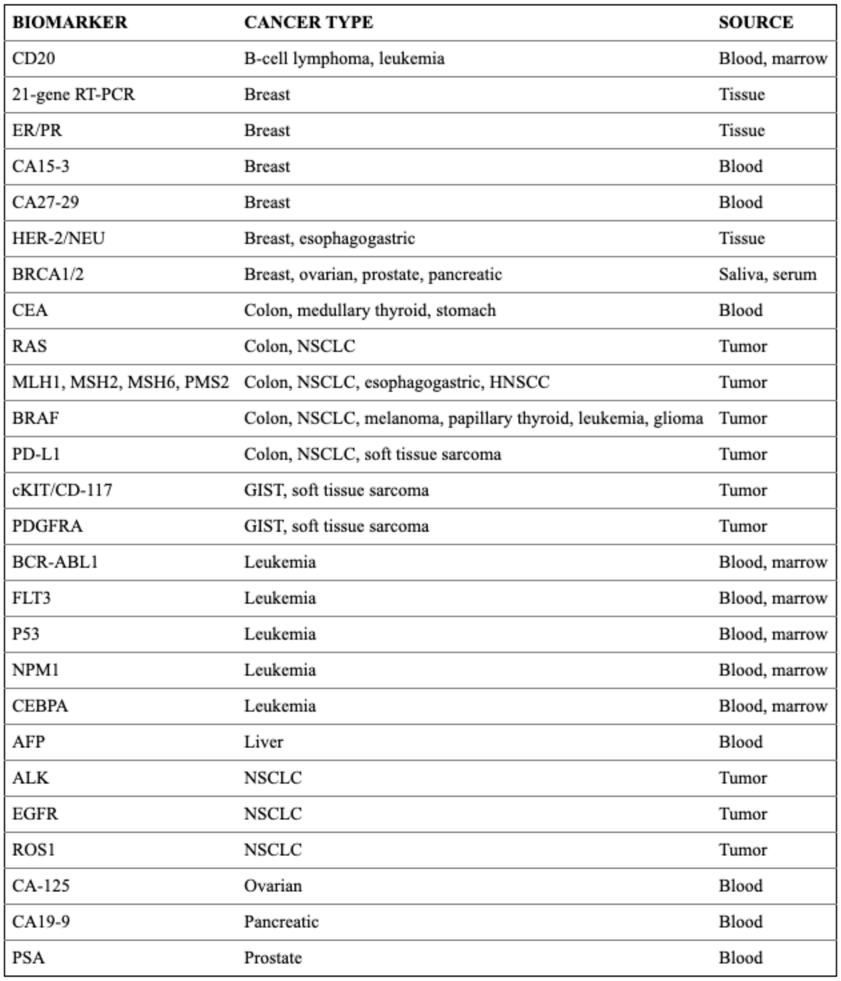
Hypertension, complicated	402.10, 402.90, 404.10, 404.90, 405.1, 405.9
Paralysis	342.0. 342.1, 342.9-344.x
Other neurological disorders	331.9, 332.0, 333.4, 333.5, 334.x, 335.x, 340.x, 341.1-341.9, 345.0, 345.1, 345.4, 345.5, 345.8, 345.9, 348.1, 348.3, 780.3, 784.3
Chronic pulmonary disease	490-492.8, 493.00-493.91, 494.x-505.x, 506.4
Diabetes, uncomplicated	250.0-250.3
Diabetes, complicated	250.4-250.7, 250.9
Hypothyroidism	243-244.2, 244.8, 244.9
Renal failure	403.11, 403.91, 404.12, 404.92, 585.x, 586.x,V42.0, V45.1, V56.0, V56.8
Liver disease	070.32, 070.33, 070.54, 456.0, 456.1, 456.2, 571.0, 571.2-571.9, 572.3, 572.8, V42.7
Peptic ulcer disease excluding bleeding	531.70, 531.90, 532.70, 532.90, 533.70, 533.90, 534.70, 534.90, V12.71
AIDS/H1V	042.x-044.x
Lymphoma	200.x-202.3x, 202.5-203.0, 203.8, 238.6, 273.3, V10.71, V10.72, V10.79
Metastatic cancer	196.x-199.x
Solid tumor without metastasis	140.x-172.x, 174.x, 175.x, 179.x-195.x, V10.x
Rheumatoid arthritis/ collagen vascular diseases	701.0, 710.x, 714.x, 720.x, 725.x
Coagulopathy	286.x, 287.1, 287.3-287.5
Obesity	278.0
Weight loss	260.x-263.x
Fluid and electrolyte disorders	276.x
Blood loss anemia	280.0
Deficiency anemia	280.1-281.9, 285.9
Alcohol abuse	291.1, 291.2, 291.5-291.9, 303.9, 305.0, V113
Drug abuse	292.0, 292.82-292.89, 292.9, 304.0, 305.2, 305.9
Psychoses	295.x-298.x, 299.1
Depression	300.4, 301.12, 309.0, 309.1, 311

Appendix B: SAPS II Variables and Score Assignment



Variables		Maximum scores
Acute physiology	Temperature	3
	Heart rate	11
	Systolic blood pressure	13
	WBC	12
	Bilirubin	9
	Serum sodium	5
	Serum potassium	3
	Serum bicarbonate	6
	BUN	10
	Urine output	11
	PaO_2[a]or FiO_2[a]	11
	GCS	26
Chronic health status	AIDS[a]	17
	Haematologic malignancy	10
	Metastatic cancer	9
Other	Age	18
	Type of admission	8
Overall score		182

Appendix C: Some Biomarkers used in Cancer Treatment



BIOMARKER	CANCER TYPE	SOURCE
CD20	B-cell lymphoma, leukemia	Blood, marrow
21-gene RT-PCR	Breast	Tissue
ER/PR	Breast	Tissue
CA15-3	Breast	Blood
CA27-29	Breast	Blood
HER-2/NEU	Breast, esophagogastric	Tissue
BRCA1/2	Breast, ovarian, prostate, pancreatic	Saliva, serum
CEA	Colon, medullary thyroid, stomach	Blood
RAS	Colon, NSCLC	Tumor
MLH1, MSH2, MSH6, PMS2	Colon, NSCLC, esophagogastric, HNSCC	Tumor
BRAF	Colon, NSCLC, melanoma, papillary thyroid, leukemia, glioma	Tumor
PD-L1	Colon, NSCLC, soft tissue sarcoma	Tumor
cKIT/CD-117	GIST, soft tissue sarcoma	Tumor
PDGFRA	GIST, soft tissue sarcoma	Tumor
BCR-ABL1	Leukemia	Blood, marrow
FLT3	Leukemia	Blood, marrow
P53	Leukemia	Blood, marrow
NPM1	Leukemia	Blood, marrow
CEBPA	Leukemia	Blood, marrow
AFP	Liver	Blood
ALK	NSCLC	Tumor
EGFR	NSCLC	Tumor
ROS1	NSCLC	Tumor
CA-125	Ovarian	Blood
CA19-9	Pancreatic	Blood
PSA	Prostate	Blood

References

Quan H, Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.C., Saunders, L.D., Beck, C.A., Feasby, T.E., Ghali, W.A. (2005/11). Coding algorithms for defining Comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care, 43(11)*, 1130–1139.

Anand, R. S., Stey, P., Jain, S., Biron, D. R., Bhatt, H., Monteiro, K., … Chen, E. S. (2018). Predicting Mortality in Diabetic ICU Patients Using Machine Learning and Severity Indices. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2017*, 310–319.

Arabi, Y., Shirawi, N. A., Memish, Z., Venkatesh, S., & Al-Shimemeri, A. (2003). Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: a prospective cohort study. *Critical Care* , *7*(5), R116.

Aravind, M., & Chung, K. C. (2010). Evidence-based medicine and hospital reform: tracing origins back to Florence Nightingale. *Plastic and Reconstructive Surgery*, *125*(1), 403–409.

Austin, S. R., Wong, Y.-N., Uzzo, R. G., Beck, J. R., & Egleston, B. L. (2015). Why summary comorbidity measures such as the Charlson Comorbidity Index and Elixhauser score work. *Medical Care*, *53*(9), e65-72.

Awad, A., Bader-El-Den, M., McNicholas, J., & Briggs, J. (2017). Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International Journal of Medical Informatics*, *108*, 185–195.

Balachandran, V. P., Gonen, M., Smith, J. J., & DeMatteo, R. P. (2015). Nomograms in oncology: more than meets the eye. *The Lancet Oncology*, *16*(4), e173-80.

Begum, M. (2017). Identification of the Risk Factors Associated with ICU Mortality. *Biometrics & Biostatistics International Journal*, *6*(1). doi:10.15406/bbij.2017.06.00157

Braun, K. F., Otter, W., Sandor, S. M., Standl, E., & Schnell, O. (2012). All-cause in-hospital mortality and comorbidity in diabetic and non-diabetic patients with stroke. *Diabetes Research and Clinical Practice*, *98*(1), 164–168.

Breitling, R. (2010). What is systems biology? *Frontiers in Physiology*, *1*, 9.

Brown, S.-A. (2016). Patient Similarity: Emerging Concepts in Systems and Precision Medicine. *Frontiers in Physiology*, *7*, 561.

Campion, F. X., Carlsson, G., & Francis, F. (2017). *Machine Intelligence for Healthcare*.

Ceriani, R., Mazzoni, M., Bortone, F., Gandini, S., Solinas, C., Susini, G., & Parodi, O. (2003). Application of the sequential organ failure assessment score to cardiac surgical patients. *Chest*, *123*(4), 1229–1239.

Chan, L., Chan, T., Cheng, L., & Mak, W. (2010, December). Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy. *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, 467–470.

Chen, Y. C., Lin, S. F., Liu, C. J., Jiang, D. D., Yang, P. C., & Chang, S. C. (2001). Risk factors for ICU mortality in critically ill patients. *Journal of the Formosan Medical Association = Taiwan Yi Zhi*, *100*(10), 656–661.

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., … Jiang, Q. (2019). Computational methods for identifying similar diseases. *Molecular Therapy. Nucleic Acids*, *18*, 590–604.

Ciechanowski, P. S., Katon, W. J., & Russo, J. E. (2000). Depression and diabetes: impact of depressive symptoms on adherence, function, and costs. *Archives of Internal Medicine*, *160*(21), 3278–3285.

Diabetes, Y. C. M. (2001). WHAT CAN YOU DO? *Changes in Lifestyle among Subjects with Impaired Glucose Tolerance. N Engl J Med*, *344*, 1343–1350.

Doerr, F., Badreldin, A. M. A., Heldwein, M. B., Bossert, T., Richter, M., Lehmann, T., … Hekmat, K. (2011). A comparative study of four intensive care outcome prediction models in cardiac surgery patients. *Journal of Cardiothoracic Surgery*, *6*(1), 1–8.

Dorr, D. A., Jones, S. S., Burns, L., Donnelly, S. M., Brunker, C. P., Wilcox, A., & Clayton, P. D. (2006). Use of health-related, quality-of-life metrics to predict mortality and hospitalizations in community-dwelling seniors. *Journal of the American Geriatrics Society*, *54*(4), 667–673.

Dubois, R. W., Chawla, A. J., Neslusan, C. A., Smith, M. W., & Wade, S. (2000). Explaining drug spending trends: does perception match reality? *Health Affairs*, *19*(2), 231–239.

Dworzynski, P., Aasbrenn, M., Rostgaard, K., Melbye, M., Gerds, T. A., Hjalgrim, H., & Pers, T. H. (2020). Nationwide prediction of type 2 diabetes comorbidities. *Scientific Reports*, *10*(1), 1776.

Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical Care*, *36*(1), 8–27.

FDA-NIH Biomarker Working Group. (2021). *FDA-NIH Biomarker Working Group*. Food and Drug Administration (US).

Ferreira, F. L., Bota, D. P., Bross, A., Mélot, C., & Vincent, J. L. (2001). Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA: The Journal of the American Medical Association*, *286*(14), 1754–1758.

Geneviève, L. D., Martani, A., Mallet, M. C., Wangmo, T., & Elger, B. S. (2019). Factors influencing harmonized health data collection, sharing and linkage in Denmark and Switzerland: A systematic review. *PloS One*, *14*(12), e0226015.

Gorin, S. S. (2014). *Prevention Practice in Primary Care*. Oxford University Press.

Gottlieb, A., Stein, G. Y., Ruppin, E., Altman, R. B., & Sharan, R. (2013). A method for inferring medical diagnoses from patient similarities. *BMC Medicine*, *11*, 194.

Haendel, M. A., Chute, C. G., & Robinson, P. N. (2018). Classification, ontology, and precision medicine. *The New England Journal of Medicine*, *379*(15), 1452–1462.

Hall, M. A., & Smith, L. A. (1999). Feature Selection For Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper. Retrieved November 17, 2016, from http://www.aaai.org/Library/FLAIRS/1999/flairs99-042.php

Hall, Mark A., & Smith, L. A. (1998). *Practical feature subset selection for machine learning*. Retrieved from http://researchcommons.waikato.ac.nz/handle/10289/1512

Halpern, Y., Choi, Y., Horng, S., & Sontag, D. (2014). Using anchors to estimate clinical state without labeled data. *AMIA Annual Symposium Proceedings*, *2014*, 606–615.

Hu, Z., Melton, G. B., Arsoniadis, E. G., Wang, Y., Kwaan, M. R., & Simon, G. J. (2017). Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics*, *68*, 112–120.

Huerta, L. E., Wanderer, J. P., Ehrenfeld, J. M., Freundlich, R. E., Rice, T. W., Semler, M. W., & SMART Investigators and the Pragmatic Critical Care Research Group. (2018). Validation of a sequential organ failure assessment score using electronic health record data. *Journal of Medical Systems*, *42*(10), 199.

Hunter, J., Freer, Y., Gatt, A., Logie, R., McIntosh, N., van der Meulen, M., … Sykes, C. (2008). Summarising complex ICU data in natural language. *AMIA Annual Symposium Proceedings*, 323–327.

Ibrahim, J. G., Chu, H., & Chen, M.-H. (2012). Missing data in clinical studies: issues and methods. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *30*(26), 3297–3303.

Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). Involvement of Machine Learning Tools in Healthcare Decision Making. *Journal of Healthcare Engineering*, *2021*, 6679512.

Jia, P., Dai, J.-H., Pan, Y.-H., & Zhu, M.-L. (2006). Novel algorithm for attribute reduction based on mutual-information gain ratio. *JOURNAL-ZHEJIANG UNIVERSITY ENGINEERING SCIENCE*, *40*(6), 1041.

Johnson, A. E. W., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., & Clifford, G. D. (2016a). Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, *104*(2), 444–466.

Johnson, A. E. W., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D., & Clifford, G. D. (2016b). Machine learning and decision support in critical care. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, *104*(2), 444–466.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., … Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035.

Jones, A. E., Trzeciak, S., & Kline, J. A. (2009). The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Critical Care Medicine*, *37*(5), 1649–1654.

Juneja, D., Singh, O., Nasa, P., & Dang, R. (2012). Comparison of newer scoring systems with the conventional scoring systems in general intensive care population. *Minerva Anestesiologica*, *78*(2), 194–200.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271–277.

Katon, W. J., Rutter, C., Simon, G., Lin, E. H. B., Ludman, E., Ciechanowski, P., … Von Korff, M. (2005). The Association of Comorbid Depression With Mortality in Patients With Type 2 Diabetes. *Diabetes Care*, *28*(11), 2668–2672.

Kattan, M. W., & Gerds, T. A. (2020). A framework for the evaluation of statistical prediction models. *Chest*, *158*(1S), S29–S38.

Keegan, M. T., Gajic, O., & Afessa, B. (2011). Severity of illness scoring systems in the intensive care unit. *Critical Care Medicine*, *39*(1), 163–169.

Kelly, F. E., Fong, K., Hirsch, N., & Nolan, J. P. (2014). Intensive care medicine is 60 years old: the history and future of the intensive care unit. *Clinical Medicine*, *14*(4), 376–379.

Knaus, W. A., Draper, E. A., Wagner, D. P., & Zimmerman, J. E. (1985). APACHE II: a severity of disease classification system. *Critical Care Medicine*, *13*(10), 818–829.

Kollef, M. H., & Schuster, D. P. (1994). Predicting intensive care unit outcome with scoring systems. Underlying concepts and principles. *Critical Care Clinics*, *10*(1), 1–18.

Koskinen, S. V., Reunanen, A. R., Martelin, T. P., & Valkonen, T. (1998). Mortality in a large population-based cohort of patients with drug-treated diabetes mellitus. *American Journal of Public Health*, *88*(5), 765–770.

Kramer, A. A., Higgins, T. L., & Zimmerman, J. E. (2014). Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV hospital mortality models: implications for national benchmarking*. *Critical Care Medicine*, *42*(3), 544–553.

Kuzniewicz, M. W., Vasilevskis, E. E., Lane, R., Dean, M. L., Trivedi, N. G., Rennie, D. J., … Dudley, R. A. (2008). Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. *Chest*, *133*(6), 1319–1327.

Le Gall, J. R., Lemeshow, S., & Saulnier, F. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA: The Journal of the American Medical Association*, *270*(24), 2957–2963.

Lee, J., Maslove, D. M., & Dubin, J. A. (2015). Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric. *PloS One*, *10*(5), e0127428.

Lehman, L.-W., Saeed, M., Long, W., Lee, J., & Mark, R. (2012). Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2012,* 505–511.

Lemeshow, S., Teres, D., Klar, J., Avrunin, J. S., Gehlbach, S. H., & Rapoport, J. (1993). Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA: The Journal of the American Medical Association*, *270*(20), 2478–2486.

Li, G., Thabane, L., Cook, D. J., Lopes, R. D., Marshall, J. C., Guyatt, G., … Levine, M. A. H. (2016). Risk factors for and prediction of mortality in critically ill medical-surgical patients receiving heparin thromboprophylaxis. *Annals of Intensive Care*, *6*(1), 18.

Mathur, S., & Dinakarpandian, D. (2012). Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, *45*(2), 363–371.

Maxwell, C. A., Moreno, V., Solé, X., Gómez, L., Hernández, P., Urruticoechea, A., & Pujana, M. A. (2008). Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment. *Molecular Cancer*, *7*(1), 4.

McCormick, K. A., & Gugerty, B. (2013). *Healthcare information technology exam guide for CompTIA healthcare IT technician and HIT pro certifications*. New York, NY: McGraw-Hill Professional.

Mehta, H. B., Sura, S. D., Adhikari, D., Andersen, C. R., Williams, S. B., Senagore, A. J., … Goodwin, J. S. (2018). Adapting the Elixhauser comorbidity index for cancer patients. *Cancer*, *124*(9), 2018–2025.

Menendez, M. E., Neuhaus, V., van Dijk, C. N., & Ring, D. (2014). The Elixhauser comorbidity method outperforms the Charlson index in predicting inpatient death after orthopaedic surgery. *Clinical Orthopaedics and Related Research*, *472*(9), 2878–2886.

Menendez, M. E., Valentin, N., van Dijk, C. N., & David, R. (2014). The Elixhauser Comorbidity Method Outperforms the Charlson Index in Predicting Inpatient Death After Orthopaedic Surgery. *Clinical Orthopaedics and Related Research*, *472*(9), 2878–2886.

Michalia, M., Kompoti, M., Koutsikou, A., Paridou, A., Giannopoulou, P., Trikka-Graphakos, E., & Clouva-Molyvdas, P. (2009). Diabetes mellitus is an independent risk factor for ICU-acquired bloodstream infections. *Intensive Care Medicine*, *35*(3), 448–454.

Minne, L., Abu-Hanna, A., & de Jonge, E. (2008). Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. *Critical Care / the Society of Critical Care Medicine*, *12*(6), R161.

Mosko, J. D., Leiman, D. A., Ketwaroo, G. A., Gupta, N., & Quality Measures Committee of the American Gastroenterological Association. (2020). Development of quality measures for acute pancreatitis: A model for hospital-based measures in gastroenterology. *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association*, *18*(2), 272-275.e5.

Mukhopadhyay, A., Tai, B. C., See, K. C., Ng, W. Y., Lim, T. K., Onsiong, S., … Phua, J. (2014). Risk factors for hospital and long-term mortality of critically ill elderly patients admitted to an intensive care unit. *BioMed Research International*, *2014*, 960575.

Mull, H. J., Borzecki, A. M., Chen, Q., Shin, M. H., & Rosen, A. K. (2014). Using AHRQ patient safety indicators to detect postdischarge adverse events in the Veterans Health

Administration. *American Journal of Medical Quality: The Official Journal of the American College of Medical Quality*, *29*(3), 213–219.

Nassar, A. P., Jr, Mocelin, A. O., Nunes, A. L. B., Giannini, F. P., Brauer, L., Andrade, F. M., & Dias, C. A. (2012). Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. *Journal of Critical Care*, *27*(4), 423.e1-7.

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262–e273.

Pai, S., & Bader, G. D. (2018). Patient Similarity Networks for Precision Medicine. *Journal of Molecular Biology*, *430*(18 Pt A), 2924–2938.

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of Global Health*, *8*(2), 020303.

Parimbelli, E., Marini, S., Sacchi, L., & Bellazzi, R. (2018). Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*, *83*, 87–96.

[PDF]National Diabetes Statistics Report, 2014 - CDC. (n.d.). Retrieved from https://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *9*, 157–166.

Piette, J. D., & Kerr, E. A. (2006). The Impact of Comorbid Chronic Conditions on Diabetes Care. *Diabetes Care*, *29*(3), 725–731.

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., & van der Laan, M. J. (2015). Mortality prediction in intensive care units with the Super ICU Learner

Algorithm (SICULA): a population-based study. *The Lancet. Respiratory Medicine*, *3*(1), 42–52.

Quail, J. M., Lix, L. M., Osman, B. A., & Teare, G. F. (2011). Comparing comorbidity measures for predicting mortality and hospitalization in three population-based cohorts. *BMC Health Services Research*, *11*(1), 1.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., … Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, *43*(11), 1130–1139.

Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M., & Van Den Berghe, G. (2007/7). Mining data from intensive care patients. *Advanced Engineering Informatics*, *21*(3), 243–256.

Sanders, D., Burton, D. A., & Protti, D. (2013). The healthcare analytics adoption model: A framework and roadmap. *Health Catalyst*. Retrieved from http://healthsystemcio.com/whitepapers/HC_analytics_adoption.pdf

Selleck, M. J., Senthil, M., & Wall, N. R. (2017). Making Meaningful Clinical Use of Biomarkers. *Biomarker Insights*, *12*, 1177271917715236.

Sharafoddini, A., Dubin, J. A., & Lee, J. (2017). Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Medical Informatics*, *5*(1), e7.

Simon, R. (2011). Genomic biomarkers in predictive medicine: an interim analysis. *EMBO Molecular Medicine*, *3*(8), 429–435.

Smith, G., & Nielsen, M. (1999). ABC of intensive care: Criteria for admission. *BMJ* , *318*(7197), 1544–1547.

Steinmeyer, C., & Wiese, L. (2020). Sampling methods and feature selection for mortality prediction with neural networks. *Journal of Biomedical Informatics*, *111*(103580), 103580.

van der Schaar, M., Alaa, A. M., Floto, A., Gimson, A., Scholtes, S., Wood, A., … Ercole, A. (2020). How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning*, *110*(1), 1–14.

Vincent, J. L., de Mendonça, A., Cantraine, F., Moreno, R., Takala, J., Suter, P. M., … Blecher, S. (1998). Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on "sepsis-related problems" of the European Society of Intensive Care Medicine. *Critical Care Medicine*, *26*(11), 1793–1800.

Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., … Thijs, L. G. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Medicine*, *22*(7), 707–710.

Wang, F., Hu, J., & Sun, J. (2012, November). Medical prognosis based on patient similarity and expert feedback. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 1799–1802.

Wang, N., Huang, Y., Liu, H., Fei, X., Wei, L., Zhao, X., & Chen, H. (2019). Measurement and application of patient similarity in personalized predictive modeling based on electronic medical records. *Biomedical Engineering Online*, *18*(1), 98.

Wells, B. J., Jain, A., Arrigain, S., Yu, C., Rosenkrans, W. A., & Kattan, M. W. (2009). Predicting 6-Year Mortality Risk in Patients With Type 2 Diabetes: Response to Paul et al. *Diabetes Care*, *32*(5), e61–e61.

Xia, J., Pan, S., Zhu, M., Cai, G., Yan, M., Su, Q., … Ning, G. (2019). A Long Short-Term Memory Ensemble Approach for Improving the Outcome Prediction in Intensive Care Unit. *Computational and Mathematical Methods in Medicine*, *2019*, 8152713.

Yande, S., Gohil, S., & Johnson, M. L. (2020). Pcn49 adapting the elixhauser comorbidity index for cancer patients by developing cancer-specific weights using seer-medicare data. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *23*, S31.

Ye, J., Yao, L., Shen, J., Janarthanam, R., & Luo, Y. (2020). Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Medical Informatics and Decision Making*, *20*(Suppl 11), 295.

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. *ICML*, *3*, 856–863. aaai.org.

Yun Chen, & Hui Yang. (2014). Heterogeneous postsurgical data analytics for predictive modeling of mortality risks in intensive care units. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2014*, 4310–4314.

Zhang, P., Wang, F., Hu, J., & Sorrentino, R. (2014). Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2014*, 132–136.

Zhang, Y., Luo, H., Wang, H., Zheng, Z., & Ooi, O. C. (2020). Validation of prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among cardiac-, thoracic-, and vascular-surgery patients admitted to a cardiothoracic intensive care unit. *Journal of Cardiac Surgery*, *35*(1), 118–127.

Zhou, L., & Hripcsak, G. (2007). Temporal reasoning with medical data—A review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, *40*(2), 183–202.