# A Quantitative Approach to Patient Risk Assessment and Safety Optimization in Intensive Care Units



by

Yiqun Hu

B.S., University of Michigan, Ann Arbor, 2013

Submitted to the Center for Computational Science and Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017



Signature redacted

Author .....
Center for Computational Science and Engineering
July 10, 2017

Signature redacted

Certified by .....
Retsef Levi
J. Spencer Standish Professor of Operations Management
Thesis Supervisor

Signature redacted

Accepted by .....
Youssef M. Marzouk
Co-Director, Computation for Design and Optimization

# A Quantitative Approach to Patient Risk Assessment and Safety Optimization in Intensive Care Units

by

Yiqun Hu

Submitted to the Center for Computational Science and Engineering
on July 10, 2017, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

## Abstract

Health care quality and patient safety has gained an increasing amount of attention for the past two decades. The quality of care nowadays does not only refer to successful cure of diseases for patients, but a much broader concept involving health care community, inter-relationships among care providers, patients and family, efficiency, humanity and satisfaction. The intensive care units (ICU) typically admit and care for the most clinically complex patients. While much effort has been put into patient safety improvement, the critical care system still continuous to see many human errors occur each day, despite the fact that people who work in such environment have received exceptional training. Traditional interventions to mitigate patient harm events in ICU generally focus on individual patient harms, and highly underestimate the overall risk patient face during their stay. This thesis aims to establish a new framework that more accurately account for patient risk and is capable of providing recommendations for operational decision making in launching intervention strategies that improve care quality and patient safety.

Our approach is based on theories regarding the underlying causes of human errors and a system engineering as well as analytics perspectives. We use various statistical methodologies to output rigorous but clinically intuitive insights. The core concept is to study and utilize how system-level conditions, including both human and environmental factors, can affect the likelihood of harm events in ICU. These can hopefully be used to reduce patient harms and promote patient safety by eliminating unfavorable conditions that are in higher correlation with these events, or promoting safe conditions.

We first create a quantitative metric to assess the total burden of harm that patients face, including both high frequency harms, which are typically measured in ICUs today, as well as harms that can bring highly negative outcomes to the patient but ignored due to low frequency. It is an aggregated measure that aims to reflect the true risk level in the ICUs. Then, unlike the traditional approach that motivates

intervention strategy to specific harms, we depend on the concept of risk drivers, which describe relevant ICU system conditions, and investigate what drivers affect the probability for harm events in the ICU. These conditions are defined as *Risky States*, and suggested by the model for elimination to avoid a variety of consequent risk and improve patient safety. The underlying assumption is that the same risk drivers (risky state) may affect many harms. Finally, we propose a new ensemble statistical learning algorithm based on regression trees that is not only powerful in examine the relationship between drivers and outcomes, but also being descriptive defining the risky states.

The framework was applied to the retrospective data of 2012 and 2013 from 9 ICUs at the Beth Israel Deaconess Medical Center (BIDMC), with both clinical and administrative records of more than ten thousand patients. Based on our analysis, we see a strong evidence that system conditions are associated with harm events, which include, for example, ICU patient flow (e.g., how many patients are admitted to and discharged from unit), patient acuity level, nurse workload, and unit service type, etc. The model is capable of providing insights such as "when a medical unit has more than 3 newly admitted patients during a day shift, its risk level is approximately 35% higher than the average day shift risk levels in medical units", which can motivate decisions such as assigning a new patient to some other medical unit when the current one has already admitted 3 patients during the shift, in order to avoid the above risky state from occurring.

The model output is further presented to BIDMC experts for validation through a clinical perspective. It is also being implemented and integrated to BIDMC ICU tablet application to provide guidance to ICU staff as an alerting system. The Risky State framework is unique in its innovative approach to assess patient risk and capability to offer leverage for overall patient safety improvement, and at same time designed to be compatible and spreadable with different hospital settings.

Thesis Supervisor: Retsef Levi
Title: J. Spencer Standish Professor of Operations Management

# Acknowledgments

I would first like to thank my thesis advisor Prof. Retsef Levi for his continuous support through my journey at MIT. His patience, motivation and immense knowledge has led me to the successful completion of this work. Thank you for your precious academic advice that steered me to the right direction, for the countless times you inspired me with new ideas, for your recommendation letter that enables me to advance my study at MIT and your caring and willingness to help at all times.

I would also like to express my gratitude to Dr. Pat Folcarelli, Dr. Daniel Talmor, Dr. Jennifer Stevens and Dr. Victor Novack for their passionate participation and input to this project, and for their expertise always guiding me to stay on the correct path. Without assistance from them, it would be impossible to conduct this research. I am also grateful for the enthusiastic support from the nursing staff at BIDMC, Lynn Mackinson, Veronica Kelly and Juliann Corey, who helped me learn tremendously about ICU and are always ready to help.

My sincere thanks to my project colleagues Adam Traina and Yiyin Ma. Their insightful comments and simulating discussions have enlightened me to accomplish this project. I appreciate the seamless collaboration from Adam, which inspired my teamwork spirit, and the thoughtfulness and backing from Yiyin that helped me overcoming many obstacles.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Yiqun

July, 2017

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Beth Israel Deaconess Medical Center

On February 5, 1896, New England Deaconess Hospital was opened in Boston, MA by religious women dedicating themselves to the care of the sick and the poor from an outgrowth of the Methodist Deaconess movement. In 1916, Beth Israel Hospital opened its facility in Roxbury, MA, with an initial targeted population of growing suburban Jewish due to language barriers and lacking in kosher food in other hospitals. Advancing in patient care, medical education and research in both organizations and with a success relationship formed with Harvard Medical School, two neighboring hospitals merged in 1996 to maintain their leading position in today's rapidly changing health care environment. As one of the top four recipients of biomedical research funding from the National Institutes of Health, BIDMC's research funding totals nearly $200 million annually with more than 850 active sponsored projects and 200 clinical trials.

BIDMC has earned numerous awards for its excellence in patient care quality improvement including:

**Truven Healthcare "Top 100 Hospitals":** recognition as one of 15 top academic medical centers nationally,

**"A" in Leapfrog's First-ever Hospital Safety Grades:** one of the top 65 hos-

pitals nationally,

**Society of Critical Care Medicine's 2010 Family-Centered Care Award:** award for innovation to improve care quality for critically ill patients and their families (one winner per year),

**American Hospital Association - McKesson "Quest for Quality" Award:** award for national best hospital demonstrating progress among all six dimensions of quality as defined by the Institute of Medicine.

In 2011, BIDMC launched the Center for Health Care Delivery Science (HDS) with its mission to "lead the medical center's efforts in applying rigorous, high-quality science to the evaluation of real-world innovations aimed at improving the quality, safety and value of health care". HDS collaborates with various BIDMC investigators who possess research interests in scientific evaluation of operational innovations to conduct projects that aim to improve the value of health care.

## 1.2 BIDMC ICUs

An Intensive Care Unit (ICU), is a special department of a hospital or healthcare facility that takes care for patients with severe and life-threatening illnesses and injuries, who require constant, and close monitoring, and support from specialized equipment and medications in order to ensure normal bodily functioning. ICU patients are cared by a team of extensively trained doctors and nurses specializing in critical care. The clinical condition of the patients requires a much higher staff-to-patient ratio. ICU has exclusive access to advanced medical resources and equipment and usually treat conditions, such as trauma, multiple organ failure and sepsis [34].

BIDMC has 9 adult intensive care units, specializing in 7 different areas, with a total of 77 ICU beds, as summarized in Table 1.1 below.

Units with the same prefix intuitively offer the same service type and are similar in nature. For instance, SICU-A and SICU-B (CVICU-A and CVICU-B) are on the same floor, taking mostly surgical patients and share the same care team. However, it is possible for the suffix to make a difference, as of MICU-6 and MICU-7, which are

| Symbol | Unit Name | Services Offered to | Capacity |
|---|---|---|---|
| FICU | Finard ICU | Medical and general critical care patients | 12 |
| TSICU | Trauma Surgical ICU | Trauma and surgical patients | 10 |
| SICU-A | Surgical ICU A | Surgical patients | 8 |
| SICU-B | Surgical ICU B | Surgical patients | 7 |
| MICU-6 | Medical ICU 6 | Medical patients | 8 |
| MICU-7 | Medical ICU 7 | Medical patients | 8 |
| CVICU-A | Cardiovascular ICU A | Cardiovascular patients | 8 |
| CVICU-B | Cardiovascular ICU B | Cardiovascular patients | 8 |
| CCU | Coronary Care Unit | Heart patients | 8 |

Table 1.1: BIDMC ICU Service Coverage and Capacity

on different floors, have two separate care teams during each shift and take care of patients under different medical condition categories. Therefore, SICU-A and SICU-B, CVICU-A and CVICU-B will be modeled as one big unit, while MICU-6 and MICU-7 will be modeled separately in this project.

Patients could be admitted to ICU from post-anesthesia care units (PACU), emergency department, regular floor wards, or directly from outside the hospital. Therefore, it is very possible that patients get to assign to a unit that does not regularly offer the relevant service type due to capacity limits. These patients are defined as boarders and will be further discussed in Chapter 3.

## 1.3 Gordon and Betty Moore Foundation

Gordon and Betty Moore Foundation is a grant-making foundation established in 2000 by Gordon and Betty Moore to generally promote environmental conservation, scientific research, higher education, and *patient care*, in order to make significant impact and create positive outcomes for our world and future generations. For the last 15 years, the Moore Foundation has awarded more than 2,000 grants, totaling more than $3 billion dollars.

In September 2013, BIDMC's proposal in "Optimizing ICU Safety through Patient Engagement, System Science and Information Technology" received a grant award of $5.3 million dollars through the Gordon and Betty Moore Foundation. The project contains three different work streams including:

**MyICU:** create a bi-directional interface that pushes relevant information between clinicians and family in patient portal,

**Risky States:** identify correlation between environmental condition and patient harm with goal of better understanding risk and improve patient safety,

**Context Sensitive Checklist:** establish a checklist system that is automatically adjusted based on unit conditions.

The work summarized under this thesis is part of the second work stream, which aims at supporting the hospital management team with decision making related to improving care quality and patient safety through a recommendation system that is built upon an innovative and quantitative risk assessment methodology.

In order to adopt new innovations and create a reliable system of care that is broadly scalable and spreadable to various medical settings, BIDMC is working closely with the Libretto Consortium to conquer barriers from different perspectives. The Libretto ICU Consortium, a group of four member hospitals including Beth Israel Deaconess Medical Center, Brigham and Women's Hospital, Johns Hopkins University Hospital and University of San Francisco Medical Center, oversees several grants awarded by the Moore Foundation as part of its initiative to create new architectures for the healthcare systems that offer better care quality, reduce patient harms and related costs. Since these grants share similar goals while taking different approaches, the Consortium have decided to establish some standard definitions to systematically evaluating the work of different centers and allowing horizontal comparisons. Table 1.2, for example, is a list of harms that are conventionally measured in ICUs, whose formal definitions are re-investigated and set by the representatives from each other hospitals.

[1]VAC: ventilator associated condition; IVAC: infection-related ventilator associated complication; VAP: ventilator associated pneumonia; PVAP: possible VAP; ProbVAP: probable VAP

| | Harm |
|---|---|
| i. | Central Line Associated Bloodstream Infections (CLABSI) |
| ii. | Ventilator Associated Events (VAE), incl. VAC, IVAC, PVAP and ProbVAP [1] |
| iii. | Deep Venous Thrombosis (DVT) and Pulmonary Embolism (PE) |
| iv. | ICU-acquired delirium and weakness |
| v. | High tidal volume and Acute Respiratory Distress Syndrome (ARDS) |
| vi. | Loss or diminution of respect and dignity afforded to patients and families |
| vii. | Inappropriate care and excessive intensity of care |

Table 1.2: Consortium Standardized Harm List

## 1.4 ICU Harm Events

The assumption in this thesis is that *harm events* is a broad concept that includes any undesired circumstances occurring to patients, regardless whether a real harmful consequence occurs or not. As the conventionally measured ICU harm events are known to only capture a small fraction of the total harm event that patients face everyday, BIDMC is motivated to expand this list and develop a broad concept of harm event in ICU, called the *Total Burden of Harm.* In particular, the goal is to capture other harms that may be relatively infrequent but can add up for a large number.

According to *To err is human: building a safer health system*[11], the famous report published by the Institute of Medicine in 1999, up to 98,000 patients die because of human errors in U.S hospitals each year. ICU patients experience many of these events, partially due to "the complexity of their conditions, need for urgent interventions, and considerable workload fluctuation"[8] and it is estimated that each patient experience 1.7 medical errors per day in ICUs [4]. In a study carried out at three ICUs in a large, urban teaching hospital affiliated to a university medical school by Lori Andrews et al., 45.8% percent of ICU admissions were associated with a patient harm event [1] and nearly all suffer a potentially life-threatening error at some point during their stay [26]. The past decade has seen many publications that increase the awareness and concern of unsatisfactory health care quality and patient safety across the country. However, despite the impact of these errors on patients, care quality, and cost, there is a limited number of researches that have identified the

human factors or system conditions that contribute to errors in the ICU.

The Risky States project aims to introduce a quantitative approach to assess the overall patient risks in ICUs and model the relationship between the likelihood of harm events and the *states* of ICUs - including its environment, the systems in the ICU and people in the ICU. The term "risky states" refers to conditions associated with a given ICU at a given shift that are likely to increase the likelihood of harm events and therefore increase the overall risk level. This model aims to bring several benefits to BIDMC with:

- Leverages for harm event prevention by avoiding, mitigating and eliminating risky states through decision making, or alternatively identify safe states
- An alerting scheme for awareness of increased likelihood of harm events
- System level mitigation strategies to the ICU risk management that can be easily adopted by other institutes.

## 1.5 Summary of Approaches and Insights

In this thesis, we first define a concrete way of measuring the *Total Burden of Harms* for patient in ICUs. It is an aggregated notion of all harm events occurred in the critical care unit environment. More precisely, we expand the definition of harm events by aggregating over all types of harm events rather than adopt a measure of harm based on some specific care plan. This does not only leads to a better understanding of where patient risk exists, but also increases the statistical power when use data to identify risk sources.

In addition, the Risky States approach models the relationship between ICU harm events - defined by the *Total Burden of Harms* above - and the states of system, the environment of the system, and the people in the system. While this certainly allows us to identify risky states, which are system conditions that will increase the likelihood of patient harms and potentially the magnitude of their outcomes, more importantly, these conditions are common to many different harm events. Therefore, by eliminating the occurrences of these risky states, we are hopefully able to reduce

the probability of a variety of harms simultaneously.

In order to identify risky states derived from statistical models, we need to build one that is more than just a predictive black-box. We choose to use regression tree because of its high interpretability by dividing shifts into clusters and characterizing them by descriptions in all ancestor nodes. However, regression trees are known to be very sensitive to the training data and thus causing overfitting. Therefore, we use random forests - an ensemble learning method that combines many regression trees - to decrease the variance of the prediction. Since random forest is not an interpretable approach, we developed an algorithm that can provide a relatively stable regression tree using data simulated from the original dataset and corresponding response variable calculated using random forest. This algorithm does not only have low variance when making predictions but is also capable of providing valid interpretations and descriptions of risky or safe states. Finally, we test the stability and significance of potential risky/safe states using the $k$-nearest neighbor algorithm and Mann-Whitney-Wilcoxn test.

The analysis identifies several combinations of environmental conditions of ICUs, called states, that are correlated with high rate of harm events. For instance, in surgical ICUs if during a day shift, patients are quite sick and there are more than 2 newly admitted patients, then there is a higher likelihood that patients in this unit and shift will experience a harm event. In general, it shows that:

- Surgical and medical units are generally different because they treat very different types of patients
- CVICU usually carries a high risk for harm events than other units due to the type of patients they admit
- Day shifts in general have high harm rate than night shifts
- Patient acuity is not always correlated with higher likelihood of harm events
- Higher nurse utilization often seems correlated with high rate of harm events

## 1.6 Outline

Chapter 2 provides a review of prior work on patient safety and risk management practices through interventions for individual harm event. It discusses the differences between those conventional approaches and the new Risky States approach that utilizes the concept of event drivers as well as aggregated metrics for describing ICU harms and system conditions.

Chapter 3 provides a precise definition of *Total Burden of Harms* in our project, including how to derive harms from the BIDMC information system and how to aggregate them into unit and shift level. It also explains in detail why aggregation of harms is the key concept in the Risky States approach.

Chapter 4 describes the proposed risk drivers in various aspects and how to quantify them. Drivers are also aggregated from patient level to unit and shift level, so they match accordingly with harms defined in Chapter 3, and together they form input pairs into our statistical model developed in succeeding chapters.

Chapter 5 describes the statistical model we developed to identify risky states. The process starts with feature selection, and moves on to a single regression tree and finally random forests. It then illustrates how to integrate and interpret the model outputs using human-understandable risky/safe states. Finally, it introduces the $k$-nearest neighbor algorithm and the Mann-Whitney-Wilcoxn test as two checks for model output stability and significance.

Chapter 6 explains how the algorithm in Chapter 5 is applied to the retrospective data from BIDMC in the calendar years of 2012 and 2013. It summaries our findings using the Risky States approach, followed by a detailed list of risky states and safe states.

Chapter 7 concludes wit a summary of our project and suggestions for future research.

# Chapter 2

# Background and New Approach

In this chapter, we will first go over the existing practices to measure harm and manage risk. We will then discuss some of its underlying problems and how our new approach is motivated by them, followed by the key concepts in our approach and their advantages.

## 2.1 History of Health Care Quality and Patient Safety

Starting in the 19th century, a variety of events helped shaping up the foundation of quality of care metrics and concepts that we see today. This includes, for instance, the study of mortality rates during the Crimean war in the 1850s, outcome measurements for surgical interventions established in 1912, medical auditing through patient chart review advanced by the Joint Commission on Accreditation of Hospitals (JCAH) around the 1950s, just to name a few of them [8]. Healthcare quality and patient safety started to draw the public's attention in the early 2000's when a series of widely publicized reports, such as *Crossing the Quality Chasm: A New Health System for the 21st Century* [23] and *To Err is Human: Building a Safer Health System* [11] got published. Over the last decade, government and medical societies has worked mostly through new regulations to establish more elaborated and transparent systems to monitor and report on patient harms, as well as safety and quality issues. This has been a trend in many countries. Australia and the United States lead the trend

in 2000 [31, 24], followed by the United Kingdom in 2003 [18] and France in 2006 [8]. In 2002, JCAH started to require accredited hospitals to report standardized quality measures for conditions, such as heart failure, acute myocardial infarction, pneumonia and pregnancy. They also developed six national measures of quality for ICUs, which correspond to a proposal earlier from Palmer and Adams calling for "various dimensions of measure of care quality" [25].

Health care quality metrics nowadays have expanded over the years to include factors such as satisfaction of patients and their families, efficiency, humanity in care and quality of care team work. Patient safety, as one important factor that defines health care quality, is typically considered as the avoidence from accidental injuries to patients as a result of failure to complete a plan as intended or a misuse of an improper plan for goal achievement, referring to errors of commission and errors of omission, respectively [11]. Even though much effort has been put into the standardization of practices related to care quality and patient safety, this has not yet led to great improvements. HealthGrades reported that more than 0.2 million deaths were due to avoidable errors between 2004 and 2006 [13], and approximately 150 thousand life-threatening errors occur in US teaching hospital ICUs annually [30].

In the year of 2012 and 2013, BIDMC recorded more than two thousand patient harm events associated with their ICU patients, including harms from the conventionally measured harm list and various events voluntarily reported by staff through the BIDMC Incident Reporting System. However, researches have shown that the reported number of patient harms is in general tremendously underrepresented comparing to the overall harm that patients face [35]. Attempting to address this issue, we seek to develop a far more comprehensive notion of patient harm to more accurately capture the extent of which harms occur in the ICU environment.

## 2.2 Traditional Approaches for Quality Assurance and Risk Management in Healthcare

There are two basic approaches, conceptually, adopted in ICUs for the purpose of care quality and safety measurement. The first one is called the *room-for-improvement* model. In this model, when a new problem occurs, plans are created to address the problem and then executed; after observing and assessing the effectiveness of the plans, new action plan might be put into effect based on what is learned. This approach is known as the *Plan-Do-Act Cycle* (PDAC) of the Institute for Healthcare Improvement (IHI) [14]. The second way is to evaluate quality indicators by consistently monitoring key factors, as fluctuations in glucose levels, and the number of unplanned extubations [40]. These two approaches are often used in parallel and the latter serves as a source for identifying rooms for improvement, which will then initiate a PDAC.

The traditional patient safety and risk management practices are harm-driven as they were developed aiming to reduce the frequency of occurrence of common pre-defined harmul events, such as *Central Line Associated Bloodstream Infections (CLABSI)*, *Ventilator Associated Events (VAE)*, *Deep Venous Thrombosis (DVT) and Pulmonary Embolism (PE)*, *ICU-acquired delirium and weakness*, and *High tidal volume and Acute Respiratory Distress Syndrome (ARDS)*, as mentioned in Table 1.2 in Chapter 1. These harms are consistently monitored because they either cause severe consequences and have direct impact on patients, or they happen very frequently in ICUs. In addition, harms are always measured at patient level in current risk models. Rareness of specific harms for specific patients makes it hard to obtain robust statistical relations between the related causes and consequences.

Numerous efforts have been invested to design and implement specific mitigation plan or intervention strategies for specific patient harms, in order to decrease the probability of occurrence or alleviate the corresponding negative outcomes. National government as well as local hospitals often launch surveillance systems to monitor quality metrics, and it creates a self-assessment scheme for quantifying what they are

doing. We will discuss some of the common practices for monitoring quality metrics and mitigation measures adopted after identification of problems, with the shortcomings that come along.

1. The medical chart review.

Patient chart is a comprehensive summary of patient's information during his/her hospitalization, which records both the evolution of the patient's medical conditions and a complete description of the sequence of care interventions he/she received. Reviews of patient charts, such as administrative information, morbidity condition or discharge summary, can often reveal at least some of the quality factors of interest. Chart review can be done manually or electronically. However, manually reviewing the massive amount of data for an entire large patient population to obtain key information extraction is generally unrealistic. In particular, electronic review of chart data requires a mature information system architecture that at the moment is not available in most hospitals. Another limitation is biased data recording and interpretation due to variability in medical terminologies used by different physicians, spelling errors, non-standardization in how medical charts are filled, etc.

2. Voluntary reporting system.

A voluntary reporting system is widely adopted in hospitals in the United States. It provides regular feedbacks from the team of care providers who actually work in the unit and possess the most accurate information for the scene. While it is a direct and comprehensive source for error learning, it inevitably suffers from underreporting due to time constraints, complexity in filing, fear of liability for one's own mistake, varied harm event judgment and lack of changes after reporting.

3. Bedside observation.

Harm events can also be detected through direct observation at the bedside by expert personnel, and is used periodically for detecting errors by omission [12, 3]. For example, medication error collected by a pharmacist at the bedside can increase

to 560/1,000 patient-days with daily routine observation of prescriptions, comparing to 7.45/1,000 patient-days from voluntary reporting [38, 29]. However, this practice cannot be followed routinely but only for short because of the enourmous amount of workforce required.

4. Root cause analysis.

A root cause analysis (RCA) is usually performed after an occurrence of a severe harm event that brings negative consequences to patients. It is followed by a "detailed mitigation plan developed to eliminate or reliably reduce the risk of another patient experiencing the same specific harm", which is designed under general quality control concepts for single process improvement [19]. This approach is expected to be effective for the same type of harm, but is limited only to that type. In addition, this approach will not usually be triggered for harms that do not cause serious health consequences to patients even the harm occurs in moderate frequency.

5. Checklist.

A checklist is a type of informational job aid used to reduce failure caused by limits of human memory and attention. It helps to ensure consistency and completeness in carrying out a task. Checklists are first used in airline industries, but have been widely adopted in the healthcare system. They are algorithmic lists made specifically for various clinical circumstances to ensure standard process steps are completed by clinical staff. One of the best known intensive care checklist protocols was created by Dr. Peter Pronovost, which is considered having "saved more lives than that of any laboratory scientist in the past decade" according to Atul Gawande in *The New Yorker* and also awarded a MacArthur Fellowship [9].

While researches show that checklist, with a theoretical basis in principles of human factors engineering, has achieved substantial successes for patient safety improvement [26, 33], it provides hardly any help for newly emerging or unpredictable patient risks because it is too specific. It is a common dilemma whether ICU should enforce more checklists, which can potentially impose a higher risk due to the in-

creased complexity and a greater workload or burden to staff, or whether ICU should limit the number of checklists to achieve simpler workflow but suffer from a higher risk level resulted from non-standard practices carried out by clinicians.

A combination of these methods is a great start to monitor and mitigate harm events for assessing care delivery quality and optimizing risk management in the ICU. However, there are still many problems yet need to be resolved. First of all, it targets only at a subset of patient harms, mainly the ones with severe outcome and frequent occurrence. Therefore, a lot of potential harms are ignored, including harms that have low occurrences yet high impact to the patients or harms that we are not aware of, which we call "the unknown unknowns". This could cause a misleading picture of what the overall patient risk level is. Secondly, by creating specific interventions with respect to certain harm, we are making the system more and more complex as the number of harms we consider gradually increases. Finally, the corrective actions are mostly reactive and post-problem mechanisms which bring little prediction power. As it is critical to be able to launch early interventions to minimize the number of harm events and their consequences, lacking in means of understanding the current environment in a timely manner makes it difficult to design efficient preventive actions.

In order to address these problems, which cannot be quantified and reduced by the existing tools described above, we propose the Risky States approach.

## 2.3 The Risky States Approach

In the Risky States approach, there are three core concepts. First, harm events will be considered in an aggregated notion, called the *Total Burden of Harm*. The definition of harms is expanded to include all types of harm events rather than adopt a measure of harm based on some specific care plan. This does not only leads to a better understanding of where patient risk exists, but also increases the statistical power when use data to identify risk sources. Detailed discussion on how we construct the *Total Burden of Harm* is in Chapter 3.

Second, we propose the idea of risk drivers. Harm events involve uncertainty in two ways, namely uncertainty regarding occurrence, and uncertainty of the resulting outcomes if they indeed did occur. They correspond to *event drivers* and *outcome drivers*, respectively:

**Event Drivers:** conditions of the environment, the system and the people in the system, that affect the likelihood of the harm event to occur

**Outcome Drivers:** conditions of the environment, the system and the people in the system, that affect the magnitude of the outcome of the risk event if it indeed occurred

To illustrate the ideas of both drivers, consider the case of a car accident. In general, a specific accident could be due to a variety of direct reasons such as loss of control of vehicle, failure to stop or the existence of foreign object on the road. *Event drivers* in this case could be road conditions, badly maintained car or tired driver. These are examples of event drivers related to the condition of the environment, the system, and the people in the system, respectively, which increase the likelihood of a car accident to occur. On the other hand, lack of safety features in the car could affect the magnitude of the injury from the crash, and is considered as an *outcome driver.*

In our setting, risky states are defined using these risk drivers, as a combination of system conditions that when they are present, there is an increase in the likelihood of patient harms and potentially in the magnitude of their outcomes. This is a powerful approach as such risky states are common to many harms. And by eliminating the occurrences of these risky states, we are able to reduce the likelihood of a variety of harms simultaneously. For example, high workload of nurses could lead to not paying enough attention to individual patients, and resulting various harms, such as patient falling out of bed, giving the wrong medication, etc.

To utilize this framework, we need to take two key steps. First is to identify the possible drivers that are correlated with the total burden of harm. Second is to test and validate the relationship, if any, between the drivers and the harms. Detailed

derivation and definition of drivers are discussed in chapter 4.

Last but not least, both harm events and drivers are not measured at patient level, but at unit and shift level, which also adds statistical robustness to our model since the impact from some rare and special patient becomes smaller. Aggregation, in terms of harm events and in terms of measurement level, is the key idea in our Risky States approach.

With such implementation of drivers and harm events, what we get for each shift at each ICU, is a set of aggregated drivers, consisting of both patients and staffing conditions as well as environment related factors, and a set of aggregated harms, that carries clinical meanings. The objective is to then develop a descriptive model through data analysis and statistical modeling that can provide us with an extensive understanding between system factors and overall patient risks. The results will call for interventions that reduce the likelihood of harm events through eliminating risky states, thus improving patient safety. For instance, if many harms occur when the unit receives more than three admissions during a shift, we should not assign the fourth patient to that ICU but seek for other possible arrangement. This mechanism will work as a recommendation system that promotes wise operational decisions.

The idea of aggregation increases the robustness of the model in that

- It resolves the issue with rare harms or even unknown harms because it explains harm as a phenomenon, not harms of specific types;
- It passes the barriers to use many powerful statistical methods that are not applicable for rare events;
- It improves the statistical power because we now work with a sample that have more observations labeled with harm (i.e. the extreme unbalance in data is relieved);
- It is less impacted by some particular patients;
- It enhances statistical reliability from a much smaller predictor space with respect to number of observations, comparing to modeling using patient level predictors;
- It makes possible to predict the future.

In summary, the Risky States approach aims at creating a more systematic and holistic patient safety and risk management mechanism that is:

- Comprehensive and consistent, meaning it can capture many different aspects of the system and it can be adapted to different settings
- Able to incorporate rare or not pre-defined harm events
- Reflective to the risk changes in the ICU environment in a timely manner
- Capable of capturing the interdependencies among different harm events
- Communicating various stakeholders of the system
- Adapted to various care quality and risk metrics quickly

# Chapter 3

# The Total Burden of Harm

As discussed in Chapter 1, one of the underlying assumptions of the approach developed in this thesis is that different harm events occur in the context of common underlying environmental, system-level drivers, called risk states. Based on this assumption and to enable statistical predictive analysis, we desire to develop a framework to quantify the *Total Burden of Harm* in a given ICU unit on a given shift.

The BIDMC Information System is an electronic database that stores various kinds of patient information. With different focuses and functionality, these databases were able to provide us with both clinical data on patients as well as environmental information through administrative records. A complete list of the databases that were utilized in the project can be found in Appendix A.1. These were used to retrospectively collect data from calendar years 2012 and 2013, which enables us to calculate each harm event frequency, occurring time and associated locations.

In Chapter 3.1, we first describe the principles guiding us in defining specific harm events, followed by a detailed list of harm events that we consider in this thesis. Chapter 3.2 will discuss harm characteristics, and show how they help transform individual harms into the notion of the *Total Burden of Harms* in unit and shift levels. Chapter 3.3 illustrates the idea of aggregating harm events into bigger groups and how it can help adding statistical power to the risky states analysis. It also includes a brief comparison between the Total Burden of Harm and the traditionally measured harm events in ICUs.

## 3.1 Harm Events

As discussed in Chapter 1.4, *harm events* is a broad concept that includes any undesired circumstances occurring to patients, regardless whether a real harmful consequence occurs or not. The Joint Commission on Accreditation of Hospitals divide harm events in terms of patient outcomes [24]:

**Near miss:** occurrence of an error without causing adverse consequence;

**Actual harm:** occurrence of an error that resulted in real patient harm.

In the Risky States approach, we include both of these scenarios because the underlying risk drivers are in fact common to scenarios with or without measurable patient harm. In addition, the definition that is used does not distinguish necessarily between preventable and non-preventable harms. In other words, if there is a doubt regarding the preventability, we include the harm event in our measure. This is because from the patients' perspective, both will result adversarial outcomes, as well as it is likely to expose systematic problems. In particular, lowering the measurement threshold for what is considered as a harm event is likely to ultimately lead to safer environment.

In general, the *Total Burden of Harm* is the aggregation of unexpected or unplanned outcome, regardless of the severity, consequence and preventability.

Based on these principles, the *Total Burden of Harms* that we define consists of 37 different harm events (see Table 3.1). There are four different sources of data used to define and identify these harm events (with the numbers in the box indicating how many types come from the source):

- The Libretto Consortium [5]
- A modified variant of The Institute for Healthcare Improvement (IHI) Global Trigger Tool [22] [15]
- BIDMC Incident Reporting System [16]
- Recommendation by BIDMC clinicians [1]

Among the standard seven "Consortium harms" defined by members of the Libretto Consortium (see Chapter 1.2 and Table 1.2), five of them are included in our

definition of Total Burden of Harm.[1] These harm events have relatively frequent occurrences and severe adversarial patient outcomes. Moreover, high level of attention has been given to develop mitigation plans in ICUs to reduce their frequencies. Indeed, in the past couple of years, CLABSI and VAE occurrences has been dramatically reduced at BIDMC and other ICUs across the country.

Traditional efforts to detect harm events have focused on voluntary reporting and tracking of errors. The BIDMC Incident Reporting System is a voluntary tool designed for staff members to report any harm events, plus any type of concerns they have related to care providing, including the environment, the system, other staff, patients, and family. It has some well-defined harm event types, such as a fall or a skin pressure ulcer. It contains some general categories which allow for various concerns or very rare or specific events to be recorded. Hence, it is an important source of data related to potential patient harm events.

In need of a more comprehensive way to identify harm events beyond the seven that are traditional measured, a group of Institute for Healthcare Improvement (IHI) staff member consisting of clinical experts and other professionals developed the first IHI Trigger Tool in 2000 [22]. The conditions introduced in this Global Trigger Tool do not only include harm events that are associated with visible adverse consequences to patients, such as cardiac or respiratory arrest, but puts more emphasis on capturing symptoms of possibly subtle harm events. Such conditions include "near misses" or "process complications" can be highly undesirable and potentially dangerous for patients. For instance, unplanned extubation does not necessarily make a patient measurably sicker, but it could. In addition, it might result in a reintubation, which is an elevated complication to the patient and should be avoided. The IHI Trigger Tool utilizes many of patients' standard electronic records, such as their vital signs and lab results. The IHI trigger tool is traditionally applied as a manual process of chart review done by nurses on a very small sample of patients selected at random

[1]Note that "weakness" in (to be measured by Function Mobility Scale) item iv. [ICU-acquired delirium and weakness], as well as item vi. [Loss or diminution of respect and dignity afforded to patients and families] and vii. [Inappropriate care and excessive intensity of care] are not currently measured by BIDMC, thus excluded in the analysis.

around 20. As part of this work, we have automated this process by calculating these harm events using the extracted data from the hospital IT system. Thus, the process can easily be scaled up to include all patients. Indeed, we were able to apply a version of the trigger tool to over ten thousand patients over two year period.

The definition and identification of the 37 harm events was also discussed in detail by a large group of physicians and nurses at BIDMC. The team developed appropriate metrics to define from data when a harm event occurred. In addition, there was a discussion under what situations harm events should be associated with the ICU. For example, there could be scenarios in which an harm event is presented when the patient is in the ICU but because of activities occurred prior to the arrival of the patient to the ICU. In such case, this harm will not be considered as an ICU harm event.

BIDMC clinicians also recommended including harm events that are not yet captured by these three sources. These four sources complement each other well to provide a relatively good measure for the Total Burden of Harm in ICU. Table 3.1 lists all the 37 harm events we consider within the scope of this thesis. The first column gives the harm name, the second source, and the third a short description.

Table 3.1: Harm Descriptions

| **Harm** | **Source** | **Description** |
|---|---|---|
| CLABSI | Consortium | Central line associated blood stream infection |
| VAE | Consortium | Ventilator associated events |
| DVT-PE | Consortium | Venous thromboembolism |
| ARDS | Consortium | High tidal volume exceeding patient's ideal range |
| Delirium | Consortium | ICU acquired delirium |
| CAUTI | BIDMC | Catheter associated urinary tract infection |
| Code Blue | IHI Trigger | Cardiac/respiratory arrest |

Continued on next page

Table 3.1: Harm Descriptions

| Harm | Source | Description |
|---|---|---|
| Positive blood culture | IHI Trigger | Possible infection |
| Positive C. difficile | IHI Trigger | Possible infection |
| Oversedation | IHI Trigger | Overdose of sedative drips |
| Abrupt drop in Hemoglobin | IHI Trigger | Possible internal bleeding |
| Bleeding | IHI Trigger | Bleeding caused by overdose of Warfarin, measured by INR |
| Bleeding | IHI Trigger | Bleeding caused by overdose of Heparin, measured by PTT |
| Hypoglycemia | IHI Trigger | Low blood glucose caused by overdose of Insulin drip |
| Administer Vitamin K | IHI Trigger | Reversal for overdose of Warfarin |
| Administer Naloxone | IHI Trigger | Reversal for overdose Narcotics |
| Chest tube insertion at bedside | IHI Trigger | Possible Iatrogenic Pneumothorax |
| Doubled creatinine | IHI Trigger | Kidney failure |
| Readmission | IHI Trigger | Possible uninformed decision at discharge, unnecessary complexity |
| Reintubation | IHI Trigger | Possible uninformed decision at extubation, unnecessary complexity |
| Unplanned extubation | IHI Trigger | Extubation issued by patients |
| Fall | IRS | Patients fall out of bed |
| Medication error | IRS | Any type of medication error |
| Skin tissue, infection | IRS | ICU acquired pressure ulcer |
| Handoff, communication, service coordination | IRS | Breakage of important information that results undesired outcomes |
| Lab specimen | IRS | Mislabeling or other reasons caused invalidation of drawn sample |

Continued on next page

Table 3.1: Harm Descriptions

| Harm | Source | Description |
|---|---|---|
| Code Purple | IRS | Coded safety issue |
| Line, tube, vascular access, drain | IRS | Issues related to these types of device |
| Diagnosis, treatment, testing | IRS | Issues related to plan of care |
| Airway manage | IRS | Issues with respiratory aiding device |
| Safety | IRS | Conditions that call for security |
| Nutrition | IRS | Diet related issues |
| Identification | IRS | Misidentification |
| Blood, blood product | IRS | Possible sign for overdose of anticoagulants |
| Environment | IRS | Issues observed by staff about environment concern, equipment failure and supply shortage |
| Consent | IRS | Issues related to plan of intervention or surgery and consents with patients and family |
| Restraints | IRS | Patients escape from restraints needed for their intubation |

## 3.2 Harm Characteristics

In this section, we discuss some characteristics of harm events that are later used in the analysis for the following purpose:

1. Convert the absolute number harm event to a *harm event rate* with appropriate normalization
2. Focus the predictive analysis on a more homogeneous set of harm events that are more likely to have common underlying risk drivers

These characteristics include:

**Relevance Patient Cohort:** Certain harm events could only affect a subset (cohort) of the patients in the ICU (e.g., only ventilated patients will have the risk of encountering VAE harms). This is important to obtain appropriate normalization and convert the measurement of harm into a rate. For example, 2 VAE harm events in a unit with 4 ventilated patients is worse situation then 3 VAE harm events with 8 ventilated patients. Indeed, we say a rate of $\frac{1}{2}$ VAE per qualified patient in the former unit is considered as riskier than $\frac{3}{8}$ VAE per qualified patient in the latter.

**Type:** This indicates whether a harm event is caused by environmental conditions occurred in the same shift. *Instantaneous* harm events have precise timestamps and usually occurs due to some ongoing conditions in the neighboring time window (e.g., medication error is due to an error of a nurse at the very moment and is unlikely related to the shift before). In contrast, *long-term* or *evolving* harms take time to develop and it is usually impossible for us to say when exactly the harm event occurred, or what conditions at what time caused it to occur (e.g., an infection can gradually evolve for several days, and it is hard to say exactly the infection happens).

Based on these characteristics, we are able to calculate the *average harm event rate per patient* in a given unit $u$ during a given shift $j$:

$$r_j^u = \frac{1}{n} \sum_{i=1}^{n} \frac{N_{ij}^u}{\mathtt{RP}_{ij}^u} \qquad (\star)$$

where $n$ is the total number of harm event types taken into consideration, $N_{ij}^u$ is the number of occurrences of harm event of type $i$ in the unit $u$ during the shift $j$, and $\mathtt{RP}_{ij}^u$ is the number of patients relevant to harm event $i$ presented in unit $u$ during shift $j$.

The ratio $\frac{N_{ij}^u}{\mathtt{RP}_{ij}^u}$ calculates the *harm event rate per patient* for a specific type of harm event $i$ in shift $j$ at unit $u$. We sum up over all $i$ and divide the result by the total number of types $n$ to obtain the *average harm event rate per patient.*

Note that the difference in nature for harms with different *Type* makes counting the actual number of occurrences of an harm event ($N_{ij}^u$) different. Whereas it is easy

to associate instantaneous harm to the shift when it occurs (e.g., if 2 of instantaneous harm event of type $i$ occur during shift $j$ at unit $u$, then $N_{ij}^u = 2$), an additional step is needed to associate evolving harms to appropriate shifts. Since no definite conclusion can be made regarding when an evolving harm actually happens, it is suggested by BIDMC experts to evenly attribute an evolving harm to all the shifts that bare a positive probability that the harm may happen, starting from the "first eligible shift", referring to either the shift when the patients start to carry the risk for a harm event or the admission shift, which ever is later. For example, a central line infection can happen during any shift between the insertion of the central line in the ICU and the identification of an infection by Infection Control at BIDMC. If there are 5 shifts in between, then the number of central line infection harm is 0.2 for each of the 5 shifts. Therefore, we alter equation ($\star$) slightly to obtain our final way of calculating *average harm event rate per patient*:

$$r_j^u = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\mathrm{RP}_{ij}^u} \sum_{k=1}^{N_{ij}^u} \frac{1}{\#\text{ Attributed shifts}_k} \qquad (\star\star)$$

If all harms are instantaneous harms, we have $\sum_{k=1}^{N_{ij}^u} \frac{1}{\#\text{ Attributed shifts}_k} = N_{ij}^u$, which is the same as ($\star$). If some harms are evolving, only a fraction of those harms will be added to the total burden of harm for the given shift.

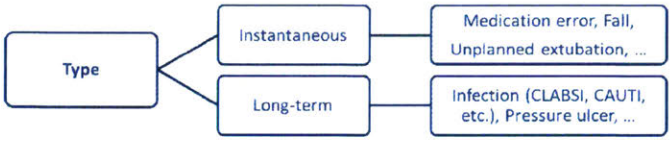
Table B.1 in Appendix B offers a detailed list of harm characteristics for each harm, including the two discussed in this section, as well as their definitions, metrics, and shift attribution.

## 3.3 Harm Aggregation

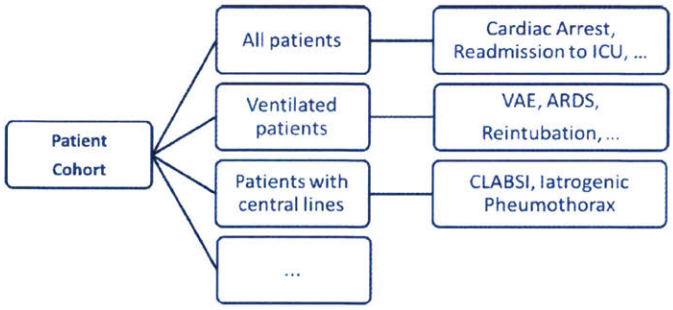
In this section, we propose several ways to partition the harm events into aggregated groups that are more likely to have common underlying risk drivers. The goal is to increase the statistical power of predictive risk models.

The three different ways of aggregation we propose include 1) by *Type*, 2) by *Relevance Patient Cohort*, and 3) by *Process Relevancy*, as illustrated in Figure 3-1. Type

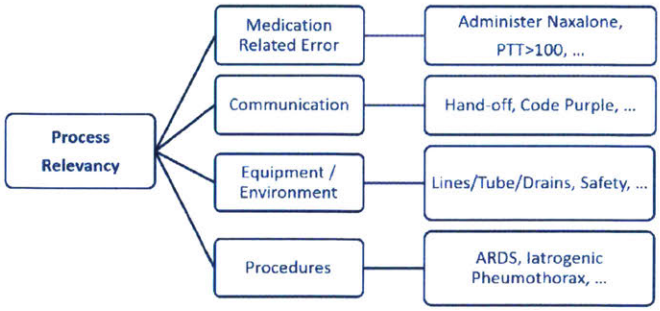
divides the Total Burden of Harm exhaustively into two groups with one containing all the *instantaneous* harms and the other *long-term/evolving* harms (Figure 3-1a). This is particularly crucial in the succeeding modeling steps because instantaneous harms, associated with one single shift, are likely caused by the factors of the same shift, so we can model this using a direct match between drivers and harm events for that shift and assume no correlation with the others. Evolving harms are usually attributed across multiple consecutive shifts, so we need to take a different approach to incorporate this into a model by considering the states of the units over multiple shifts.




(a) Harm Groups by *Type*




(b) Harm Groups by *Patient Cohort*

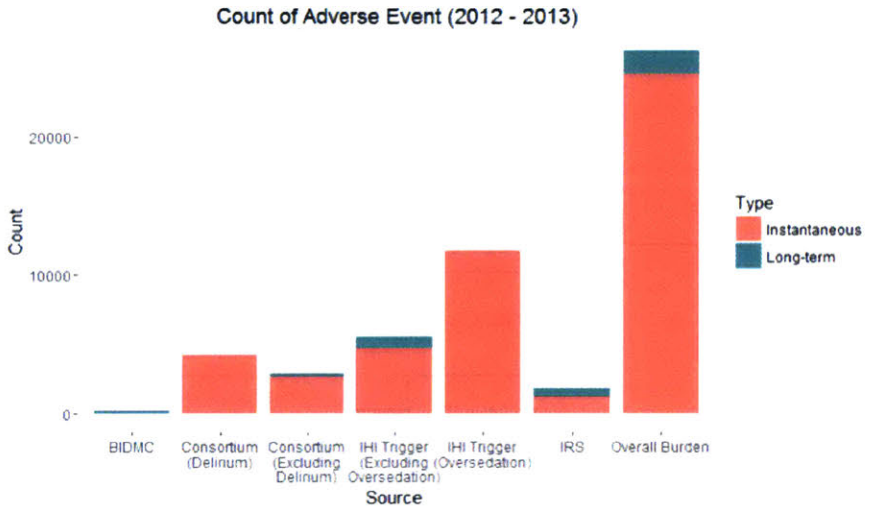



(c) Harm Groups by *Process Relevancy*

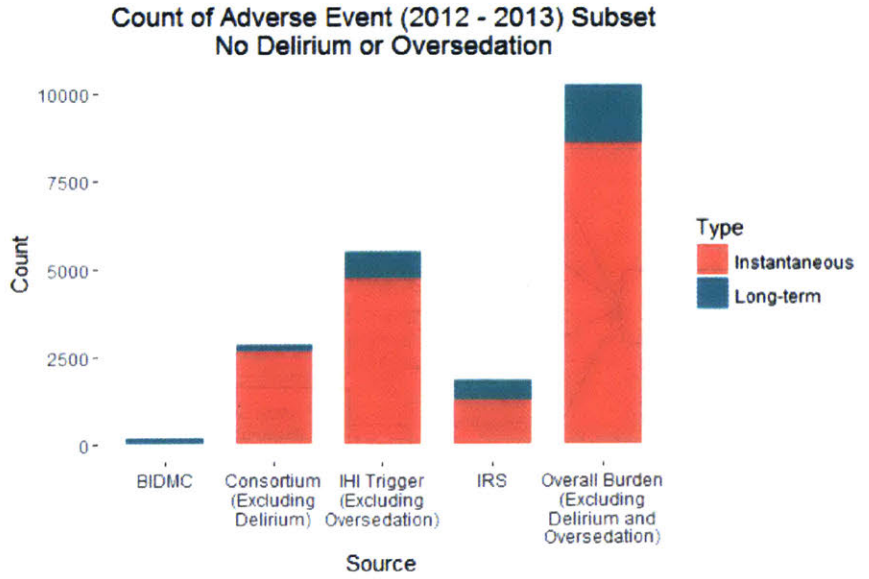
Figure 3-1: Harm Groups

The second way of aggregation depending on *Relevance Patient Cohort* puts clin-

ically similar patients into the same group (Figure 3-1b), which could catch possible dependencies of harm events on type of patients that might not be captured using *Type*. Figure 3-1c provides harm group rules based on process relevancy, specifically the similarities and differences in terms of actions taken or equipment used. For example, medication errors are probably affected by similar drivers (e.g., high workload caused distraction), while communication related harm events might be caused




(a) Harm Volume: Total Burden of Harm v.s. Conventional




(b) Harm Volume: Total Burden of Harm v.s. Conventional (Subset)

Figure 3-2: Harm Volume: Total Burden of Harm v.s. Conventional

by another set of conditions (e.g., unfamiliar with standard handoff protocol). The latter two proposals are not necessarily exhaustive as some harm events may possess very special characteristics that would not put it into any group.

However, there are many more ways of aggregating harm events as long as the groups bear clinical meaningfulness or possess operational similarities. In this thesis, we will model the Total Burden of Harm by adding up *all instantaneous* harms with respect to unit and shift level.

Figure 3-2 shows that the number of harm events from January, 1st, 2012 to December 31st, 2013 at BIDMC is four times larger than what we would obtain using a conventional approach, which only includes the Libretto Consortium harms. Each bar except the last one represents the total number of occurrences of harm events collected from the corresponding source indicated on the $x$-axis, while the last bar is the sum of all the others. Figure 3-2b plots a subset of harm events, excluding delirium and oversedation, comparing to Figure 3-2a. While these two harm events are well defined through standard metrics, it is perceived by BIDMC clinicians that the numbers we obtained do not agree with their understanding of the system and need

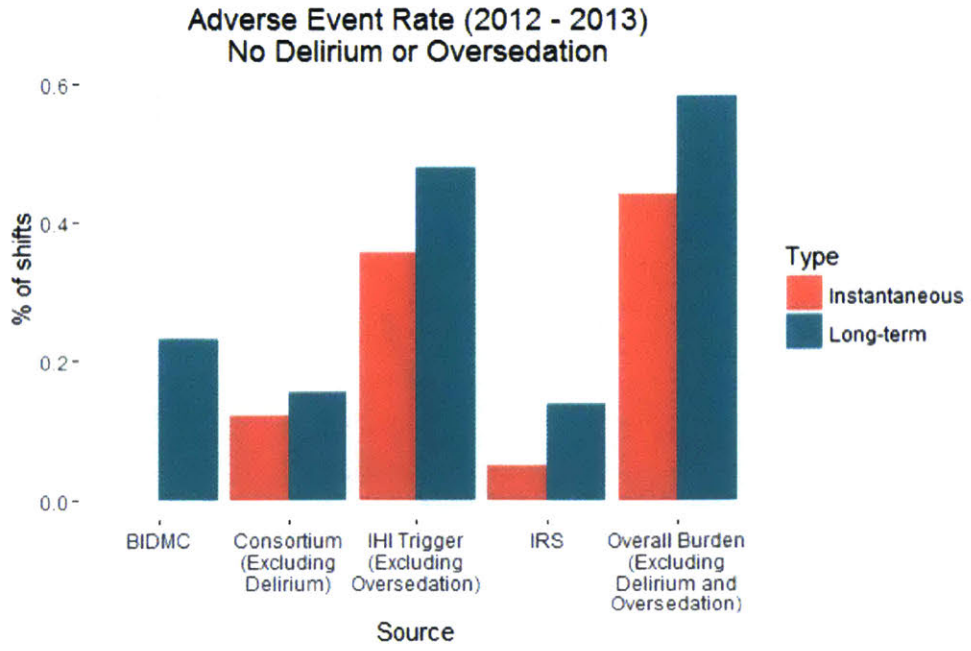



Figure 3-3: Shift Harm Rate: Total Burden of Harm v.s. Conventional

further investigation. Therefore, they will be excluded in the analysis presented in the subsequent chapters. Nevertheless, both present a roughly four-time relationship between the traditional approach and the Total Burden of Harm. This suggests that current harm measurement approaches are very limited and do not provide a reliable picture of the overall harm in the ICU environment.

Figure 3-3 shows how many shifts in the 2-year range observed at least one harm event, from different sources. We can also see an elevated figure in the Total Burden of Harm approach, which will help alleviate the unbalancedness of our data as inputs to the model and grant for a higher statistical power.

# Chapter 4

# Risk Drivers

In this chapter, we will describe the set of risk drivers that will be input to our models. The development of the set of hypothesized drivers was done in collaboration with a multidisciplinary team from MIT and BIDMC. Study of human errors and case review of the thirty past adverse events enabled us to hypothesize a set of possible factors that could impact the likelihood of harm events in ICUs based on instinctive, experiences and perspectives. These factors are called *risk drivers*, describing the states of system, the environment of the system and the people in the system, which will affect the likelihood of harm events and/or the potential likely magnitude of their outcomes.

Prior to discussing the specific risk drivers hypothesized with respect to the ICUs at BIDMC, we provide relevant backgrounds.

## 4.1 Backgrounds

The motivation of proposing risk drivers starts from understanding human errors in general. Human errors usually refers to actions "not intended by the actor; not desired by a set of rules or an external observer; or that led the task of system outside its acceptable limits", which deviates from intention, expectation and desirability [32]. Integrated framework for understanding why human errors occur first occurred when Reason established a theoretical cognitive operating mode classification in his

book *Human Errors* in 1990 [28]. Reason proposed two structural features of human cognition as the *workspace* and the *knowledge bases*. The *workspace* is identified with the schematic control mode, which is responsible for the majority of human daily activities. These activities are controlled unconsciously and automatically by the brain without involving intensive thinking, due to their preconceived templates, so the level of activation required is relatively modest and parallelism is often possible. The *knowledge base*, on the other hand, is identified with the attentional control mode, which is mainly used when conducting activities such as setting goals, selecting the means to achieve them, monitoring progress, detecting and recovering from error, as well as overseeing the schematic control mode. Such activities require a rather great amount of mental attention and hardly operates in parallel.

Reason then connected the operating modes with three types of errors that are closely related to the three cognitive modes proposed by Rasmussen and Jensen [27] including *skill-based*, *rules-based* and *knowledge-based* level.

Skill-based operation is a phase when activities are controlled by the schematic control mode. Errors at the skill-based level, called *slip*, mostly occur due to small failure from the operator, such as distraction, over-attention or under-attention. Rules-based operation involves both schematic control and attention control. When faced with new problems, a problem solver first seeks a solution from known rules and then goes back to skill-based thinking once he/she gets familiar with the routine and can perform the task effortlessly. Rule-based falls, called *lapse*, can occur in two aspects, namely misapplication of good rules, and application of bad rules. Knowledge-based operation usually occurs when heavy reliance on the attention control mode is required

| | | **Operating Modes** | | **Error Type** |
|---|---|---|---|---|
| | | Schematic control mode | Attention control mode | |
| **Cognitive Modes** | Skill-based | ✓ | | Slips |
| | Rules-based | ✓ | ✓ | Lapse |
| | Knowledge-based | | ✓ | Mistake |

Table 4.1: Relationship between Operation Modes, Cognitive Modes, and Error Types

to seek new solutions in unfamiliar environments, for a complicated process that depends on various schemata and rules applications. Therefore, mistakes can occur at multiple stages in multiple forms regarding the knowledge-base level, like "biased memory (psychological biases that enhance or impair information recall), availability heuristic (over-reliance on immediate examples), confirmation bias (searching to confirm one's beliefs) and overconfidence (over-reliance on personal judgment)" [15].

We summarize the relationship between operation modes, cognitive modes, and error types discussed above in Table 4.1 and formal errors defined by Reason, categorized as active failures, in Table 4.2.

| Error | Definition |
|---|---|
| **Slip** | A failure to execute an action due to a routine behavior being misdirected |
| **Lapse** | A failure to execute an action due to lapse in memory and a routine behavior being omitted |
| **Mistake** | A knowledge-based error due to an incorrect thought process or analysis |

Table 4.2: Human Error Types

Healthcare delivery system and ICUs in particular experience human errors on a daily basis despite the fact that people who work in these environments have received exceptional training and are highly qualified and motivated. There are also different ways to classify human errors in ICUs for different modeling aspects, for example, as stated in *To Err is Human* [11]:

**Error of omission:** Failure to perform an appropriate action (e.g., not to wash hands)

**Error of commission:** Performing an inappropriate action (e.g., administering wrong dose of medication)

As one of the most complex system in hospitals, ICU utilizes a numerous member of care teams with different functionality to provide totally specific services, and is a place where all kinds of errors can occur. By understanding the theoretical base of errors that cause harms in critical care, we are motivated to propose the concrete

related factors, and attribute them to different error categories, which we will discuss in the next section.

## 4.2 Risk Drivers

A preliminary review of 15 serious adverse events in the ICU at BIDMC revealed several core event risk drivers, such as:

1. Fundamental conditions of care providers when facing special patients that challenge their expertise outside their routine practice (e.g., when a neurological ICU patient is placed into a unit that typically does not deal with that type of patient, or when an interventional procedure is done in an ICU that typically does not host such procedure)
2. Irregular operations and communication barriers between stakeholders that participate in care delivery (e.g., the allergy information of patient, known to the surgeon team, did not get transferred to the ICU care team due to incapability of communication between existing IT systems, which eventually resulted in a harmful medication dispensing)
3. Heavy workload (e.g., interruptions and distractions in workflow that creates time windows when nurses are called to perform other tasks and not able to pay attention to their patients' conditions).

The identification and measurement of risk drivers largely depends on the work Traina has done [37] in 2014 and 2015, including a modernized and customized version of nursing workload scoring system as a joint work with Ma [17]. Like harms, all drivers are aggregated to unit and shift levels to reflect the operational features BIDMC has and obtain generalization. Patient level data are pulled from BIDMC Information System Electronic Database (Table A.1) for calculation of patient level metrics, which are then aggregated under certain rules to give the unit and shift level measurement. While the set of drivers generally stay unchanged, the proper definitions and calculation details underwent some major modifications, and we will present the fully updated collection of drivers in this section.

Drivers are categorized into four groups: *Acuity*, *Unfamiliarity*, *Utilization*, and *Other*.

- *Acuity* is a measure that describes the illness or clinical severity of patients, which include a pre-defined scoring system to assess patient's organ failure, proportion of patients within their 24-hour of care after admission to an ICU and overall length of stay in the intensive care environment.

- *Unfamiliarity* covers a set of drivers that are recognized from BIDMC clinical team, based on their experience or observations of practices that fall out of the comfort zone of the care provider. For example, this includes nurses working in the environment they are not used to or performing rare procedures to patients.

- *Utilization* is a metric we intend to use to capture the business of nurses in the ICUs, as high workloads, if beyond clinician's manageable capacity, is expected to increase the probability for an error which may result in adverse events. Drivers under this category includes nursing workload, a scoring system based on BIDMC-customized TISS-28 [21, 20] and patient movement across ICU such as admissions and discharges.

- *Other* includes a number of drivers that are inspired from case reviews such as time impact or patients with unknown medical history.

The idea of these three categories is motivated from the concept of three types of human errors in the previous section, namely skill-based, rule-based and knowledge-based. High *acuity* of patients generally requires an extensive and complicated care plan for the patient, where knowledge-based errors are likely to occur. *Unfamiliarity* challenges the nurse in performing standard practices, which may result in rule-based errors. *Utilization* is closely related to skill-based errors from nurses not paying enough attention to patients when being overloaded with other works.

Table 4.3 summarizes such relationships between types of human errors and categories of risk drivers, as well as the specific risk drivers falling under each category, which will be discussed in detail in the next section.

Table 4.3: Types of Human Errors and Risk Drivers

| **Error Type** | **Category** | **Driver** |
|---|---|---|
| Knowledge-based | Acuity | SOFA |
| | | First 24 hour |
| | | Length of stay in ICU |
| Rule-based | Unfamiliarity | Float nurse |
| | | New nurse |
| | | Boarding patient |
| Rule-based | Utilization | Nursing workload |
| | | Nurse to patient ratio |
| | | Admissions |
| | | Discharges |
| | | Close-in-time movement indicator |
| | Other | EU Critical |
| | | Night & Day |
| | | Weekend |
| | | Unit |

To simplify the representation of driver definitions in our model, we first introduce the following notations:

$\texttt{AT}_k$: admission time for patient $k$ into an ICU from a non-ICU location

$\texttt{DT}_k$: discharge time for patient $k$ from ICU

SS: start time of shift, 7 am or 7 pm

ES: end time of shift, 7 am or 7 pm

$T_a$: total number of admissions during a shift

$T_d$: total number of discharges during a shift

$T$: total number of movements during a shift (i.e., $T_a + T_d = T$)

$t_i$: time stamp when the $i$-th movement (i.e. admission and discharge) of patients happens during a shift, $i = 1, 2, \ldots, T$, $t_0 = \texttt{SS}$

$N_i$: number of patients in the unit during a shift at $t_i$ before the movement, i.e., excluding the one to be admitted and including the one to be discharged

$P_k^i(\cdot)$: patient level measure/indicator for some drivers during a shift at $t_i$

$S$: total number of nurses during a shift

The reason we record time stamps for each patient movement is that when a patient is admitted to or discharged from ICU, it might change the risk driver states in the ICU during a shift. For example, we want to calculate the faction of first 24-hour patients in the unit. It is 2 out of 8 (i.e., 0.25) at the beginning of the shift. After 4 hours, one new patient is admitted, and the fraction becomes 3 out of 9 (i.e., 0.333) and stays the same until the end of the shift (for another 8 hours). Instead of using one of these two numbers, we choose to calculate a time-weighted average of the two for a more reasonable measure, namely 0.25*4/12 + 0.333*8/12 = 0.306, since each shift is 12-hour long.

Such time weighted average is one solution to incorporate all values of a driver and weigh them based on how long each scenario lasts within a shift, and gives a robust way to capture the dynamic changes that have occurred throughout the shift.

### 4.2.1 Acuity

1. **Sequential Organ Failure Assessment (SOFA) Score**

The SOFA score is a scoring system to determine the extent of a person's organ function or rate of failure. It is developed by a group working on sepsis-related problems in 1996 [39], and is currently widely used in assessing ICU patients. The score is based on patients' organ performance in six different areas including the respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems. With the score range between 0 to 24, a value only as high as 11 can give predictive mortality close to or above 95% while a score below 9 can decrease the rate to 33% [39].

Patient Level:

To calculate the SOFA score for one patient during a particular shift, we take the

following 4 steps:

1) Extract necessary data from relevant datasets for the patient with time stamps between the start of the previous shift and the end of the current shift of interest (i.e., need to span the 24-hour window)
   - If no records exist for the time window, choose the one from before the starting point cutoff that is the closes to the current shift
   - If no records exist at all for the patient for all time, assume the patient's record for that field is within the normal range
2) Calculate the metric for each component with data obtained from previous step
3) Compare the calculated metric with the values/conditions in Table C.2 and assign the corresponding points to that component for the patient
   - If multiple point values are obtained during this step, choose the worst one
4) Add up points for all six components to obtain the SOFA score for the patient

Note that $PaO_2$ and $FiO_2$, required to calculate the respiratory component, are usually not documented for all patients, so an alternative implementation based on a slightly different scheme called Modified SOFA is used in our project [10]. The calculation details are listed in Appendix C.2.1.

Unit Level:

To aggregate the drivers into unit level, we first divide a shift (12 hours in length) into segments according to movements of patients. As previously explained, this is because whenever a patient is admitted into or discharged from the unit, it changes the unit level condition. We then average over all time segments and over all patients for the patient level measure to obtain the unit level measure. There are several ways to include the aggregated measure of patient level SOFA to unit level:

$$\text{Average:} \quad \mu(\texttt{SOFA}) = \frac{1}{12}\sum_{i=1}^{T}(t_i - t_{i-1})\left[\frac{1}{N_i}\sum_{k=1}^{N_i} P_k^i(\texttt{SOFA})\right]$$

$$\text{Standard Deviation:} \quad \delta(\texttt{SOFA}) = \sqrt{\frac{1}{12}\sum_{i=1}^{T}(t_i - t_{i-1})\left[\frac{1}{N_i}\sum_{k=1}^{N_i} P_k^i(\texttt{SOFA}) - \mu(\texttt{SOFA})\right] \cdot \frac{T}{T-1}}$$

$$\text{Highest:} \quad H(\texttt{SOFA}) = \max_k \{P_k^i(\texttt{SOFA})\}$$

$$\text{Quantile:} \quad Q(\texttt{SOFA}) = \frac{1}{12}\sum_{i=1}^{T}(t_i - t_{i-1})\mathbb{1}_{\{P_k^i(\texttt{SOFA}) \geq 9\}}$$

The "Average" of SOFA score first calculates the average SOFA score of all *current* patients ($N_i$ of them) during each time segment, i.e., between $t_i$ and $t_{i-1}$ for $i = 1, 2, \ldots$. It then averages over all time segments with the length of the segment as its weight.

The other three measures follow a similar pattern, but measuring the acuity level using SOFA from a different perspective. While "Standard Deviation" and "Highest" is self-explanatory, "Quantile" is essentially trying to capture, on average, how many patients are really sick (SOFA≥9), which is an alternative to standard deviation to model the dispersion of SOFA score in the unit.

2. **First 24 Hour**: weighted average of fractions of patients who are within the first 24 hour from admission to ICU during a shift.

Patient Level:

$$P_k^i(\texttt{F24H}) = \mathbb{1}_{\{t_i - \texttt{AT}_k \leq 24\}}$$

Unit Level:

$$\texttt{F24H} = \frac{1}{12}\sum_{i=1}^{T}(t_i - t_{i-1})\left[\frac{1}{N_i}\sum_{k=1}^{N_i} P_k^i(\texttt{F24H})\right]$$

3. **Length of Stay in ICU**: weighted average duration of time the patients who have been in the ward since they were admitted to an ICU.

Patient Level:

$$P_k^i(\texttt{LS}) = t_i - \texttt{AT}_k$$

Unit Level:

$$\texttt{LS} = \frac{1}{12}\sum_{i=1}^{T}(t_i - t_{i-1})\left[\frac{1}{N_i}\sum_{k=1}^{N_i} P_k^i(\texttt{LS})\right]$$

### 4.2.2 Unfamiliarity

Some of these drivers have consistent values because they are related to conditions of nurses, which does not change throughout the shift.

1. **Float Nurse**: fraction of nurses in the unit who are considered as float.[1] This is flagged in the staffing data.

$$\texttt{FN} = \frac{1}{S}\sum_{j=1}^{S} \mathbb{1}_{\{\text{Nurse } j \text{ is "Floating"}\}}$$

2. **New Nurse**: fraction of nurses in the room who are considered as unexperienced because they have been hired to work in the ICU for less than one year. This information is calculated by the difference between the nurse's hire date and the date of the shift, available in the staffing data.

$$\texttt{NN} = \frac{1}{S}\sum_{j=1}^{S} \mathbb{1}_{\{\text{Nurse } j \text{ is "New"}\}}$$

3. **Boarding Patient**: average fraction of boarding patients who are assigned to ICU that does not usually provide the type of service patients are associated with.

[1] Float nurses are nurses who work in ICU wards other than the one they usually work in, for instance, a medical ICU nurse working in a surgical unit. These nurses are less familiar with the new unit, and are likely to make errors, which is why we consider it as a risk driver. BIDMC also has "float nurse pool", of which the nurses are trained to work in various ICU environment. This type of float nurses are not included in our definition of float nurse.

Boarding into an ICU requires the satisfaction of two conditions, namely it is possible, in real practice, to assign the patient with some service to certain unit AND the assigned unit is not a home unit for the service.

Patient Level:

A patient is a boarding patient if his/her service type and assigned unit combination gets value 1 in Table C.1 and not if 0.

Unit Level:

$$\texttt{Boarder} = \frac{1}{12}\sum_{i=1}^{T}(t_i - t_{i-1})\left[\frac{1}{N_i}\sum_{k=1}^{N_i} P_k^i(\texttt{Boarder})\right]$$

### 4.2.3 Utilization

Utilization related drivers generally measures nurses' workload, i.e., the utilization of nursing hours.

1. **Nursing Workload (TISS point)** :average nursing workload based on the duties they need to perform in terms of required patient activities, measuring by the Therapeutic Intervention Scoring System (TISS). It gives different number of points to different activities based on its intensiveness. For example, a lab draw is 1 point for the nurse, while ventilating a patient is 5 points. A nurse can do up to 46 TISS points per 8-hour shift, or 69 in terms of BIDMC 12-hour shift. In real practice, a shift with score of 0-18 means it has light workload, 18-24 moderate, and 24+ being high. Details are discussed in Traina's thesis [37].

Patient Level:

To obtain patient level workload, first use relevant data and Table C.4 to acquire all the activity based points one should receive during a shift, them sum them up.

Unit Level:

Sum all patients workload and normalize by the total number of patients.

2. **Nurse to Patient Ratio**: ratio between nurse working hours and patient hours, a measure for average utilization of nurses' time during a shift. This is a unit level driver.

Unit Level:

Each nurse has 12 hours available and for each patient, its patient hours for the shift is determined by the corresponding admission and discharge time stamps.

$$\texttt{PH}_k = \min\{\texttt{DT}_k, \texttt{ES}\} - \max\{\texttt{AT}_k, \texttt{SS}\}$$

Then sum up both to calculate the ratio:

$$\texttt{NPR} = \frac{\sum_{k=1}^{N} \texttt{PH}_k}{12 \cdot S}$$

where $N$ is the total number of patients who have been physically in the unit at least once, no matter for how much time, during the shift and $S$ is the total number of nurses. This number is theoretically supposed to be between 0.5 and 1, with 0.5 meaning all nurses are taking care of two patient throughout the shift, and 1 meaning all 1 to 1 assignment throughout the shift. However, with resource nurses and training nurses, this number can go above 1.

3. **Admission**: number of admissions to the unit, modified to be normalized by nurse to patient ratio in order to capture the distraction for the nurse from their current work to care for patients to accept the new admission. For each admission $i$, we have:

$$\texttt{AD}_i = \frac{1}{\frac{N_i}{S}} = \frac{S}{N_i}$$

where $S$ is the number of nurses working in the unit for the shift.

Unit Level:

$$\texttt{AD} = \sum_{i=1}^{T_a} \texttt{AD}_i$$

4. **Discharge**: number of discharges to the unit, modified to be normalized by nurse to patient ratio in order to capture the distraction for the nurse from their current work to care for patients to handoff the patient and clean the bed. For each discharge $i$, we have

$$\texttt{DS}_i = \frac{1}{\frac{N_i - 1}{S}} = \frac{S}{N_i - 1}$$

where $S$ is the number of nurses working in the unit for the shift.

Unit Level:

$$\texttt{DS} = \sum_{i=1}^{T_d} \texttt{DS}_i$$

5. **Close-in-time Movement Indicator**: unit level binary or continuous indicator.

**Binary:** 1 if there are multiple ($\geq$ 2) admissions or discharges happening within 2-hour window, 0 otherwise

$$\mathbb{1}_{\{t_i - t_{i-1} < 2\}} \qquad \forall i = 1, 2, \ldots, T$$

**Continuous:** count the number of *cases* where multiple ($\geq$ 2) admissions or discharges happen within a 2-hour window, 0 if none

$$\sum_{i=1}^{T} \mathbb{1}_{\{t_i - t_{i-1} < 2\}} \qquad i = 1, 2, \ldots, T$$

### 4.2.4 Others

1. **EU Critical**: average fraction of patients who are flagged as "EU Critical". These patients get admitted to BIDMC either with an unknown identity of no records of medical history.

Patient Level:

Patient is counted toward EU Critical if they identifier is listed in the EU Critical database and is within the first 48 hours of admission to hospital. Status changes back to non-EU Critical automatically after 48 hours of any kind of hospitalization.

Unit Level:

$$\texttt{EUC} = \frac{1}{12}\sum_{i=1}^{T}(t_i - t_{i-1})\left[\frac{1}{N_i}\sum_{k=1}^{N_i} P_k^i(\texttt{EUC})\right]$$

2. **Night and Day**: shift level binary indicator. 1 for night shifts (7pm to 7am) and 0 for day shifts (7am to 7pm).

3. **Weekend**: shift level binary indicator. 1 for weekend shifts occurring between 7pm Friday and 7am on Monday and 0 for the rest.

4. **Unit**: categorical variable that identifies the unit information for other drivers. 7 possible values for this driver include FICU, SICU, TSICU, MICU6, MICU7, CVICU, and CCU.

Driver processing yields, for each shift, a set of values describing the conditions from the above four aspects of an ICU for that shift.

# Chapter 5

# Statistical Methodology

By preprocessing the harms and drivers, we obtained, for each shift and in each unit, a set of 15 drivers that describe the unit conditions and the corresponding *average harm rate per patient* (for all instantaneous harms) in that shift (12 hour in length) and unit. Table 5.1 shows the sample data structure after the preprocessing, with each row representing a shift, and each column representing a variable (last column is the response variable and the rest are independent variables). For example, the first row is one observation of a shift in MICU6, it has a SOFA score of 8, fraction of float nurses at 25% and a TISS score of 18, etc.

| **SOFA** | **Float Nurse** | **TISS** | **Unit** | ... | **Harm Rate** |
|---|---|---|---|---|---|
| 8 | 0.25 | 18 | MICU6 | ... | 0.036 |
| 6 | 0 | 15 | SICU | | 0.008 |
| ... | ... | ... | ... | ... | ... |

Table 5.1: Sample Data Snapshot

Its associated average harm rate is 0.036 harms per patient in MICU6 of that shift. This is calculated using equation ($\star$) in Chapter 3.2. Specifically, we first obtain the harm rate per patient for each harm event by calculating the ratio between the absolute occurrence of each harm and the number of relevant patients in the unit during the shift for that harm. We then sum up all these ratios and divided by 25 since there are 25 different types of instantaneous harms.

In total, we have around 10,000 such observations (around 1,400 shifts for each unit). With the existence of response variable, we use supervised learning algorithms to model the relationship between drivers (independent variables/features) and harm rate (response), and identify how one or more drivers impact the harm rate in BIDMC ICUs. The output of the algorithms describes a set of *risky (safe) states*, which are represented by a cluster of shifts that share similar feature value ranges (states) and have a higher (lower) rate of harm compared to the overall harm rate across all shifts and units. The hope is that understanding what kind of feature values are correlated with high rate of harms, we can 1) make intervention plans to prevent such states from happening, or 2) alert on them to call for extra attention in the clinical operation, thus reducing likelihood of harm events and providing a safer care environment to patients. However, we face several challenges in such model construction:

- The model output must be descriptive.

  There are many statistical models that can capture the relationship between a set of independent variables and the response, simple ones like linear regression or sophisticated ones like neural networks. These methods are built with the purpose of predicting new outcomes given new data, and are not necessarily able to provide descriptive explanation of how feature range values are correlated with the response variable, which is the key information we are seeking in order to launch mitigation plans. Therefore, we need develop an algorithm more than just a simple application of some black-box type of statistical model.

- The results need to be statistically robust.

  With the idea of aggregation embedded, it is hard to find already validated clinical evidence to support the innovative Risky States approach other than clinical intuitions from experts at BIDMC. Therefore, we need to be very careful about the credibility of the statistical tools we use and looking for consistencies across statistical methodologies.

- We work with an unbalanced dataset.

  Since specific ICU harms are typically very rare events, the data we obtained for modeling statistical relationship between harm events and drivers would

be unbalanced with many shifts seeing no harm events (i.e., the response is 0). We alleviate this problem by first aggregating over different types of harm events, but we still need to take this into consideration as unbalanced dataset will cause problems in various statistical models.

The new approach designed in the thesis attempts to address these challenges by employing a combination of regression trees, random forest and $k$-neareast neighbor (KNN). Next we provide a brief outline of our approach:

1. Run multiple regression trees to reduce dimensionality of the driver space
   - Eliminate drivers that are not chosen often by the trees
2. Train a random forest model using the remaining drivers
   - Each regression tree built on a 75% randomly chosen sample of the data
3. Create samples that are uniformly drawn from the driver space
   - Predict the harm rate for each sample using random forest model
4. Build a regression tree on the combination of real data and simulated data
5. Identify risky/safe states
6. Verify the stability of risky/safe states using KNN
7. Test the significance of the risky/safe states using Mann-Whitney-Wilcoxon hypothesis test

## 5.1 Regression Tree and Random Forest

The idea of linear regression is to find a linear relationship between a real-valued dependent variable $Y$ using a set of independent variable $\mathbf{X}$. A regression tree is to utilize this simple and straightforward idea of linear regression, and at the same time achieve manageable interaction among drivers for non-linearity through recursive partition and offer descriptive statement about independent variables [16]. The algorithm works in the following way:

1. For each independent variable, find all possible values as the split points for all independent variables

2. Choose one independent variable and one split point, fit two simple linear regressions to the split data, respectively
3. Repeat for all independent variables at all split points
4. Choose what independent variable to split at what value:
    (a) Calculate the the sum of squared errors (SSE) between the predicted value and the actual values, for each variable and each split
    (b) Compare the error across all the independent variables and all the splits
    (c) Choose the one yielding the lowest SSE and split the data into two partitions using the corresponding variable and split point
5. Repeat step 1 to 4 until stop criterion

Ultimately, the predictor space is partitioned to multiple subspaces, and each space is associated with a leaf node in the tree representation, with a fitted value of the response variable for all the observations inside the node determined by tree partition rules. These rules are essentially what we will use to describe the states. Regression tree has many advantages including, for example, fast computation with large dataset, insensitive to scaling, indirect interaction discovery, resistance to irrelevant variables, capable of taking categorical variables, handling missing values and tuning for parameter. Most importantly, it offers us interpretable model representations.

In our case, a regression tree will partition the driver space recursively into smaller subspaces, and for each partition, it will give a fitted patient harm rate value, obtained from the average of all observations falling in that partition. Figure 5-1 gives a simple example using two drivers *SOFA* and *Admission*.

In this naive example, the tree is basically telling us that if the unit average SOFA score is below 5, the fitted rate of harm is 0.5 per patient, which is relatively safe (lower than average). On contrary, high SOFA and high admission together yields a much high rate at 2 harms per patient, making the unit and patients in relative danger.

In general, we use regression trees for two purposes. First is feature selection. To do this, we build 10,000 regression trees on different sets randomly drawn, with

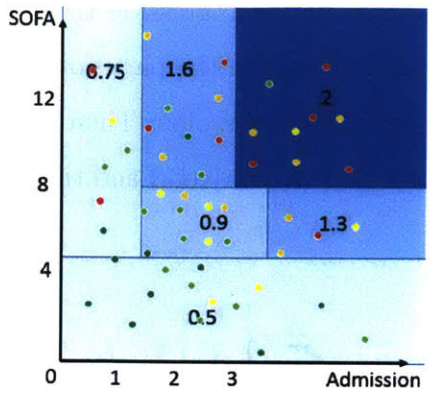



Figure 5-1: How a Regression Tree Works

replacement, from the dataset as shown in Table 5.1, with each set 75% of the total number of observations (around 7,500 data points), and check how many drivers are consistently chosen by the trees to split on. The trees are pruned so that each node has at least 150 observations (shifts) to avoid over-fitting within the tree. Since the tree will ignore irrelevant variables, we decide to only include the ones chosen by at least 50% of the time by the trees, namely, *SOFA, Admission, Discharge, TISS, Night, NP Ratio,* and *unit*[1], as shown in Figure 5-2.

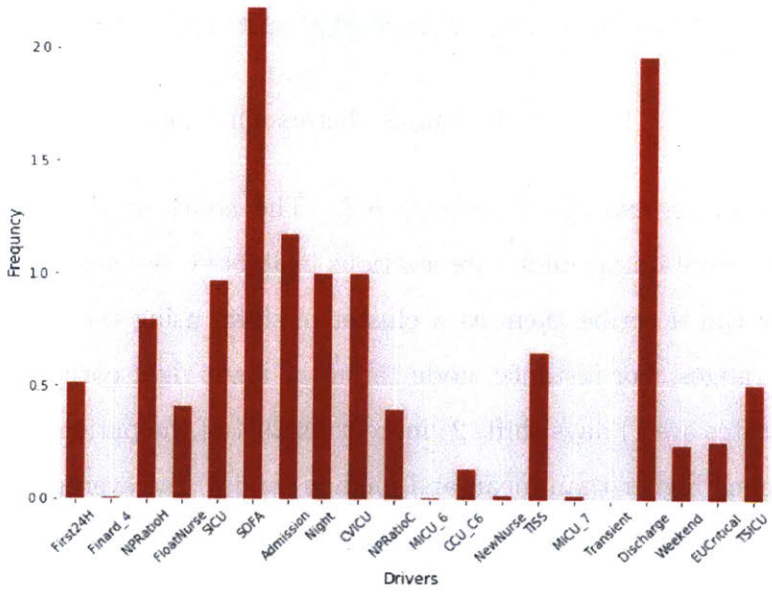



Figure 5-2: Feature Frequency

[1]Note that *unit* is a categorical variable, and we created dummy variables for easy visualization.

The second purpose of constructing regression trees is to learn how ICU states affect patient harm rate by understanding the partition rules, which describe the way to split a set of drivers at particular thresholds. Then multiple subspaces are formed in terms of the selected dimensions (i.e. drivers) and their respective ranges of values, what we call "states".

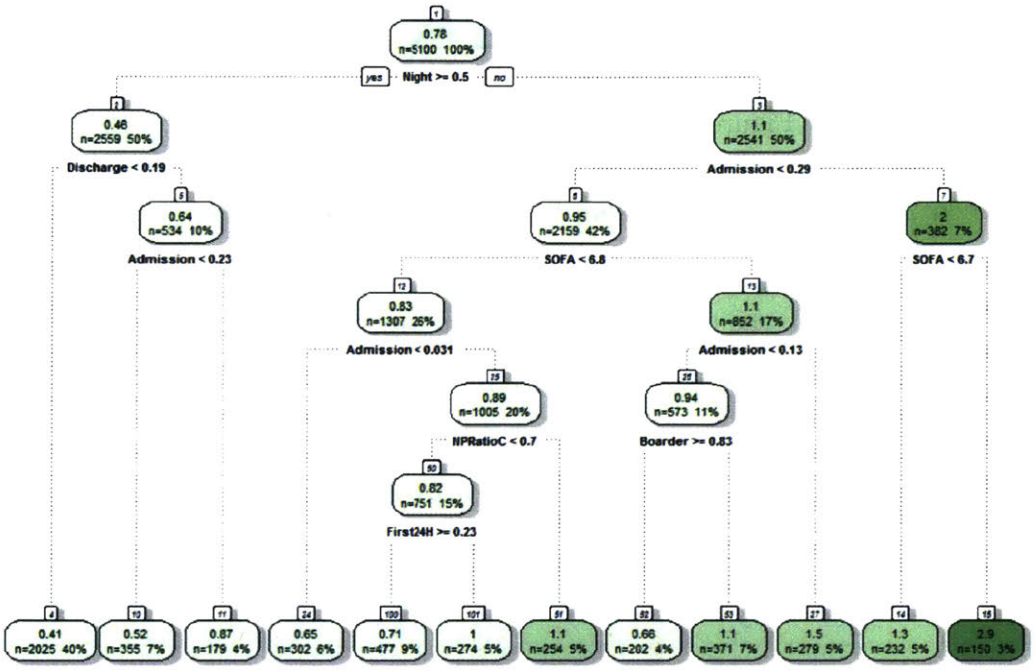



Figure 5-3: Sample Regression Tree

Let us look at an example in Figure 5-3. The nodes at the bottom refer to leaf nodes, each containing many observations that obey the same partition rules. This means we can describe them as a cluster of shifts using the same drivers and corresponding ranges. For instance, node 15 on the lower right corner consists of 150 shifts, whose states are 1) days shift, 2) more than 29% of the patients in the unit are new admits (much higher than mean and median) and 3) the average SOFA score of the unit is higher than 6.7 (higher than mean and median), as illustrated in Figure 5-4. Since it is associated with a harm rate of 2.89 while the population average is 0.78, this is defined as a risky state.

Other risky states and safe states are also identified by the tree, with some presented in Figure 5-5. In general, we see less workload (in terms of number of admissions and discharge) and less acuity (in terms of SOFA) is associated with less harm, which is clinically valid.

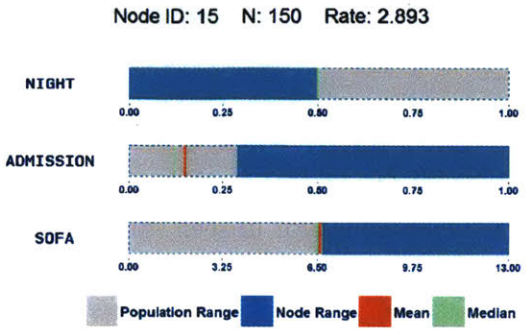



Figure 5-4: Node 15 Description

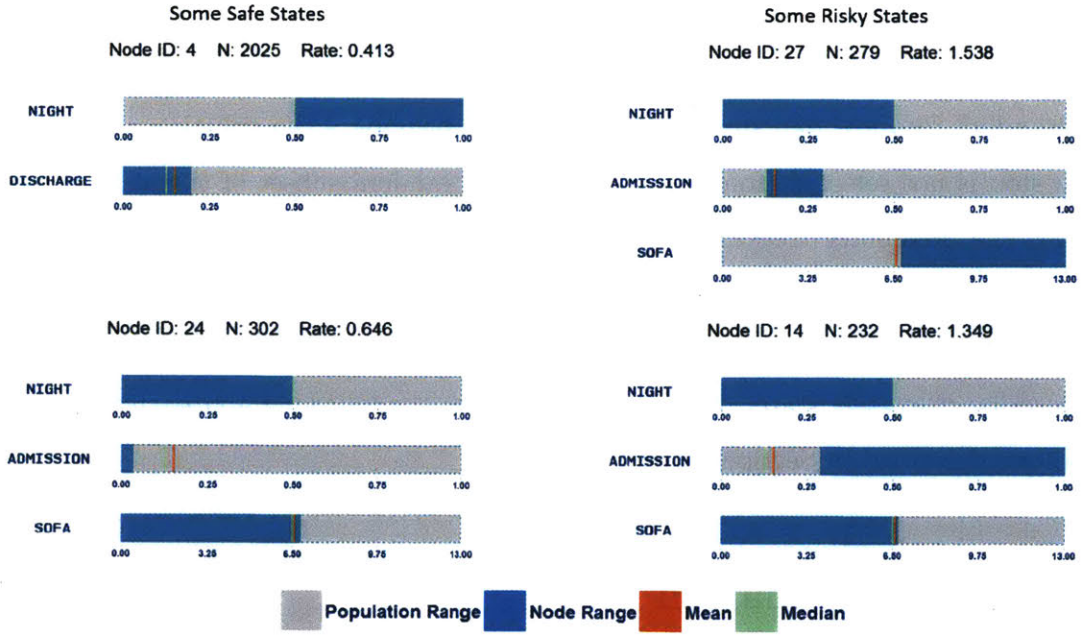



Figure 5-5: Safe States and Risky States

As we can see, regression trees are easy to understand and provide us with clear descriptions of states associated with different level of response values. However, it also comes with drawbacks. Regression tress are known to be very sensitive to data they are trained on. This means a slight change in some variable could yield totally

differently shaped trees. Therefore, a regression tree trained on one random split of the data is likely to be overfitting and will not give very reliable descriptions of risky states. To solve for this problem, many ensemble algorithms are used such as tree bagging and random forest.

Tree bagging is a combination of regression trees. More specifically:

for $b = 1, \ldots, B$:

1. Sample, with replacement, $B$ examples from $\mathbf{X}$, $Y$, call each of these $\mathbf{X}_b, Y_b$
2. Train a regression tree $f_b$ on $\mathbf{X}_b, Y_b$
3. After training, predict the response for unseen samples $x$ by averaging the predictions from all individual regression trees on $x$:

$$\hat{f}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} f_b(\mathbf{x})$$

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias [36].

Random forest is one step further from tree bagging - the splitting variable at each step is not selected from all features, but a random subset of them, to further reduce the possibility trees being correlated [2]. Therefore, random forest resolves the problem of overfitting. However, since each tree in random forest splits using different features at different threshold, it then becomes unclear what we should use to describe risky/safe state. Therefore, we propose the following approach to fix this problem, described in Section 5.2 below.

## 5.2 Data Simulation

We start with simulating data samples according to the current distribution of drivers in the feature space. More specifically, for each driver that is continuous, we do the following step:

1. Generate the histogram based on the historical data
   - Set the number of bins to 25

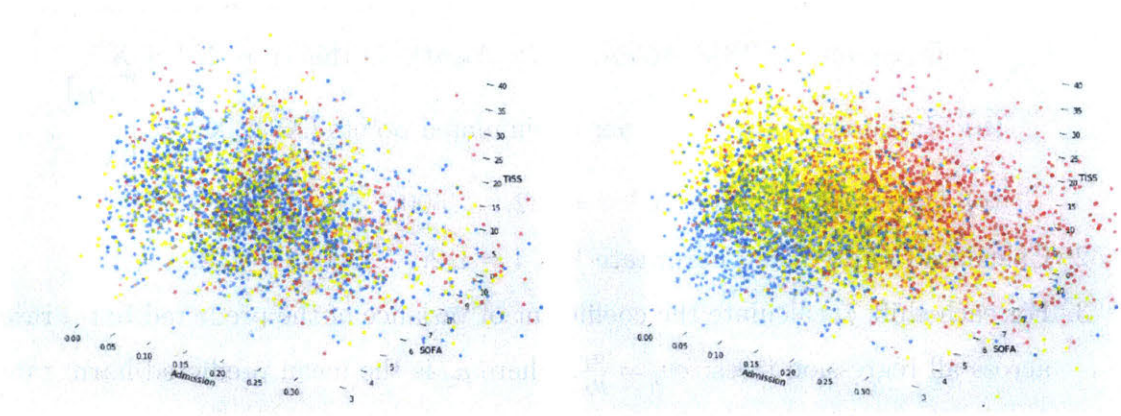
2. Calculate the empirical distribution for the mid-point $m_i$ of each bin
   - $P(m_i) = \frac{\text{height of bar}_i}{\text{total number of observations}}, i = 1, 2, \ldots, 50$
3. Sample a point from the empirical distribution from step 2
4. Add some randomness to the sample point
   - Calculate the driver range: $range = \max - \min$
   - Sample a number uniformly from $[-2\% \times range, 2\% \times range)$
   - Add the random number to the midpoint sample drawn at step 3

For drivers that are discrete, we directly sample from the empirical distribution based on the historical data. For instance, the empirical distribution for the driver *Night* is $P(\text{day shift}) = P(\text{night shift}) = 0.5$.

To generate a complete observation, we repeat the above steps for all drivers and calculate the corresponding response using the random forest model trained based on all real data from last section.

At this point, we obtained a new set of data points, with driver values and harm rate in the same format as the real dataset. We denote the original real data $(\mathbf{X}^R, Y^R)$, the simulated data $(\mathbf{X}^S, Y^S)$ and the combination of the two $(\mathbf{X}^A, Y^A)$.

Figure 5-6 gives a comparison of distribution of shifts in a 3-dimensional driver subspace (*Admission*, *SOFA*, and *TISS*) between the real data $\mathbf{X}^R$ (Figure 5-6a) and the combination of both real and simulated $\mathbf{X}^A$ (Figure 5-6b). Color of different points bear the following meaning: blue - $< 30\%$ less than population average harm rate;




(a) Real Data ($\mathbf{X}^R$) (b) Real data + Simulated Data ($\mathbf{X}^A$)

Figure 5-6: Scatter Plot of Shifts in a 3-dimentional Driver Subspace

red: $> 30\%$ more than population average harm rate; yellow - within 30% interval of population average harm rate.

Figure 5-6b is able to depict a more obvious edge between lower-rate-of-harm shifts and higher-rate-of-harm shift immediately. In addition, by adding more data points in the feature space, we lower the risk of sensitivity in regression tree construction caused by different sampling of the data. This is because, with data points being very dense in feature space, even multiple times of random sampling will result in very similar data point distributions in different samples, which in turn should give very similar regression trees. We use the following three metrics to measure the consistency of regression trees trained using different samples drawn from the same dataset:

- Average coefficient of variance of predicted values across different regression trees for all original data observations ($\tilde{Y}^R$).
- Standard deviation of the number of times each driver is selected by the regression tree across all trees.
- Distribution of split thresholds for each driver across all regression trees.

We will show an increase in consistency in the regression trees after data simulation with these three metrics, using the following steps.

1. Train 500 regression trees using a random sample (75%) drawn from a set of shift observations:
   (a) The initial set of shift observations only consists of real data (a total of $n = 10200$ observations and 6 independent variables: *Admission*, *Discharge*, *NPRtio*, *SOFA*, *TISS*, *Night*). In this case, $\mathbf{X}^A = \mathbf{X}^R$.
   (b) Gradually add the number of simulated points ($\mathbf{X}^S$). $\mathbf{X}^A = \begin{bmatrix} \mathbf{X}^R \\ \mathbf{X}^S \end{bmatrix}$.
   (c) Denote each sample $\mathbf{X}_i^A$, $i = 1, 2, \ldots, 500$.
2. Calculate the predicted harm rate $\tilde{Y}_{ij}^A$, $i = 1, 2, \ldots, 500$, $j = 1, 2, \ldots, n$.
3. For each shift $j$, calculate the coefficient of variance of the predicted harm rate across all regression trees: $cv_j = \frac{\sigma_j}{\mu_j}$, where $\mu_j$ is the mean predicted harm rate for $j$ and $\sigma_j$ is the standard deviation of predicted harm rate for $j$.
4. Calculate the average coefficient of variance: $\bar{cv} = \frac{1}{n}\sum_{j=1}^{n} cv_j$.

The number of simulated points ranges from $10^2$ to $10^6$. Figure 5-7 shows how the average coefficient of variance decreases as the number of simulated data points increases.

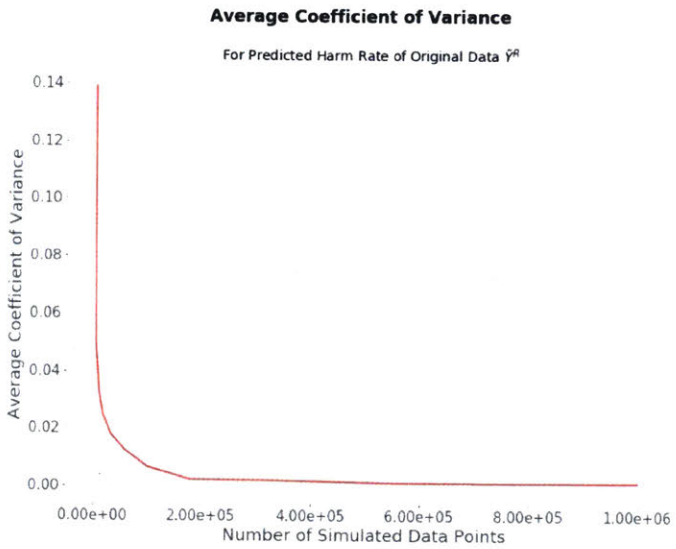



Figure 5-7: Average Coefficient of Variance

When the number of simulated data points is approximately $2 \times 10^5$, the average coefficient of variance converges to a small stable value. The following comparison of the other two consistency metrics will be conducted using $n = 2 \times 10^5$.

Figure 5-8 compared the mean and the standard deviation of the number of times each driver is selected by the regression tree, between the original data set $\mathbf{X}^R$ and the combined dataset $\mathbf{X}^A$. The black line in the middle of each bar represents one standard deviation away from the mean. We can see from the plot that the number of times selected by different trees for each driver (except *Night*) varies a lot when we only used the original dataset to train the random forest. In contrast, there is almost no variation when we used the combined dataset with both original and simulated data (the standard deviation is 0 so there is no visible black line for corresponding bars). More specifically, each tree from the latter case split once for driver *Admission*, *SOFA*, and *Night*, twice for driver *Discharge*.

Finally, we take a look at where each driver is being split. Since regression trees

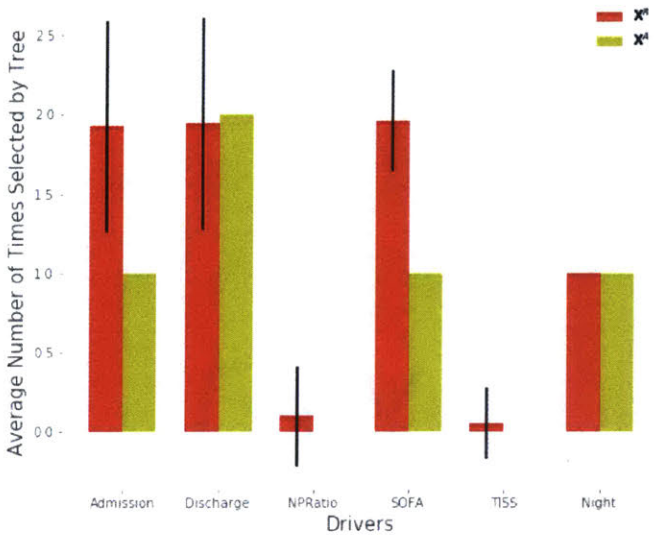



Figure 5-8: Average Driver Frequency Selected By Trees

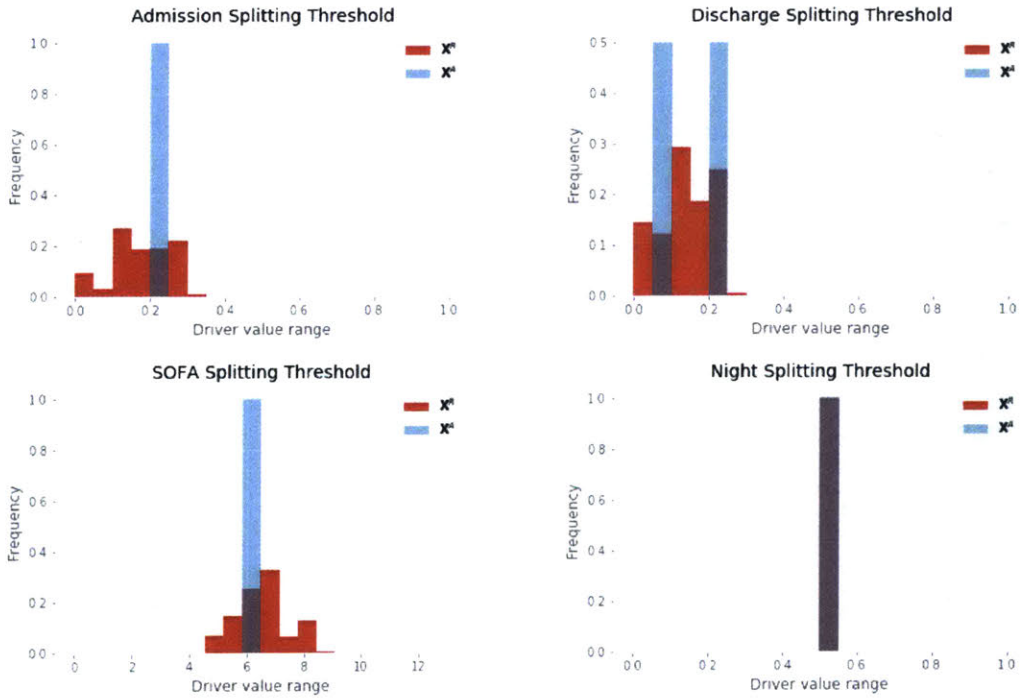



Figure 5-9: Distribution of Splitting Values for Different Drivers

trained using $\mathbf{X}^A$ do not split with respect to drivers *NPRatio* and *TISS*, we present here the four other drivers. From Figure 5-9, we can see that, except for *Night*, the

splitting thresholds for the other three drivers concentrate at a very small and stable range for $\mathbf{X}^A$ comparing to $\mathbf{X}^R$.

All three metrics we investigated above have shown that training regression trees using random samples drawn from the combined dataset yields a much higher consistency. Therefore, after we generated the simulated data, we can obtain relatively reliable clusters even by fitting one regression tree using the simulated data. Each cluster represented by a leaf node in the tree with an average harm rate higher (lower) than the population average describes a risky (safe) state through its defining drivers and corresponding thresholds.

To further check the stability and credibility of risky states and safe states obtained from the regression tree in this way, we will do two more checks using the $k$-nearest neighbors algorithm and the Mann-Whitney-Wilcoxn test, discussed in Chapter 5.3 and 5.4, respectively.

## 5.3 *K*-Nearest Neighbors

The $k$-nearest neighbors algorithm (KNN) is a non-parametric method used for classification and regression. In this thesis, we use it for regression, namely the property value of a data point is the average of that of its $k$ nearest neighbor points. For example, let's assume $k = 3$. In Figure 5-10, the predicted value for point $x$ is the

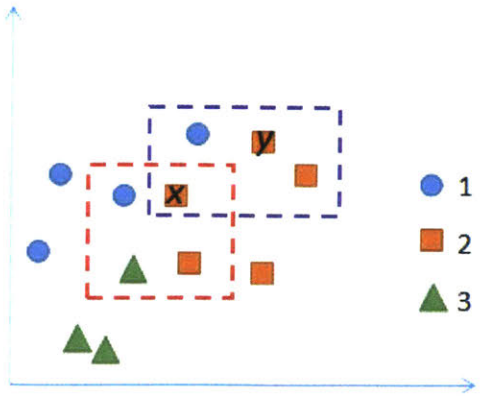



Figure 5-10: $K$-Nearest Neighbor for Regression

average of the value of the three points in the red dashed square $((1+2+3)/3 = 2)$, while that for point $y$ is the average of the three points in the purple dashed square $((1+2+2)/3 = 1.67)$.

In our setting, the value of interest is the harm rate per patient in the shift, so each shift's rate is calculated by averaging its nearest $k$ shift observations.

The distance metric used in KNN is the euclidean norm for relevant independent variables, the ones discussed in the previous section. Generally, a small $k$ number will make the model very sensitive to small changes and the edges of separation very rough. On the other hand, a large $k$ can decrease the sensitivity, but has less power of distinguish different data points. In our analysis, we use $k = 30$, which is generally believed to provide a good balance between the two.

After we obtain clusters of shifts from the regression tree trained in the last step of Chapter 5.1, we apply KNN to the shifts within each cluster. This gives an average harm rate for all shifts in the cluster through the KNN approach, which we denote as $\mu_s^K$ for a cluster $s$. This is independent from the average harm rate obtained from the tree algorithm for the same cluster, which we denote as $\mu_s^T$. To check whether a cluster of shifts is *stable* in the sense that the KNN algorithm agrees with the regression tree when the cluster yields higher/lower than population average harm rate, we define

$$h(\mu, \mu_s^K, \mu_s^T) = \begin{cases} 1 & \text{if } \operatorname{sgn}(\mu_s^K - \mu) = \operatorname{sgn}(\mu_s^T - \mu) \\ 0 & \text{otherwise} \end{cases}$$

where $\mu$ is the population average harm rate. When $h$ is 1, KNN agrees with the tree, so we conclude that the corresponding cluster is a stable cluster, and the combination of feature value ranges defined by the tree are valid descriptions of risky/safe states. Otherwise, the cluster is not stable and to be discarded.

## 5.4 Mann-Whitney-Wilcoxn Test

Finally, we will test the statistical significance of clusters obtained from the regression tree. Recall that risky (safe) states are represented by clusters of shifts whose average

harm rate within cluster is higher (lower) than the population average. However, without a rigurous test, it is hard to conclude whether the difference between the two average is statistically significant. If indeed the distribution of harm rate in cluster and that of the general population are not different and the average differs slightly only due to sampling, then the risky/safe states provided by such clusters are not reliable and should not be taken into consideration.

The Mann-Whitney-Wilcoxon (MWW) test is a non-parametric statistical test determine the significance of the null hypothesis ($H_0$) that two samples come from the same population versus the alternative hypothesis ($H_a$) that the two samples are from different populations. Under the assumption of continuous responses, this essentially tests the alternative hypothesis whether one distribution is greater than the other.

Specifically in this thesis, if the harm rate distribution over shifts within a certian cluster is statistically different from that of the population by conducting the MWW test and comparing the results with $p$-value of 0.05, we claim the cluster is a well-founded cluster whose feature value ranges give a valid risky/safe state. Otherwise, the cluster cannot be considered as different from the population, thus not able to identify risky/safe states.

# Chapter 6

# Results

## 6.1 Overview

The results presented in this chapter is from the application of the statistical framework described in Chapter 5 to identify risky states for all *instantaneous* harms. The harms included in the analysis are:

Table 6.1: All Instantaneous Harms

| **Harm** | **Relevance Patient Cohort** |
|---|---|
| Administer Nalaxone | Received Narcotics |
| Administer Vitamin K | Received Warfarin within previous 48H |
| Airway management | Being venilated |
| Bleeding (Abrupt drop in Hemoglobin) | All |
| Bleeding (INR $>$ 6) | All |
| Bleeding (PTT $>$ 100) | All |
| Blood product | All |
| Consent | All |
| Code Blue | All |
| Code Purple | All |
| Diagnosis, treatment, testing | All |

Continued on next page

Table 6.1: All Instantaneous Harms

| **Harm** | **Relevance Patient Cohort** |
|---|---|
| Environment, equipment, supply | All |
| Fall | All |
| Glucose | On Insulin drip |
| Handoff, communication | All |
| High tidal volume (ARDS) | Being ventilated |
| Iatrogenic Pneumothorax | With qualified central lines |
| Identification | All |
| Lab specimen | All |
| Medication error | All |
| Readmission | Discharged within previous 48H |
| Reintubation | Extubated within previous 12H |
| Restraints | All |
| Safety | All |
| Unplanned Extubation | Being ventilated |

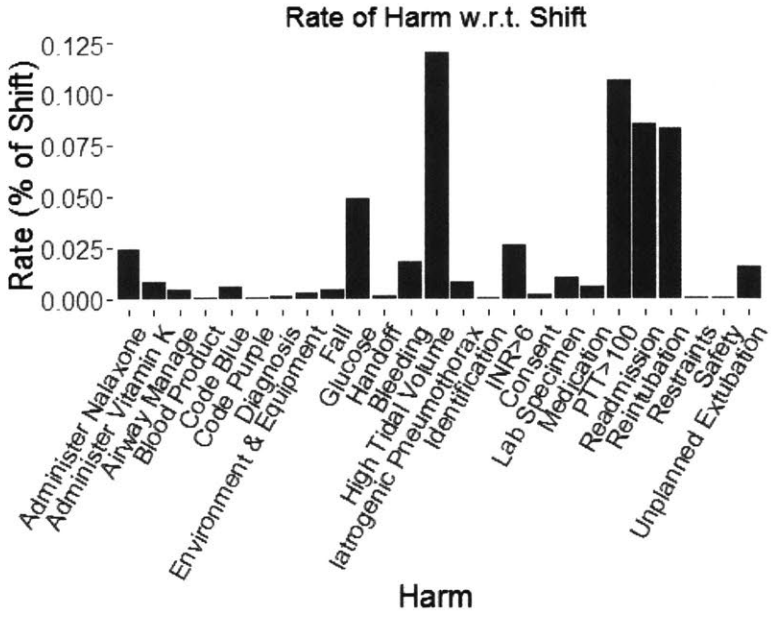



Figure 6-1: Percentage of Shifts with Harm in 2012 and 2013

The single harm rate with respect to shift, i.e., the percentage of shifts encountered harm in 2012 and 2013, is fairly low, with an average of 2.34%. See individual rates in Figure 6-1. After harm aggregation, we obtain that 44% of the shifts has a least one of these harms in the two years range, which helps alleviate the problem of rare events and unbalancedness in our original dataset. The average harm rate per patient across all shift is 0.032, with standard deviation of 0.05 and a maximum value of 0.14.

## 6.2 List of Risky and Safe States

We applied the statistical framework introduced in Chapter 5 to the following 6 subsets of shift observations:

| Unit | Shift Type | # of Shifts | # of Patients | Mean Harm Rate |
|---|---|---|---|---|
| Medical units | Day | 2918 | 6717 | 0.031 |
| Medical units | Night | 2919 | 6698 | 0.022 |
| Surgical units | Day | 1452 | 3746 | 0.048 |
| Surgical units | Night | 1452 | 3757 | 0.018 |
| CVICU | Day | 730 | 1711 | 0.093 |
| CVICU | Night | 729 | 1695 | 0.019 |

Medical units includes Finard 4, MICU 6, MICU 7, and CCU. Surical units includes TSICU and SICU.

The reason why we did not build one single model using all retrospective data from year 2012 and 2013 is that we observed the following two phenomena, when investigating different regression trees trained using random samples drawn from the dataset:

- The root node is mostly always *Night*, meaning it always divides the dataset into two subsets, with one containing only day shifts and the other containing only night shifts.
- The tree in general will split the data into medical units (Finard 4, MICU 6, MICU 7, CCU), surgical units (TSICU, SICU) and CVICU at some top level of the trees.

These two phenomena show that 1) day shifts and night shifts are different in nature (e.g., physician team is usually on call at night, almost no patients are admitted from the operating room at night, etc.), and 2) medical units, surgical units and CVICU are very different units in terms of the type of patients they admit, type of care the nurses need to provide, etc. So we divided the dataset into three different types of units, as well as day/night shifts, and performed analysis separately.

This also makes the comparison of mean harm rate between states and population more meaningful. For example, if there is some state in the medical units at night with mean harm rate of 0.03 and if we compare it to the overall harm rate across all shifts and units (which is 0.032), we would not consider it as a risky state. However, this rate is approximately 40% higher than the mean harm rate of all night shifts in medical units (0.022 as shown in the table above), which intuitively is a legitimate risky state.

Next we present a list of risky and safe states, for the 6 different sets, followed by their states description via driver value ranges.

### 6.2.1 Medical Units

- Day shifts, population average harm rate: 0.031

| State No. | # Shifts | $\mu^T$ | $\mu^K$ | $p$-value | Type |
|---|---|---|---|---|---|
| 6 | 184 | 0.019 | 0.020 | 3.957e-04 | Safe |
| 8 | 171 | 0.019 | 0.019 | 2.393e-05 | Safe |
| 15 | 201 | 0.025 | 0.025 | 1.919e-02 | Safe |
| 20 | 252 | 0.038 | 0.038 | 1.781e-02 | Risky |
| 22 | 163 | 0.041 | 0.041 | 4.597e-04 | Risky |
| 24 | 338 | 0.042 | 0.041 | 1.564e-04 | Risky |

Stable and Significant Model Outputs

| State No. | Driver | Lower Threshold | Upper Threshold |
|---|---|---|---|
| 6 | Admission | | 0.29 |
| | SOFA | | 6.22 |
| | Discharge | | 0.29 |
| | NP Ratio | | 0.80 |
| | TISS | | 14.17 |
| 8 | Admission | | 0.29 |
| | SOFA | | 6.22 |
| | Discharge | | 0.29 |
| | NP Ratio | 0.80 | |
| 15 | Admission | | 0.11 |
| | SOFA | 6.22 | |
| | Discharge | 0.10 | 0.23 |
| 20 | Admission | 0.11 | 0.19 |
| | SOFA | 6.22 | |
| | TISS | 20.58 | |
| 22 | Admission | 0.19 | 0.29 |
| | NP Ratio | | 0.72 |
| 24 | Admission | 0.29 | |

Description of States[1]

We can see that nurse utilization and patient acuity have impact on the likelihood of harm events in medical units. Both state 6 and 8 are safe states because 1) there are not many admissions and discharges and 2) average SOFA score is less than average. State 15, with fewer admissions/discharges but a higher SOFA score, is still a safe state. This means there is a tradeoff between the two types of drivers: while lowering nurse utilization may help decrease the overall risk level, having sicker patients in the unit may do the opposite.

State 20 shows a high nurse utilization, measured by the number of admissions as

[1]Blank indicates no lower or upper thresholds for the driver, i.e., can be as small as the minimal or as large as the maximum.

well as the nurse workload score TISS, together with high patient acuity, increases the likelihood of harm events. On the other hand, even if patients were not particularly sick in the units, the unit risk level could still be high when more nurses have to take care of two patients instead of one (i.e., the NP ratio is closer to 0.5), based on state 22. State 24 shows that when the number of new admissions is extremely high, the unit is in danger regardless of other conditions, possibly due to additional work for the nurses or nurses being unfamiliar with new patients, etc.

- Night shifts, population average harm rate: 0.022

| State No. | # Shifts | $\mu^T$ | $\mu^K$ | $p$-value | Type |
|---|---|---|---|---|---|
| 10 | 143 | 0.015 | 0.014 | 1.438e-02 | Safe |
| 11 | 192 | 0.013 | 0.012 | 2.373e-05 | Safe |
| 13 | 429 | 0.017 | 0.018 | 1.940e-02 | Safe |
| 17 | 196 | 0.026 | 0.026 | 3.566e-02 | Risky |
| 19 | 180 | 0.027 | 0.026 | 4.075e-02 | Risky |
| 24 | 289 | 0.036 | 0.036 | 6.187e-13 | Risky |

Stable and Significant Model Outputs

| State No. | Driver | Lower Threshold | Upper Threshold |
|---|---|---|---|
| 10 | Discharge | | 0.11 |
| | SOFA | | 8.72 |
| | TISS | | 24.60 |
| | Admission | | 0.15 |
| | NP Ratio | 0.72 | |
| 11 | Discharge | | 0.11 |
| | SOFA | | 8.72 |
| | TISS | | 24.60 |
| | Admission | 0.15 | 0.28 |
| | NP Ratio | 0.60 | |

Description of States

| State No. | Driver | Lower Threshold | Upper Threshold |
|---|---|---|---|
| 13 | Discharge | 0.11 | 0.21 |
| | SOFA | | 8.72 |
| | TISS | | 18.49 |
| | Admission | | 0.28 |
| | NP Ratio | 0.60 | |
| 17 | Discharge | | 0.21 |
| | SOFA | | 6.57 |
| | TISS | 24.60 | |
| 19 | Discharge | | 0.21 |
| | SOFA | 8.72 | |
| 24 | Discharge | 0.21 | |
| | Admission | 0.22 | |

Description of States

Night shifts in medical units share some similar patterns as day shifts. For example, in the safe states 10, 11 and 13, we can see that low nurse utilization (represented by small number of admissions/discharges, low TISS and closer to one-to-one NP Ratio) as well as low patient acuity (low SOFA score) result in a lower harm rate.

State 17 depicts the scenario where when nurses are extremely busy, the unit risk level increases regardless of patient acuity, while state 19 shows that even if nurses are not particularly busy, having more sicker patients will make the unit less safe. State 24 is similar as above, indicating high level of patient movement will have an adverse impact to the unit risk level.

### 6.2.2 Surgical Units

- Day shifts, population average harm rate: 0.048

| State No. | # Shifts | $\mu^T$ | $\mu^K$ | $p$-value | Type |
|---|---|---|---|---|---|
| 4 | 165 | 0.040 | 0.040 | 1.830e-02 | Safe |
| 5 | 168 | 0.026 | 0.026 | 2.998e-05 | Safe |
| 11 | 251 | 0.064 | 0.061 | 2.081e-05 | Risky |

Stable and Significant Model Outputs

| State No. | Driver | Lower Threshold | Upper Threshold |
|---|---|---|---|
| 4 | SOFA | | 5.62 |
| | NP Ratio | 0.69 | |
| | TISS | | 18.71 |
| 5 | SOFA | | 5.62 |
| | NP Ratio | 0.69 | |
| | TISS | 18.71 | |
| 11 | SOFA | 5.62 | |
| | NP Ratio | 0.64 | |
| | TISS | | 21.77 |
| | Discharge | 0.17 | |

Description of states

Patient acuity has a larger impact on unit risk level in surgical units comparing to medical units. We can see that regardless of TISS, both state 4 and 5 indicate that a lower SOFA will result in a safer unit. Even though the NP ratio is closer to one-to-one for these two states, state 11 shows that a higher average SOFA score increases the likelihood of harm events under similar NP ratio conditions.

- Night shifts, population average harm rate: 0.018

| State No. | # Shifts | $\mu^T$ | $\mu^K$ | $p$-value | Type |
|---|---|---|---|---|---|
| 3 | 232 | 0.009 | 0.009 | 3.344e-06 | Safe |
| 13 | 174 | 0.026 | 0.026 | 7.209e-04 | Risky |
| 14 | 149 | 0.028 | 0.029 | 4.301e-05 | Risky |

Stable and Significant Model Outputs

| State No. | Driver | Lower Threshold | Upper Threshold |
|---|---|---|---|
| 3 | Discharge | | 0.06 |
| | SOFA | | 6.11 |
| 13 | Discharge | 0.06 | |
| | SOFA | 7.02 | |
| 14 | Discharge | 0.21 | |

Description of states

Night shifts in surgical units are much safer than day shifts because surgical units typically admit new patients from operating room during the day, resulting high patient acuity. In contrast, surgical units rarely admit new sick patients at night since surgeries are usually not conducted at night, which could be why other drivers also play important roles at night. As we can see in the table above, in additional to SOFA score, number of discharges have a relatively big correlation with the likelihood of harm events in surgical units.

### 6.2.3 CVICU

- Day shifts, population average harm rate: 0.093

| State No. | # Shifts | $\mu^T$ | $\mu^K$ | $p$-value | Type |
|---|---|---|---|---|---|
| 2 | 131 | 0.035 | 0.036 | 5.231e-16 | Safe |
| 3 | 93 | 0.062 | 0.058 | 1.041e-03 | Safe |
| 7 | 150 | 0.147 | 0.147 | 1.578e-11 | Risky |
| 10 | 132 | 0.111 | 0.110 | 4.158e-03 | Risky |

Stable and Significant Model Outputs

| State No. | Driver | Lower Threshold | Upper Threshold |
|---|---|---|---|
| 2 | Admission | | 0.15 |
| | SOFA | | 7.29 |

Description of states

| State No. | Driver | Lower Threshold | Upper Threshold |
|---|---|---|---|
| 3 | Admission | | 0.15 |
| | SOFA | 7.29 | |
| 7 | Admission | 0.15 | |
| | NP Ratio | | 0.86 |
| | SOFA | 6.85 | |
| 10 | Admission | 0.15 | |
| | NP Ratio | 0.86 | |
| | SOFA | 7.18 | |

Description of states

CVICU sees a huge harm rate for day shifts, much higher than the overall average harm rate of all units, because of the special type of patients the unit takes care for. During the day, number of admissions is the dominant driver affecting the likelihood of harm events in the unit. Regardless of patient acuity, the unit is safe when there is a small number of admission, according to state 2 and 3. State 7 and 10, on the other hand, show that a large number of admission together with high patient acuity results in a higher unit harm rate.

- Night shifts, population average harm rate: 0.019

| State No. | # Shifts | $\mu^T$ | $\mu^K$ | $p$-value | Type |
|---|---|---|---|---|---|
| 2 | 181 | 0.010 | 0.010 | 9.623e-05 | Safe |
| 7 | 106 | 0.026 | 0.025 | 2.659e-02 | Risky |

Stable and Significant Model Outputs

Harm rate is much lower during night shifts in CVICU and is affected by both nurse utilization and patient acuity. State 2 shows that the unit is safe when both SOFA and number of admissions are low. High risk level is, on the other hand, associated with state 7, when the number of discharges is higher and when patients are sicker (SOFA scores are higher and more nurses need to take care of one patient).

| State No. | Driver | Lower Threshold | Upper Threshold |
|---|---|---|---|
| 2 | SOFA | | 7.46 |
| | Admission | | 0.08 |
| 7 | SOFA | 7.46 | |
| | Discharge | 0.08 | |
| | NP Ratio | 0.78 | |

Description of states

## 6.3 Summary of Insights

In general, we discover the following insights:

- Surgical and medical units are generally different because they treat very different types of patients
- CVICU usually carries a high risk for harm events than other units due to the type of patients they admit
- Day shifts in general have high harm rate than night shifts
  - There are more activities occurring during the day shift comparing to the night shift, such as surgery operations, drawing lab specimen, moving patients to other departments for check-ups, etc., which adds complexity to the unit and thus resulting a higher likelihood of harm events
  - The care team is quite different between day shifts and night shifts. For example, doctors would stay in unit during the day and check on patients regularly, while at night they are usually on call.
- Patient acuity is not always correlated with higher likelihood of harm events
  - SOFA score has less impact in medical units comparing to surgical units and CVICU
  - There are some states in which SOFA has no impact on the unit risk level
- Higher nurse utilization often seems correlated with high rate of harm events.

Nurse utilization could be measured through the following drivers:

- Higher TISS score
- A lower NP ratio (fewer nurses taking care of more patients)
- Frequent patient flow (large number of admissions and discharges causing additional work)

The latter point is particularly important because patient flow, unlike patient acuity, is something that can be modified and optimized. Therefore, the risky (safe) states that are observed and summarized in this project are essentially valuable to the hospital since they can make feasible and realistic intervention plans with controllable items ahead of time to eliminate risky states, create safe states, and improve patient safety.

# Chapter 7

# Conclusion

The Risky States approach is a more comprehensive approach for ICU patient risk assessment and the Risky States model we designed in this project establishes a more robust framework that does not only model the relationship between ICU system conditions and adverse events faced by patients, but is also able to provide what and how a set of conditions are affecting patient risk level. With such property, the method provides some major useful information needed to make effective and efficient intervention and mitigation plans in a timely manner.

This approach is built upon the theoretical understanding in human error recognition and system engineering, and is capable of providing rigorous yet clinically intuitive insights through the utilization of various statistical methodologies. We hope that this model does not only provide a robust tool for the hospital to assess patient risk and improve patient safety in ICUs, but also motivates for better operational strategic planning and inspires other innovative researches ideas.

We encountered several challenges during the designing process, but we have management to find solutions to solve these problems. While there is still room to improve the model, it is now being implemented by Aptima on the BIDMC ICU tablet application and will be further validated through human interactions and integrated with the two other work streams under the Moore grant. We hope that it can help ICU clinicians and staffing members with a more comprehensive understanding of their surrounding environment and enables them to bring better care quality to patients.

## 7.1 Next Steps

We identify several possible next steps for different stakeholders in this project.

**BIDMC Leadership:** by understanding how environmental conditions can influence individual's wellbeing, try to adopt changes to the overall ICU operational plan, which may benefit the risk management practice.

**Bed Assignment Team:** create certain type of automated and optimized patient assignment strategy with guidelines suggested by the model output since patient flow has a high impact on ICU risk level.

**Staffing Assignment Team** : establish an effective way to keep track of the care team members throughout patients' stay, which is considered as an important risk driver but was not included in current analysis because no historical data can be obtained.

**ICU Staff:** 1) utilize the model to understand current unit conditions on a regular basis; 2) adjust working protocols or priorities according to environmental changes.

**IT System:** with a large number of problems encountered during data extracting and processing, it is advised that a better IT infrastructure be created or upgraded that enhances the communication across different software or database.

**Project researchers:** 1) work out a better metric to measure the nurse workload as the current version of TISS only captures a subset of activities that the nurses perform each day; 2) it is crucial to validate the model through clinical trials and gain insights from working clinicians, who can help identify model misses and drawbacks; 3) more work need to be done in order to improvement model performance as well as compatibility for more general ICU settings.

## 7.2 Future Research

In addition to using risky state model as a recommendation system for eliminating risky states, BIDMC is also interested in other functionalities of the models such as

real time harm prediction. This requires a prospective approach in that using current known information to predict future unit risks, which is more than a direct adoption of our model into the real time database. Based on active discussion with BIDMC experts, we started to seek possibilities of aggregating harms and drivers into a 4-hour window instead of to the shift level. This will allow capturing state changes in a more timely manner, is compatible with many clinical practices and potentially offer the possibility to model beyond concurrent states and risks.

# Appendix A

# BIDMC Information System

| Database Name | Functionality |
|---|---|
| Admission Table (ADT) | Administrative database recording patient service type, movements (admissions, transfers, discharges), origins and destinations, with timestamps |
| MetaVision | Clinical and vital sign monitoring application database |
| Omnicell | Automated medication dispensing database |
| RLSolution | Voluntary incident reporting database |
| Infection Control | Patient infection history database |
| Labs | Clinical database for lab tests and results |
| Online Medical Record (OMR) | Patient medication history database |
| Provider Order Entry (POE) | Physician-orderer medication history database |
| Codes | Emergency codes alert database |
| Staffing | Nurse working history and administrative records database |
| EU Critical | Database identifying patients admitted without any medical history |

Table A.1: BIDMC Information System Electronic Database Overview

# Appendix B

# Harm Characteristics Table



Table B.1: Overall Burden of Harm Key Characteristics

| **Harm** | **Metric** | **Carry-over Exemption** | **Relavancy Patient Cohort** | **Type** | **Shift Attribution** |
|---|---|---|---|---|---|
| CLABSI | NHSN standard and identified by Infection Control (IC) [5] | N/A | w/ central lines | Long-term | [First eligible shift, Identification by IC] |
| VAE | NHSN standard and identified by Infection Control [7] | N/A | Intubated | Long-term | [First eligible shift, Identification by IC] |

Continued on next page

Table B.1: Overall Burden of Harm Key Characteristics

| **Harm** | **Metric** | **Carry-over Exemption** | **Relavancy Patient Cohort** | **Type** | **Shift Attribution** |
|---|---|---|---|---|---|
| DVT-PE | Coded upon discharge with relevant tests taken in ICU | Test taken within 24 hours of admission | Any | Long-term | [ICU Admission, Test] |
| ARDS | Tidal volume > personal ideal tidal volume | N/A | Intubated | Instantaneous | Current |
| Delirium | CAM score: Positive | Score recorded within 12 hour of admission | Any | Instantaneous | Current |
| CAUTI | NHSN standard and identified by Infection Control [6] | N/A | w/ Foley catheter | Long-term | [First eligible shift, Identification by IC] |
| Code Blue | Report by medical staff | N/A | Any | Instantaneous | Current |
| Positive C. difficile | C. difficile lab result: Positive | Test taken within 12 hours of admission | Any | Long-term | [ICU Admission, Test] |
| Positive blood culture | Blood culture lab result: Positive | Test taken within 12 hours of admission | Any | Long-term | [ICU Admission, Test] |
| Oversedation | Goal RASS < Actual RASS | First pair of RASS ignored | On sedative drips | Instantaneous | Current |

Continued on next page

Table B.1: Overall Burden of Harm Key Characteristics

| **Harm** | **Metric** | **Carry-over Exemption** | **Relavancy Patient Cohort** | **Type** | **Shift Attribution** |
|---|---|---|---|---|---|
| Bleeding | Abrupt drop in 2 consecutive Hemoglobin lab values (diff > 4 within 24 hours) | N/A | Any | Long-term | [Previous test, Detecting test] |
| Bleeding | PTT blood test result > 100 seconds | Test within 6 hours of admission | Received Heparin | Instantaneous | Shift Heparin given |
| Bleeding | INR blood test result > 6 | Test within 48 hours of admission | Received Warfarin | Instantaneous | Shift Warfarin given |
| Hypoglycemia | Glucose lab/fingerstick < 50 mg/dl | Test within 2 hours of admission | On Insulin drip | Instantaneous | Current |
| Administer Vitamin K | Vitamin K being administered within 48 hours of Warfarin | N/A | Received Warfarin | Instantaneous | Shift Warfarin given |
| Administer Naloxone | Nalaxone injection being administered | N/A | Received Narcotics | Instantaneous | Shift Narcotics given |
| Doubled Creatinine | Creatinine blood test result > 2× Creatinine upon admission and > 2 mg/dL | N/A | Any | Long-term | [Previous test, Detecting test] |

Continued on next page

Table B.1: Overall Burden of Harm Key Characteristics

| **Harm** | **Metric** | **Carry-over Exemption** | **Relavancy Patient Cohort** | **Type** | **Shift Attribution** |
|---|---|---|---|---|---|
| Chest tube insertion at bedside | Chest tube insertion within 24 hour of placement of central line | N/A | w/ qualified central lines | Instantaneous | Shift of central line placement |
| Readmission | Time diff b/w ICU discharge & readmission $\leq$ 48 hours | N/A | Any | Instantaneous | Shift of discharge |
| Reintubation | Time diff b/w extubation & reintubation $\leq$ 12 hours | N/A | Intubated | Instantaneous | Shift of extubation |
| Unplanned extubation | Report by medical staff | N/A | Intubation | Instantaneous | Current |
| Skin tissue, infection | Report by medical staff | Coded for pressure ulcer upon admission | Any | Long-term | [Admission, Detection] |
| Other IRS events | Report by medical staff | N/A | Any | Instantaneous | Current |

# Appendix C

# Driver Calculation References

## C.1 Boarding Patients

| Service | FICU | TSICU | SICU-B | SICU-A | MICU-6 | CVICU-A | CVICU -B | MICU-7 | CCU |
|---|---|---|---|---|---|---|---|---|---|
| MED | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| CSURG | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| SURG | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| NSURG | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| VSURG | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| NMED | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CMED | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| GU | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ORTHO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| TRAUM | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| TSURG | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| OMED | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GYN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| OBS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PSURG | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| ENT | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

Table C.1: BIDMC ICU Service Matching Overview[1]

[1]Note that there are two possibilities when a pair of service and ICU gets a value 0: 1) the assigned unit is considered as a home unit to the patient with that service; 2) it is never done in practice to assign patients with the service to that ICU. For example, it is never possible for FICU to have any boarding patients due to the second reason.

## C.2 Sequential Organ Failure Assessment (SOFA) Score

| System | Respiratory | Neurological | Cardiovascular | Hepatic | Coagulation | Renal | Points |
|---|---|---|---|---|---|---|---|
| **Measurement** | $PaO_2/FiO_2$ (mmHg) | Glasgo Coma Scale (GCS) | Mean arterial pressure or administration of vasopressors | Total Bilirubin (mg/dl) | Platelets $\times 10^3$/mcl | Creatinine (mg/dL) (or urine output) | |
| **Conditions** | $\geq 400$ | 15 | No hypotension | $< 1.2$ | $\geq 150$ | $< 1.2$ | 0 |
| | $< 400$ | 13 - 14 | MAP $< 70$ mm/Hg | 1.2 - 1.9 | $< 150$ | $1.2 - 1.9$ | 1 |
| | $< 300$ | 10 - 12 | dopamine $\leq 5$ or dobutamine (any dose) | 2.0 - 5.9 | $< 100$ | $2.0 - 3.4$ | 2 |
| | $< 200$ and mechanically ventilated | 6 - 9 | dopamine $> 5$ or epinephrine $\leq 0.1$ or norepinephrine $\leq 0.1$ | 6.0 - 11.9 | $< 50$ | $3.5 - 4.0$ (or $< 500$ ml/d) | 3 |
| | $< 100$ and mechanically ventilated | $< 6$ | dopamine $> 15$ or epinephrine $> 0.1$ or norepinephrine $> 0.1$ | $\geq 12.0$ | $< 20$ | $\geq 5.0$ (or $< 200$ ml/d) | 4 |

Table C.2: Standard SOFA Metric Comparison Table for Points Assignment

### C.2.1 Modified SOFA Score

In case when $PaO_2$ and $FiO_2$ are not documented for patients not ventilated and/or no arterial blood gas being drawn, we use $SpO_2/FiO_2$ ratio, obtained by dividing $SpO_2$ by the fraction of inspired oxygen, $FiO_2$, as a surrogate for $PaO_2/FiO_2$, which has been separately validated [10].

When $FiO_2$ is not documented, the fraction of inspired oxygen of 0.21 is used as an estimate of $FiO_2$ for patients on Room Air. Non-ventilated patients who are receiving oxygen via nasal cannula or mask will have "$O_2$ Flow" or "$O_2$ Flow (additional cannula)" documented, which records the value of oxygen flow in Liter Per Minute. This number is usually between 1 to 6 LPM, and we multiply it by 0.03 and add the result to 0.21 of ambient air to obtain an estimate for $FiO_2$.

Using $SpO_2/FiO_2$ ratio as an alternative in SOFA calculation also requires taking into account PEEP, the positive end-expiratory pressure. This is a measure to learn how sick the lungs are: the more PEEP needed, the sicker the lungs. PEEP is generally available for all patients and will help determine what is the best comparing range to use for correctly assessing patients' respiratory condition.

To summarize, follow the next 4 rules to calculate the respiratory component of MSOFA according to Table C.3:

- For ventilated patients with $PaO_2$ documented, use $PaO_2/FiO_2$
- For ventilated patients with no $PaO_2$ documented, use $SpO_2/FiO_2$ and compare results based on their PEEP value
- For non-ventilated patients who are on nasal cannula, mask, or other respiratory device, use $SpO_2/FiO_2$ assuming PEEP $< 8$, and approximate $FiO_2$ by $0.21+0.3*O_2$ Flow
- For non-ventilated patient without any respiratory device, use $SpO_2/FiO_2$ assuming PEEP $< 8$, and set $FiO_2$ to 0.21

| $PaO_2/FiO_2$ | $SpO_2/FiO_2$ Ratio | | | Points |
|---|---|---|---|---|
| | PEEP $< 8$ | PEEP 8 - 12 | PEEP $> 12$ | |
| $\geq 400$ | $\geq 457$ | $\geq 515$ | $\geq 425$ | 0 |
| $< 400$ | $< 457$ | $< 515$ | $< 425$ | 1 |
| $< 300$ | $< 370$ | $< 387$ | $< 332$ | 2 |
| $< 200$ | $< 240$ | $< 259$ | $< 234$ | 3 |
| $< 100$ | $< 115$ | $< 130$ | $< 129$ | 4 |

Table C.3: Respiratory Metric Comparison Table under MSOFA[2]

[2]Note that according to the study by Dr. Daniel Talmor's group at BIDMC, "the original $SpO_2/FiO_2$ ratio as published would require a $SaO_2 > 110\%$", so they again modify the score in order to accept a $SaO_2$ of 96% or greater on room air as normal.

## C.3 Therapeutic Intervention Scoring System (TISS)

| **Basic Activities** | Points | | **Ventilatory Support** | Points |
|---|---|---|---|---|
| Standard Monitoring (All Patients) | 5 | | On a Ventilator | 5 |
| Routine Lab Draw (All Patients) | 1 | | O2 delivery assistance | 1 |
| Routine Medication (All Patients) | 2 | | Has a trache | 1 |
| IV insulin/ meds with extensive monitoring | 4 | | Chest CT (All Patients) | 1 |
| Routine dressing changes (All Patients) | 1 | | | |
| Care of drains (All Patients) | 3 | | **Cardiovascular Support** | |
| Pressure ulcer | 1 | | Single vasoactive medication | 3 |
| | | | Multiple vasoactive medications | 4 |
| **Renal Support** | | | 1.5L IVF/blood products per shift | 4 |
| CRRT | 8 | | Arterial catheter (in access line/invasive) | 2 |
| Measuring Urine Output | 2 | | PA Catheter, LVAD, Tandem heart | 8 |
| Diuresing (Lasix) | 3 | | Impella,PiCO, ECMO, Alsius, Arctic Sun | 8 |
| | | | Heart Mate,Blakemore, Massive Transfusion | 8 |
| **Metabolic support** | | | Central venous line | 2 |
| Acidosis/Alkalosis | 4 | | Code blue in last 24hrs | 3 |
| TPN/OPN | 2 | | | |
| Tube Feeds | 3 | | **Specific Interventions** | |
| | | | Single Procedure done in ICU | 3 |
| **Neurological Support** | | | Multiple procedures done in ICU | 5 |
| ICP Drain | 4 | | Travel (OR, Cath lab, ERCP) | 5 |

Table C.4: Scoring System for Nursing Workload

# Bibliography


[1] L. B. Andrews, C. Stocking, T. Krizek, L. Gottlieb, C. Krizek, T. Vargish, and M. Siegler. An alternative strategy for studying adverse events in medical care. *Lancet*, 349(9048):309–313, 1997.

[2] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[3] A. D. Calabrese, B. L. Erstad, K. Brandl, J. F. Barletta, S. L. Kane, and D. S. Sherman. Medication administration errors in adult patients in the icu. *Intensive care medicine*, 27(10):1592–1598, 2001.

[4] Y. Donchin, D. Gopher, M. Olin, Y. Badihi, M. Biesky, C. L. Sprung, R. Pizov, and S. Cotev. A look into the nature and causes of human errors in the intensive care unit. *Critical Care Medicine*, 23:294–300, 1995.

[5] U.S. Centers for Disease Control and Prevention. Bloodstream infection event (central line-associated bloodstream infection and non-central line-associated bloodstream infection). `http://www.cdc.gov/nhsn/PDFs/pscManual/4PSC_CLABScurrent.pdf`. Accessed: 2014-09-30.

[6] U.S. Centers for Disease Control and Prevention. Urinary tract infection (catheter-associated urinary tract infection [cauti] and non-catheter-associated urinary tract infection [uti]) and other urinary system infection [usi]) events. `http://www.cdc.gov/nhsn/PDFs/pscManual/7pscCAUTIcurrent.pdf`. Accessed: 2014-09-30.

[7] U.S. Centers for Disease Control and Prevention. Ventilator-associated event (vae). `http://www.cdc.gov/nhsn/PDFs/pscManual/10-VAE_FINAL.pdf`. Accessed: 2014-09-30.

[8] M. Garrouste-Orgeas, F. Philippart, C. Bruel, A. Max, N. Lau, and B. Misset. Overview of medical errors and adverse events. *Annals of intensive care*, 2(1):2, 2012.

[9] Atul Gawande. The checklist, if something so simple can transform intensive care, what else can it do? `http://www.newyorker.com/reporting/2007/12/10/071210fa_fact_gawande`. Accessed: 2014-09-30.

[10] C. K. Grissom, S. M. Brown, K. G. Kuttler, J. P. Boltax, J. Jones, A. R. Jephson, and J. F. Orme. A modified sequential organ failure assessment (msofa) score for critical care triage. *Disaster medicine and public health preparedness*, 4(4):10.1001, 2010.

[11] Linda T. Kohn, Janet Corrigan, and Molla S. Donaldson. *To err is human: building a safer health system.* Washington, D.C: National Academy Press, 1999.

[12] B. J. Kopp, B. L. Erstad, M. E. Allen, A. A. Theodorou, and G. Priestley. Medication errors and adverse drug events in an intensive care unit: direct observation approach for detection. *Critical care medicine*, 34(2):415–425, 2006.

[13] CO: HealthGrades Lakewood. *Medical errors cost U.S. $8.8 billion, result in 238,337 potentially preventable deaths, according to HealthGrades Study [press release].* The HealthGrades Press, 2008.

[14] G. L Langley, R. Moen, K. M. Nolan, T. W. Nolan, C. L. Norman, and L. P. Provost. *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance.* San Francisco: Jossey-Bass Publishers, 2009.

[15] Lucian L. Leape. Error in medicine. *JAMA*, 272(23):1851–1857, 1994.

[16] A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[17] Yiyin Ma. Modeling of icu nursing workload to inform better staffing decisions, 2015.

[18] Susan Mayor. English nhs to set up new reporting system for error. *BMJ*, 320:1689, 2000.

[19] Pamela Mazzocato, Carl Savage, Mats Brommels, Håkan Aronsson, and Johan Thor. Lean thinking in healthcare: a realist review of the literature. *Quality and Safety in Health Care*, 19(5):376–382, 2010.

[20] D. Reis Miranda, R. Moreno, and G. Iapichino. Nine equivalents of nursing manpower use score (nems). *Intensive care medicine*, 23(7):760–765, 1997.

[21] Dinis R. Miranda, Angelique de Rijk, and Schaufeli Wilmar. Simpli

ed therapeutic intervention scoring system: the tiss-28 items - results from a multicenter study. *Critical care medicine*, 24(1):64–73, 1996.

[22] Institute of Healthcare Improvement. Ihi global trigger tool for measuring adverse events (second edition). `http://app.ihi.org/webex/gtt/ihiglobaltriggertoolwhitepaper2009.pdf`, 2009. Accessed 2012-10-10.

[23] Institute of Medicine. *Crossing the Quality Chasm.* 2001.

[24] Joint Commission on Accreditation of Hospitals. Sentinel events data. `http://www.jointcommission.org/assets/1/18/event_type_by_year_1995-2q2013.pdf`, 2000.

[25] R. H. Palmer and M. E. Adams. Considerations in defining quality in health care. *Paper prepared for the Institute of Medicine Study to Design a Strategy for Quality Review and Assurance in Medicare*, 1988.

[26] Peter J. Pronovost, David A. Thompson, Christine G. Holzmueller, Lisa H. Lubomski, and Laura L. Morlock. De

ning and measuring patient safety. *Critical care clinics*, 21(1):1–19, 2005.

[27] Jens Rasmussen and Aage Jensen. Mental procedures in real-life tasks: a case study of electronic trouble shooting. *Ergonomics*, 17(3):293–307, 1974.

[28] James T. Reason. *Human Error.* Cambridge: Cambridge University Press, 1990.

[29] S. A. Ridley, S. A. Booth, and C. M. Thompson. Prescription errors in uk critical care units. *Anaesthesia*, 59(12):1193–1200, 2004.

[30] J. M. Rothschild, C. P. Landrigan, J. W. Cronin, R. Kasushal, S. W. Lockley, E. Burdick, P. H. Stone, C. M. Lilly, J. T. Katz, C. A. Czeisler, and D. W. Bates. The critical care safety study: the incidence and nature of adverse events and serious medical errors in intensive care. *Critical care medicine*, 33(8):1694–1700, 2005.

[31] W. B. Runciman. Lessons from the australian patient safety foundation: setting up a national patient safety surveillance system - is this the right model? *Quality and Safety in Health Care*, 11(3):246–251, 2002.

[32] John W. Senders and Neville P. Moray. *Human Error: Cause, Prediction and Reduction.* Lawrence Erlbaum Associates, 1991.

[33] Steven Q. Simpson, Douglas A. Peterson, and Amy R. O'Brien-Ladner. Development and implementation of an icu quality improvement checklist. *AACN advanced critical care*, 18(2):183–189, 2007.

[34] Intensive Care Society. What is intensive care. Accessed: 2014-09-30.

[35] James A. Taylor, Dena Brownstein, Dimitri A. Christakis, Susan Blackburn, Thomas P. Strandjord, Eileen J. Klein, and Jaleh Sha

i. Use of incident reports by physicians and nurses to document medical errors in pediatric patients. *Pediatrics*, 114(3):729–735, 2004.

[36] Ho Tin Kam. A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis and Applications*, 5(2):102–112, 2002.

[37] J. Adam Traina. Diagnosing intensive care units and hyperplane cutting for design of optimal production systems, 2015.

[38] A. Valentin, M. Capuzzo, B. Guidet, R. Moreno, B. Metnitz, P. Bauer, and P. Metnitz. Errors in administration of parenteral drugs in intensive care units: multinational prospective study. *BMJ*, 338:b814, 2009.

[39] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, 22(7):707–710, 1996.

[40] Maartje De Vos, Wilco Graafmans, Els Keesman, Gert Westert, and Peter H. J. van der Voort. Quality measurement at intensive care units: which indicators should we use? *Journal of critical care*, 22(4):267–274, 2007.