Spring 2021

Development, Implementation and Evaluation of Medical Decision Support Systems Based on Mortality Prediction Algorithms from an Operations Research Perspective.

Junchao Ma
Yale University Graduate School of Arts and Sciences, junchaoma625@gmail.com

Abstract

Development, Implementation and Evaluation of Medical Decision Support Systems Based on Mortality Prediction Algorithms From an Operations Research Perspective.

Junchao Ma

2021

Wide implementation of electronic health record systems provides rich data for personalized medicine. One topic of great interest is to develop methods to assist physicians in prognosis for example mortality. While many studies have reported on various new prediction models and algorithms there is relatively little literature on if and how these new prediction methods translate into actual benefits. My dissertation consists of three theses that aims at filling this gap between prognostic predictions and clinical decisions in end-of-life care and intensive care settings.

In the first thesis, we develop an approach to using temporal trends in physiologic data as an input into mortality prediction models. The approach uses penalized b-spline smoothing and functional PCA to summarize time series of patient data. we apply the methodology in two settings to demonstrate the value of using the ”shapes” of health data time series as a predictor of patient prognosis. The first application a mortality predictor for advanced cancer patients that can help oncologists decide which patients should stop aggressive treatments and switch to palliative care such as that provided in hospice. The second one is a real-time near term mortality predictor for MICU patients that can work as an early alarm system to guide timely interventions.

In the second thesis, we investigate the integration of a prediction algorithm with physician decision making, focusing on the advanced cancer patient setting. We design a retrospective study to compare prognoses made by doctors and those that would be recommended by the IMPAC algorithm developed in Chapter 1. We used the doctor's discharge decision as a proxy of what they predict the patient as dying in 90 days and show that doctor's predictions tend to very conservative. Although IMPAC on its own does not perform better than doctors in terms of precision and recall, we find that IMPAC and doctors identify significantly different group of positive cases. IMPAC and doctors are also good at identifying very different groups of patients in terms of survival time. We propose a new way to augment decisions of doctors with IMPAC. At the same recall, the augment method identifies 43% more patients close to death than the doctors do. We also estimate potential hospitalizations and hospital length of stays avoided if the doctors use augmented procedure instead of acting on their own beliefs.

In the third thesis, we look at the integration of a prediction algorithm with physician decision making, focusing on the ICU setting. We use a POMDP framework to evaluate how decision support systems based on ICU mortality predictions can help physicians allocate time to inspect the patients at highest risk of death. We assume physicians have limited time and seek to optimally allocate it to patients in order to minimize their mortality rate. Physicians can do Bayesian updates on observations of patient health state. A prediction algorithm can augment this process by sending alerts to physicians. We represent the algorithm by an arbitrary point on an ROC curve representing a particular alert threshold. We study two approaches to using the algorithm input: (1) Belief based policy (BBP) that integrates algorithm outputs using Bayesian updating; (2) Alarm triggered policy (ATP) where the physician responds only to the algorithm without updating, and compare them to benchmarks that do not rely on the algorithm at all. By running simulations, we explore how the accuracy of predictions can translate into lower mortality rates.

Development, Implementation and Evaluation of Medical Decision Support Systems Based on Mortality Prediction Algorithms From an Operations Research Perspective.

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Junchao Ma

Dissertation Director: Edieal J. Pinker

June, 2021

Contents

Acknowledgements **ix**

1 Introduction **1**

2 Redesigning Patient Health Status Scores using Mortality Predictions Algorithms, with Applications in End-of-Life Care and Intensive Care **3**

2.1 Introduction 3

2.2 Methodology 6

2.2.1 Functional Data Analysis 6

2.2.2 Mortality Prediction Models 10

2.3 End-of-life project: Development of Imminent Mortality Predictor in Advanced Cancer(IMPAC) 13

2.3.1 Background and Overview 13

2.3.2 Data Description 14

2.3.3 Prediction model 19

2.3.4 Model Evaluation 20

2.3.5 Determination of Potential Avoidable Care 24

2.3.6 Conclusion 26

2.4 Using the Shapes of Clinical Data Trajectories to Predict Mortality in Intensive Care Units 27

2.4.1 Overview and Background 27

2.4.2 Data Description 29

2.4.3 Comparing different models 34

2.4.4 Results 36

2.4.5 Conclusion 39

2.5 Discussion 41

3 Improving End of Life Care by Augmenting Physician Discharge Decision Process with Mortality Prediction Algorithms. 43

3.1 Introduction 43

3.2 Data Description 46

3.2.1 Study Population 46

3.2.2 Variable definitions 46

3.2.3 Prediction Model 47

3.3 Accuracy Comparison of IMPAC and Doctors 51

3.3.1 Comparison Criteria 52

3.3.2 Algorithm Performance 53

3.3.3 Overlap between Two Predictors 54

3.4 Augmentation Oncologists' End-of-Life Decision with IMPAC 55

3.4.1 Pooling Method 56

3.4.2 Agreement Method 57

3.4.3 Augmented Method 58

3.4.4 Doctors' Discharge Decision as a new Covariate to IMPAC 59

3.4.5 Evaluation Based on Precision Recall curve 61

3.4.6 Survival time analysis 62

3.5 Potential Avoidable Aggressive Treatments 63

3.6 Discussion 67

4 A POMDP Framework to Evaluate Decision Support System's Contribution to ICU patient Monitoring 68

4.1 Introduction 68

4.2 POMDP Model Setup 72

4.2.1 ICU setting 73

4.2.2 State Space 74

4.2.3 Decision Variable 76
4.2.4 State Transition Rule 76
4.2.5 Objective 78
4.3 Baseline Model 78
4.3.1 Event timeline 78
4.3.2 Bellman Equation for Baseline Model 79
4.3.3 Solving the Baseline POMDP 80
4.3.4 Numerical Experiment 81
4.3.5 Estimation of Model Parameters 84
4.4 Model with Binary Decision Support System 87
4.4.1 Policy 87
4.4.2 Event Timeline for Belief Base Policy Model 87
4.4.3 Bellman Equation for Belief Based Policy Model 88
4.4.4 Policy Iteration 91
4.4.5 Minimum Mortality Rate under Inspection Rate Constraints 92
4.5 Minimize mortality under inspection rate constraint over ROC 93
4.5.1 Receiver Operating Characteristic(ROC) 94
4.5.2 Get (α, β) plot based on ROC 95
4.5.3 Three ROC examples 96
4.5.4 Minimum Mortality Rate under Inspection Rate Constraints 98
4.6 Discussion 100

List of Figures

2.1 Empirical pdf of RI 17

2.2 B-spline Basis to Fit RI Curves 18

2.3 First 3 Principal Components of RI Curves 20

2.4 Variable Coefficients in Cox Proportional Hazard Model 21

2.5 Average Precision Recall Curve for 30,60,90,180-day Mortality Predictions. 23

2.6 Survival Time Distribution for Different Groups of Patients. 24

2.7 Observational Units and Prediction Targets. 31

2.8 B-splines Used to Fit 24-hour Time Series for Each Observational Units. . . 32

2.9 Fitted RI Curve Using Penalized B-splines. 33

2.10 Average PR Curves for Model M1-M4. 37

2.11 Average ROC Curves for Model M1-M4. 38

2.12 Variable Importance Plot for Model M4. 39

3.1 Boxplot: Number of RI Observations During 48-hour Time Window 49

3.2 Set of 13 Basis Functions 50

3.3 Average Precision Recall Plots of IMPAC and Doctors 54

3.4 Proportion Sent to Hospice among 100 Neighbors with Closest IMPAC score 56

3.5 Flow Chart for Pooling Method. 58

3.6 Flow Chart for Agreement Method. 58

3.7 Demonstration of Precision-Recall Curve for Augmented Method. 59

3.8 Coefficient and P-value for Augmented IMPAC 60

3.9 Average Precision-Recall Curves for 4 Classification Algorithms. 61

3.10 Survival Time Distribution in Box Plots 64

3.11 Empirical c.d.f. of Remaining Survival Time When Patient was Sent to Hospice for the First Time 66

4.1 State Transition Diagram. 77

4.2 Belief Cutoff $\bar{b}$ vs Cost of Inspection C_I 82

4.3 Belief Cutoff vs Cost of Inspections by Varying f 83

4.4 Belief cutoff vs Cost of Inspections by Varying m 83

4.5 Belief Cutoff vs Cost of Inspections by Varying p_2 84

4.6 RI Distribution for Those not Dying in 1 Hour 85

4.7 RI Distribution for Those Dying in 1 Hour 86

4.8 c_I vs. $\bar{b}$ Plot 92

4.9 c_I vs. $\bar{b}$ Plot 92

4.10 Inspection Rate vs Mortality Rate for combinations of α, β 93

4.11 Inspection Rate vs Mortality Rate for Combinations of α, β 94

4.12 Inspection Rate vs Mortality Rate for Combinations of α, β 94

4.13 ROC Curves 97

4.14 α, β Curves 97

4.15 Average Number of Epochs in State C per ICU visit 98

4.16 Average Number of Epochs in State P per ICU visit 99

List of Tables

2.1 Summary Statistics for Study Data Set 16

2.2 Proportion of Variance Explained by PCs 19

2.3 Raw Variables 30

2.4 Full Description of Model Variables. 35

2.5 Out of Sample Area Under ROC and Area Under Precision-Recall for 6-Hour Mortality Prediction(95% CI in Prentheses) 36

2.6 Out of Sample Area Under ROC and Area Under Precision-Recall for 12-Hour and 24-hour Mortality Prediction(95% CI in Parentheses) 36

2.7 P-value for Hosmer-Lemeshow Test of Each Split. 40

3.1 Summary Statistics for Final Data Set 48

3.2 Percentage sent to hospice binned by IMPAC 55

3.3 Maximum Achievable Recall with Guaranteed Precision 62

3.4 Percentage Sent to Hospice by Doctor and Augmented Method, Binned by Survival Time 63

4.1 Minimum Mortality Rate for Different Policies under Different Inspection Rate Constraint 100

Acknowledgements

This dissertation would not have been possible without support from my advisors, friends and families. I want to take a moment and show my sincere gratitude.

First, I would like to express my deepest appreciation to Yale Operations Department, especially my dissertation committee, for their generous time and support over the course of my PhD journey. I am truly grateful for my dissertation advisor, Professor Edieal J. Pinker for helping me set a clear research agenda and all the guidance along the way. He helped me identify my strengths and weaknesses and encourage me to grow skills that greatly benefits my career development. I am thankful for my dissertation committee members Prof. Edward H. Kaplan, Prof. Saed Alizamir and Prof. Lesley Meng for efforts in giving valuable feedback on this dissertation. I also want to thank my co-advisor during my first and second year of study, Prof. Donald Lee. I learned a ton from each one of you through great lectures, discussions, or working as your TA.

Second, I want to thank all my friends and colleagues at Yale University for support and collaborations both intellectually and spiritually over the past 5 years. My learning experience would not have been even half as joyful and meaningful without you. Big thank you to Cheng Hua, Ruiting (Dan) Dai, Zehao Liu, Rui (Ray) Zhang, Yu Shi, Yulian He, Canyao(Alex) Liu, Shawn Song and Yuang Jiang for being such amazing friends and giving me all the love and support, especially during the past year of pandemic.

Finally, I want to give my greatest appreciation to my family. I am indebted to my daughter Silvia and my wife Rose, for turning me into a more motivated and responsible person, and my parents and grandparents, for their unconditional love and support.

Chapter 1

Introduction

Wide implementation of electronic health record systems provides rich data for personalized medicine. One topic of great interest is to develop methods to assist physicians in prognosis on adverse outcomes, for example mortality. While many studies have reported on various new prediction models and algorithms there is relatively little literature on if and how these new prediction methods translate into actual benefits. My dissertation consists of three chapters that aims at filling this gap between prognostic predictions and clinical decisions in end-of-life care and intensive care settings.

In Chapter 2, we develop an approach to use temporal trends in physiologic data as an input into mortality prediction models. The approach uses penalized b-spline smoothing and functional PCA to summarize time series of patient data. We apply the methodology in two settings to demonstrate the value of using the ”shapes” of health data time series as a predictor of patient prognosis. The first application a mortality predictor for advanced cancer patients that can help oncologists decide which patients should stop aggressive treatments and switch to palliative care such as that provided in hospice. The second one is a real-time near term mortality predictor for MICU patients that can work as an early alarm system to guide timely interventions.

In Chapter 3, we investigate the integration of a prediction algorithm with physician decision making, focusing on the advanced cancer patient setting. We design a retrospective study to compare prognoses made by doctors and those that would be recommended by the IMPAC algorithm developed in Chapter 1. We used the doctor's discharge decision as a proxy of what they predict the patient as dying in 90 days and show that doctor's predictions tend to very conservative. Although IMPAC on its own does not perform better than doctors in terms of precision and recall, we find that IMPAC and doctors identify significantly different group of positive cases. IMPAC and doctors are also good at identifying very different groups of patients in terms of survival time. We propose a new way to augment decisions of doctors with IMPAC. At the same recall, the augment method identifies 43% more patients close to death than the doctors do. We also estimate potential hospitalizations and hospital length of stays avoided if the doctors use augmented procedure instead of acting on their own beliefs.

In chapter 4, we look at the integration of a prediction algorithm with physician decision making, focusing on the ICU setting. We use a POMDP framework to evaluate how decision support systems based on ICU mortality predictions can help physicians allocate time to inspect the patients at highest risk of death. We assume physicians have limited time and seek to optimally allocate it to patients to minimize their mortality rate. Physicians can do Bayesian updates on observations of patient health state. A prediction algorithm can augment this process by sending alerts to physicians. We represent the algorithm by a point on an ROC curve representing a particular alert threshold. We study two approaches to using the algorithm input: (1) Belief based policy (BBP) that integrates algorithm outputs using Bayesian updating; (2) Alarm triggered policy (ATP) where the physician responds only to the algorithm without updating, and compare them to benchmarks that do not rely on the algorithm at all. By running simulations, we explore how the accuracy of the algorithm can translate into lower mortality rates.

Chapter 2

Redesigning Patient Health Status Scores using Mortality Predictions Algorithms, with Applications in End-of-Life Care and Intensive Care

2.1 Introduction

There is a rich literature on the development of standardized scores to evaluate the health status of patients. These scores are objective metrics that provide standardized measurements of severity of disease or risk of severe adverse events.

Health status scores work primarily as a way to account for patient severity levels. The utility of developing these scores varies across different medical settings. In some clinical studies, these scores are used as proxy of severity of diseases to facilitate statistical analysis in controlled studies. For example, Sepsis Organ Failure Assessment (SOFA) has been used to assess the effects of new therapies for patients in the process of organ failure [1]. In other cases, health status scores can be used for financial and billing purposes. For example, Intensive care unit (ICU) physicians can justify a medicare beneficiaries' extended lengths of stays in ICUs with these scores as discharge readiness indicators [2].

Health status scores are also extremely important references for triaging and optimizing allocation of limited medical resources. In New York State, mortality risk assessment using SOFA is one of the three key steps in clinical ventilation protocol [3].

One category of health status score of great interest is severity scores for hospitalized and acute care patients. Existing scores include Sepsis Organ Failure Assessment (SOFA) [1], Acute Physiologic and Chronic Health Enquiry (APACHE II) [4], Mortality Probability Model (MPM II) [5], and Simplified Acute Physiologic Score (SAPS II) [6]. These prognostic tools use physiologic measures, demographics, and sometimes physician evaluations to generate scores calibrated by important patient outcomes like mortality and readmission. However, these prognostic tools are limited in using only static information taken at a fixed point in time (e.g., 24hr after ICU admission) to evaluate patients severity.

Potentially, there is valuable information conveyed by the changes in individual measurements over time. The ability to automate data extraction from electronic health record (EHR) systems provides opportunities to record patient data and analyze it as time series. This has enabled development of severity scores that can be updated in real time. For instance, the Rothman Index (RI, PeraHealth) [7, 8] is a continuous measure of patient condition based on a range of physiological measures, including labs, vitals, and nursing assessments. Dynamic scoring systems like RI can be integrated into EHR and work as a real-time visualization tool for physicians to monitor patient health conditions. These scores provide efficient and relatively accurate reference for physicians to monitor multiple patients in busy environments.

Instead of using traditional severity scores calibrated by mortality rate, recent studies have been using machine learning to generate mortality predictions as a measure of severity. Mortality predictions as severity scores have three major advantages. First, they can generate objective prediction scores based on data of large volumes and high dimensions. Second, compared to traditional severity scores, output in the form of probability can be more easily interpreted by both physicians and patients. Third, mortality predictions are easier to calibrate with existing evaluation measures for classification algorithms, like

Receiver-Operating- Characteristic (ROC) curves and Precision-Recall curves.

Recognizing the importance of temporal trends, a number of studies use summary statistics of the trajectories of the EHR measurements as inputs for mortality prediction [9]. Some do this by recalculating the scores at regular intervals using the most recent values, or a snapshot, of the measurements, whereas others recalculate the SOFA score using the worst values in the earlier 24 hours [10, 11, 12]. Sometimes, unstructured text data from patient clinical notes are also used as inputs, with features extracted from the text that have accumulated up to the current point in time [13, 14].

In chapter 2, we propose to use a machine learning approach called functional data analysis [15] to automatically extract temporal trend from time series data in the EHR to predict mortality.

Chapter 2 consists of 5 sections. In Section 2.2, we describe the procedure of functional data analysis and introduce two models for mortality prediction. These are methodologies that will be applied in the fields of end-of-life care and intensive care respectively in Section 2.3 and 2.4.

Section 2.3 is about making mortality predictions for patients with advanced cancer. We do a retrospective study of patients with advanced solid tumors identified by the Yale New Haven Hospital tumor registry. We build a new prognostic tool called IMPAC to predict 30-day, 60-day, 90-day or 180-day mortality risk of patients with advanced cancer. This tool can work as a decision support to help physicians identify patients who could benefit from palliative care.

In Section 2.4, we do a retrospective study of patients admitted to the medical ICU (MICU) of the Yale New Haven Hospital. We build a real-time mortality prediction tool for MICU patients and demonstrate that trajectory information extracted by functional data analysis gives better mortality predictions than using only snapshot or summary statistics of time series. Once implemented in EHR system, this can work as an early warning system to help nurses and physicians allocate limited medical resources.

In Section 2.5, we discuss limitations of these two applications in practice and introduce the next chapters that aim at addressing these limitations.

2.2 Methodology

2.2.1 Functional Data Analysis

We start with introducing methods to extract temporal trends in patient health conditions. One simple way to take advantage of richer time-varying data is to recalculate traditional scores like SOFA or APACHE every time there is new data available. But this will not overcome the intrinsic limitations of these scores. An alternative way is to create a new score that takes a smoothed average of past clinical measures like Rothman Index(RI; PeraHealth) [7, 16]. RI is an illness-severity index embedded within the electronic medical record. It continuously tracks 26 variables including vital signs, nursing assessments, lab results and cardiac rhythm. These data are fed into a proprietary algorithm, which gives a numerical score which measures how sick the patient is. Projects in section 2 and section 3 both utilized RI as a major input to predict mortality. Although the exact calculation of the RI is proprietary, the main idea is to calculate a 1-year mortality risk function for each of the 26 features selected by stepwise logistic regression. These univariate functions were then combined together in an additive manner to produce the RI score.

Like RI, there are other frequently updated physiologic measures recorded in EHR in the form of time series. Trajectories of these data can be seen as a function mapping from update time to measure values. Plotted on a panel, these time series can be used as visualization tool for physicians to keep track of the patient health conditions.

We are interested in using machine learning to extract temporal trend of RI and other frequently updated measurements. Ideally, the representation should be compatible input into machine learning algorithms that predicts mortality. In practice, these time series are not always measured at aligned time or frequency. Thus it's hard and inefficient to store all time series as tabular data. Linear or curve interpolation also heavily depend on the interpolation points and might lose significant amount of information about the temporal trend. Instead, we propose to use functional data analysis [17] to fit smooth curves to EHR time series.

Penalized B-Splines

Time series data stored in EHR can be represented as a discrete set of observations (o_j, t_j): At time t_j, patient's data is measured to be o_j.

Given a set of K basis functions $\{b_k(t)\}$, we can represent each time series in the form of linear combination.

$$C(t) = \sum_{k=1}^{K} c_k b_k(t) \tag{2.1}$$

We use basis spline, or B-spline as described in [18]. A B-spline function is a spline function that has minimal support with respect to a given degree, smoothness and domain partition. In order to define a set of B-spline functions, one needs to specify the order of spline basis n and a sequence interior knots $k_0, k_1, ..., k_{q+1}$, with $k_0 \leq k_1 ... k_{q+1}$. When these knots are equidistant, the splines are called uniform. Then we define an augmented knot set $k_{-(n-1)} = ... = k_0 \leq k_1 ... \leq k_{q+1} = k_{q+2} ... = k_{q+n}$ by appending the lower bound knot k_0 and upper bound knot k_{q+1} for $n-1$ times respectively. We can re-index these $q+2n$ knots by $i = 0, ..., q+2n-1$ and call them $t_0, ..., t_{q+2n-1}$.For a given sequence of knots, there is a unique set of spline functions $B_{i,n}(x), i \in \{0, 1, ..., q+2n-1\}$ satisfying:

$$B_{i,n}(x) = \begin{cases} \text{nonzero} & x \in [t_i, t_{i+n}) \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

$$\sum_i B_{i,n}(x) = 1, \forall x \in [t_0, t_{q+2n-1}] \tag{2.3}$$

Cox-de Boor recursion can be used to derive B-Spline's functional form of any order. To start with, the B-splines of order n=1(degree 0) and augmented knots $\{t_0, t_1 ... t_{q+1}\}$ can be written as the following step functions.

$$B_{i,1}(x) = \begin{cases} 1 & x \in [t_i, t_{i+1}) \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

B-splines of higher order $n = j$ are defined by the recursion below:

$$B_{i,j}(x) = \frac{x - t_i}{t_{i+j} - t_i} B_{i,j-1}(x) + \frac{t_{i+j+1} - x}{t_{i+j+1} - t_{i+1}} B_{i+1,j-1}(x) \tag{2.5}$$

Formally, any time series $\{(o_j, t_j)\}$ can be approximated as a linear combination of a set of B-splines of order d:

$$C(t) = \sum_{i=0}^{q+2n-1} c_i B_{i,d}(t) \tag{2.6}$$

While using relatively large number of knots, it's necessary to introduce some penalty on the fitted curve to avoid overfitting. We used a penalty on the second derivative of the fitted curve introduced by [19].

The optimal set of coefficients $\{c_i^*\}$ should minimize a penalized least squares objective function, which can be written as

$$\sum_j [o_j - \sum_{i=0}^{q+2n-1} c_i B_{i,d}(t_j)]^2 + \lambda \int [\sum_{i=0}^{q+2n-1} c_i B^{''}_{i,d}(t)]^2 dt \tag{2.7}$$

We are penalizing on the curvature of function $C(t)(\int [C^{''}(t)]^2 dt)$ that measures the roughness of the curve. $C^{''}(t)$ is easy to calculate because $C(t)$ is a linear combination of polynomial functions. $\lambda \geq 0$ represents the weight of penalty. The larger λ is, the more $C(t)$ will be like a straight line and the smaller λ is, the more $C(t)$ will fit closely to the observed data. In practice, λ can be tuned by visualization of data.

Functional Principal Component Analysis

After fitting curves using a set of K B-spline basis, we can represent each clinical time series with K weights. When K is large and we have multiple variables stored as time series for each individual, these functional coefficients can be high dimensional. Some mortality prediction algorithms based on linear models require us to further extract information from these variables before using them as model input. In this case, we propose to use functional principal component analysis (FPCA) [15] for dimension reduction.

FPCA is a statistical method for investigating the dominant modes of variation of functional data. FPCA has proven success in extracting temporal trends in financial [20] and environmental [21] studies. The goal of functional PCA is to identify important modes of variance among these curves. Keeping only dominant modes will improve the signal-to-noise ratio and allow machine learning models to process the information more efficiently. FPCA can be applied for displaying modes of functional variation, for modeling sparse longitudinal data and conducting functional regression and classification.

Suppose we have a frequently measured time series data x of N patients indexed by i. Assume each patient's time series is generated by a realization $X_i(t)$ of a square-integrable stochastic process $X(t)$, $t \in T$. Let $\mu(.)$ and $G(.,.)$ represent mean and covariance function of $X(t)$.

$$\mu(t) = E(X(t)) \tag{2.8}$$

$$G(s,t) = Cov(X(s), X(t)) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t) \tag{2.9}$$

$\lambda_1 \geq \lambda_2 ... \geq 0$ are the eigenvalues and $\phi_1, \phi_2, ...$ are the orthogonal eigen functions of the linear Hilbert-Schmidt Operator $H(f) = \int_T Cov(X(s), X(t)) f(s) ds$. By the Karhunen-Loève decomposition,

$$X(t) - \mu(t) = \sum_{k=1}^{\infty} \eta_k \phi_k(t) \tag{2.10}$$

where $\eta_k = \int_T (X(t) - \mu(t))\phi_k(t)dt$ is the principal component corresponding to the k-th eigenfunciton ϕ_k that satisfies $E(\eta_k) = 0, Var(\eta_k) = \lambda_k, E(\eta_k \eta_l) = 0, \forall k \neq l.$

Suppose we have $C_1(t), ..., C_n(t)$ as a set of n realization of L^2 functional C(t), the general procedure for finding principal components is as follows:

- Step 1: Find principal component weight function $\phi_1(s)$ for which the principal components scores

$$f_{i1} = \int \phi_1(s) C_i(s) ds \tag{2.11}$$

and maximize $\sum_i f_{i1}^2$ subject to constraint

$$\int \phi_1^2(s)ds = 1 \tag{2.12}$$

- Step $k(k \geq 2)$: Find principal component weight function $\phi_K(s)$ for which the principal components scores

$$f_{ik} = \int \phi_k(s)C_i(s)ds \tag{2.13}$$

and maximize $\sum_i f_{i1}^k$ subject to constraint

$$\int \phi_1^2(s)ds = 1 \tag{2.14}$$

$$\int \phi_k(s)\phi_m(s)ds = 0, \text{for all } m < k \tag{2.15}$$

The estimation of $\phi_k(t)$ and $\eta_k(t)$ in our study setting is achieved by a restricted maximum likelihood estimations through a Newton-Raphson procedure described in Peng and Paul (2009) [22] to estimate functional PCs from sparsely and irregularly observed longitudinal data.

The number of principal components can not exceed the number of B-spline basis used to fit the original curve. In practice, the proportion of variance explained by each principal component often decays fast and can guide how many principal components to keep. Only functional PCs that explains the highest proportion of variance in raw data should be included. Suppose we only include K_s functional PCs, now each time series can be approximated as linear combinations of selected PCs.

$$C(x) = \sum_{k=1}^{K_s} w_k\phi_k(x) \tag{2.16}$$

2.2.2 Mortality Prediction Models

Next, we introduce one survival model, Cox Proportional Hazard model [23], and one machine learning model, Random Forest [24] that will be used to predict patient mortality in

Section 2.3 and Section 2.4.

Cox proportional-hazards Model

We use Cox Proportional-hazards model [23] to investigate the relationship between the survival time and the encounter level variables. The key idea of this model is to relate the time that passes before certain events (like mortality in our setting) occur to covariates that may be associated with that quantity of time. A hazard function is defined as the instantaneous rate of an event occurrence and is expressed as:

$$h(t) = lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{2.17}$$

The Cox proportional-hazards model has three major inputs:

1. Event: Death, disease occurrence, disease recurrence, recovery, or other experience of interest

2. Time: The time from the beginning of an observation period (such as surgery or beginning treatment) to (i) an event, or (ii) end of the study, or (iii) loss of contact or withdrawal from the study.

3. Censoring: If a subject does not have an event during the observation time, they are described as censored. The subject is censored in the sense that nothing is observed or known about that subject after the time of censoring. A censored subject may or may not have an event after the end of observation time.

The Cox proportional-hazards model is a semi-parametric model of hazard functions. For each individual indexed i, the model assumes hazard functions of the following form:

$$h_i(t) = h_0(t) e^{\sum_k x_{i,k} \beta_k} \tag{2.18}$$

Its hazard function can be viewed as consisting of two parts:

- Underlying baseline hazard function, $h_0(t)$, that describes how the risk of event per unit time changes over time. There is no assumption about its shape. So $h_0(t)$ is estimated by taking average over the whole population.

- The log ratio $log(\frac{h(t)}{h_0(t)})$ is a stationary linear function of individual covariates.

For each individual sample indexed i, let random variable T_i denote the individual's survival time. $F_i(t)$ and $f_i(t)$ represent the c.d.f. and p.d.f. of T_i.

$$H_i(t) = \int_0^t h_i(s)ds = \int_0^t h_0(s)e^{\sum_k x_{i,k}\beta_k}ds = \int_0^t \frac{f_i(s)}{1-F_i(s)}ds = -ln(1-F_i(t)) \tag{2.19}$$

$$F_i(t) = 1 - e^{-\int_0^t h_0(s)e^{\sum_k x_{i,k}\beta_k}ds} \tag{2.20}$$

We use the Newton-Raphson algorithm to maximize the partial likelihood function given by [25]. This will give us a set of maximum likelihood estimator for $\hat{\beta}_k$. All survival analysis will be done in R language using a package called *Survival*[26]

Random Classification Forest

We use the random classification forest algorithm [27] to predict a binary outcome(mortality in T days) m for MICU patients based on a set of input variables $x_1, ...x_k$. The data set contains N independent samples, where each sample, indexed i is defined as, $S_i = (\vec{x_i}, m_i) = (x_{i,1}, x_{i,2}, ...x_{i,k}, m_i)$.

The random forest classifier consists of a combination of tree classifiers where each classifier is generated using a random vector sampled independently from the input vector and each tree casts a unit vote for the most popular class to classify an input vector. We will first describe the random forest algorithm in detail.

Repeat following procedure n_{tree} times:

1. From the original data set, draw a bootstrap sample $\{S_{i_1}, ..., S_{i_N}\}$.

2. Use this sample to grow an unpruned classification tree of depth up to d to predict m, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} out of the k features in $\vec{x}$ and choose the best split from those variables.

Each iteration j returns a classification tree $T_j(.)$, which is a mapping from feature space X to outcome space $\{0, 1\}$. Given any new data $\vec{x}_{new}$, prediction can be made by aggregating the predictions of the n_{tree} trees. The prediction score can be represented as the percentage of yes votes(positive predictions): $r = \frac{\sum_{j=1}^{n_{tree}} T_j(\vec{x}_{new})}{n_{tree}}$. By setting a cutoff c, we will predict all samples with $r \geq c$ to be positive and those with $r < c$ to be negative.

Random forests have two advantages compared to decision trees and logistic regressions. First, it can handle large and high dimensional data sets more efficiently. Second, by growing a large number of independent decision trees, it avoids overfitting by decision trees and improves accuracy. We used the *randomForest* R-package [28] to build random forests.

2.3 End-of-life project: Development of Imminent Mortality Predictor in Advanced Cancer(IMPAC)

2.3.1 Background and Overview

End-of-life care for patients with advanced cancer is aggressive, costly and often discordant with patients' wishes [2, 3, 29, 30]. Among Medicare decedents, 80% were hospitalized within 90 days of death, 27% were admitted to the intensive care unit (ICU) within 30 days, and 20% transitioned to hospice in the last 3 days of life. Thirty percent of all spending for cancer occurs in the last year of life [31, 32, 33].

Patients with advanced cancer rely on their oncologists to guide end-of-life care decisions. However, oncologists' estimation of life expectancy is often inaccurate [34, 35] and overly optimistic [36, 37]. Widely used prognostication tools are limited by dependency on subjective assessment. Instruments [38, 39, 40] have been developed to stratify the risk of dying in the near term, but most, like the Palliative Prognostic Index[40], use subjective physician assessments, static data, and statistical techniques that do not make full use of data available in electronic health record (EHR) systems.

In this section,we explain the development of a new prognostic tool, the Imminent Mortality Predictor in Advanced Cancer (IMPAC). The goal of this tool is to generate mortality prediction for patients with advanced cancer every time they get admitted to hospital. This tool requires patients to stay longer than 48 hours in order to collect enough data to evaluate each patient's health conditions. The original paper developing this tool is published on a peer reviewed medical journal.

This section is organized as follow. First, we describe the study setting and data set. Second, we describe the methodologies we used to build IMPAC, which includes penalized B-spline fitting, functional principal component analysis, and Cox proportional hazard model. Third, we will report model evaluation metrics (ROC, Precision Recall, Survival Time) and estimate the potentially avoidable cost of treatment using IMPAC.

2.3.2 Data Description

Study Population

We examined the inpatient records of 2774 unique patients with advanced solid tumors identified by the Yale New Haven Hospital tumor registry with a hospital discharge (for any cause of admission) between October 1, 2013 and June 31, 2017. These patients, in total, generated 4778 visits, during this time period.

Variable definitions

For the purposes of the analyses and discussion in this chapter we use 90-day mortality from the date of discharge as the dependent variable. Patients' survival status was collected from the institutional tumor registry as of June 31, 2019. For patients with more than one hospitalization during the study period, each visit (encounter) was treated as a separate observation. For each unique patient we have demographic data and a distinct data record for each encounter. For each encounter we have static variables and a time series of health status scores called the Rothman Index (described in detail below). The independent variables are defined as follows:

- **ED Admission** If the patient got admitted to hospital through emergency department, this field is set to 1. Otherwise, this field is set to 0.

- **Previous Admission in Last 90 days** If the patient had at least one hospitalization within the past 90 days from current hospital admission, this field is set to 1.

Otherwise, this field is set to 0.

- **Unique patient identifier**: Encrypted medical record number.
- **Gender**
- **Age** Age of the patient (integer) at the point of current hospital admission.
- **Race** Race of the admitted patient.
- **Disease Team** The specialized cancer teams that primarily treats the patient. This is a proxy for the type of cancer the patient has.
- **Discharge Disposition** This field indicated where patient is discharged to. Possible values are: Home, Hospice, Nursing Facility, Self care, Rehabilitation Center, and Other Hospital Facilities.

Summary statistics for these variables are reported in Table 2.1.

Rothman Index

We used a metric available in the Electronic Health Record at Yale New Haven Hospital, the Rothman Index (RI; PeraHealth)[7]. RI is a real-time, EHR-based scalar measure of patient acuity that is continually calculated throughout the hospitalization and has been incorporated into most commercially available EHRs. It incorporates 26 clinical data elements, including vital signs, nursing assessments, and laboratory results, and has been shown in multiple settings to predict both mortality and readmission. Lower RI scores reflect a worse health status. Table 1 shows empirical probability density function of Rothman Index.

To ensure that enough RI scores are spread across a sufficient period to inform the model, we exclude visits with less than 48 hours of RI monitoring (excluding 71 patients) and for which no RI was available between 36 and 48 hours (33 excluded). The final data set consisted of 2,774 unique patients with 4,575 inpatient encounters.

Functional Data Processing

We use the B-spline basis with $n = 4$ (degree 3, cubic function), $r = 11$(evenly divide [0,1] into 10 pieces) spanned over range from 0 to 2(representing days here). This leads

Characteristics	Total Number	Mean	Standard Deviation	Percentage
Patient Specific	2774	-	-	-
Age, years	-	63	12	-
Gender, Males	-	-	-	46.0%
Visit Specific	4575	-	-	-
Entry Through ED	-	-	-	41.0%
Prior visit in last 90 days	-	-	-	35.0%
Length of Stays, days	-	6.74	6.84	-
Type of Cancer	-	-	-	-
Breast	-	-	-	7.3%
Endocrine	-	-	-	2.4%
GI	-	-	-	24.0%
Genitourinary	-	-	-	5.8%
Gynecologic	-	-	-	15.0%
Head and Neck	-	-	-	10.0%
Melanoma	-	-	-	3.5%
Neurologic	-	-	-	7.9%
Sarcoma	-	-	-	4.9%
Thoratic	-	-	-	18.0%
Undefined and Unknown	-	-	-	0.74%

Table 2.1: Summary Statistics for Study Data Set

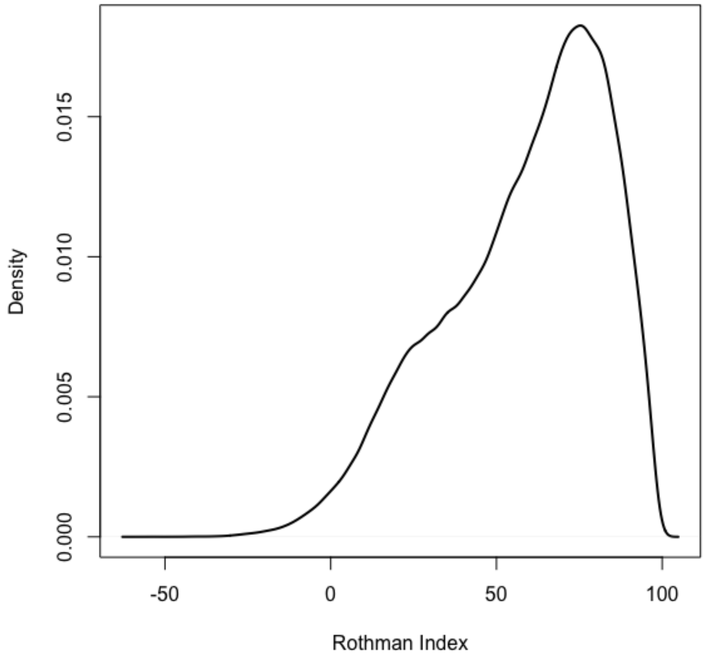


Figure 2.1: Empirical pdf of RI

to 13 functional basis, named as $b_1(t), b_2(t), ..b_{13}(t)$. We plot the B-spline basis functions that we use to fit the RI curves in figure 2.2. After that, we do functional principal component analysis on the fitted curve, and get a set of 13 Functional Principal Components $PC_1(t), ...PC_{13}(t)$. Given these functional PCs we can represent each patient encounter RI trajectory by a set of weights, then

$$RI_i(t) = w_{i,1}PC_1(t) + w_{i,2}PC_2(t) + ...w_{i,13}PC_{13}(t), \tag{2.21}$$

where $PC_1(t), ..., PC_{13}(t)$ are the principal component functions and $w_{i,1}, ..., w_{i,13}$ are the associated weights for the RI curve of encounter i.

Based on the proportion of total variance(Table 2.2) explained by each functional principal components, we only include weights of the top three functional PCs in our prediction model and plot them in Figure 2.3.

We use the Cox proportional hazard model to predict T-day mortality($T = 30, 60, 90, 180$) starting from 48 hours after admission for each encounter. We use days as unit of time for

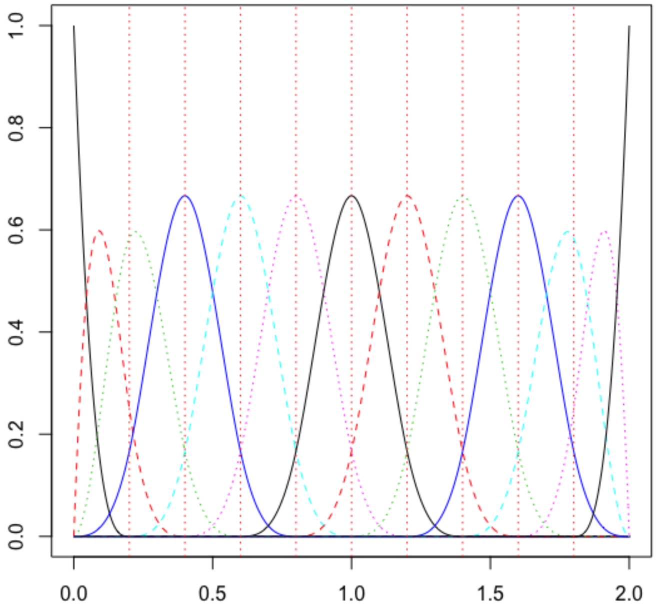


Figure 2.2: B-spline Basis to Fit RI Curves

our analysis. For any sample(hospitalization) indexed i, after 48 hours(2 days) after admission into hospital, we can observe $(t_i, \delta_i, \vec{x_i})$. t_i is the survival time between time of observation(2 days after hospital admission,A_i) and date of last contact(the day each patient's vital status was last updated).

$$t_i = D_i - (A_i + 2) \tag{2.22}$$

t_i represent actual survival time if data is not censored and a lower bound of survival time if data is censored. δ_i is the censoring indicator. $\delta_i = 1$ if the patient was dead at last observation. $\delta_i = 0$ if the patient was still alive at last observation. $\vec{x_i}$ is the patient's covariate vector which includes age, gender, weights of the first three functional PCs, entry through ED indicator, visit in prior 90 days indicator, and disease team.

Principle Component	Proportion of Variance Explained
PC1	90.1%
PC2	6.7%
PC3	1.8%
PC4	0.69%
PC5	0.41%
PC6	0.15%
PC7	< 0.1%
PC8	< 0.1%
PC9	< 0.1%
PC10	< 0.1%
PC11	< 0.1%
PC12	< 0.1%
PC13	< 0.1%

Table 2.2: Proportion of Variance Explained by PCs

2.3.3 Prediction model

We used the weights of $PC1$, $PC2$, $PC3$ (w_1, w_2, w_3) as well as 22 static variables to fit hazard function. After variable selection with Akaike Information Criteria(AIC) [41] the hazard function $h(t)$ can be written as the following formula:

$$\begin{aligned} log(h(t)) =& log(h_0(t)) + \beta_1 age + \beta_2 Gender + \beta_3 w_1 + \beta_4 w_2 + \beta_5 w_3 \\ & + \beta_6 EntryThroughED + \beta_7 VisitIn90Days + \beta_8 Breast \\ & + \beta_9 Endocrine + \beta_{10} GI + \beta_{11} Genitourinary + \beta_{12} Gynecologic \\ & + \beta_{13} HeadAndNeck + \beta_{14} Melanoma + \beta_{15} Neurologic \\ & + \beta_{16} Sarcoma + \beta_{17} Thoratic \end{aligned} \tag{2.23}$$

Figure 4 shows the coefficients and corresponding p values of each independent variables. For weights of principal components, if the regression coefficient for a particular weight is

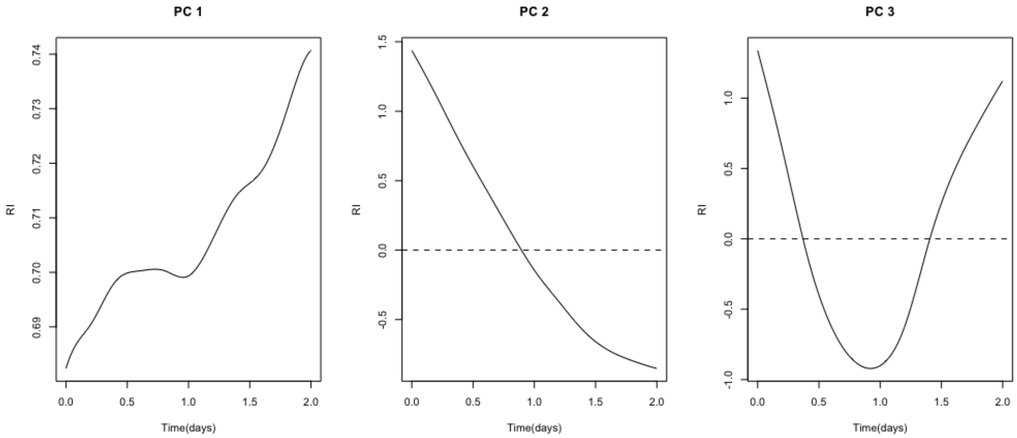


Figure 2.3: First 3 Principal Components of RI Curves

positive and significant, then trajectories that resemble the shape of the corresponding curve are associated with higher mortality. For static variables, a positive weight represents that the increase of variable value would lead to higher mortality.

- PC1 represents the average RI level and a upward trend over time. Coefficient of its weight is negative and significant. This means trajectories with more resemblance to PC1 indicate higher likelihood of death.

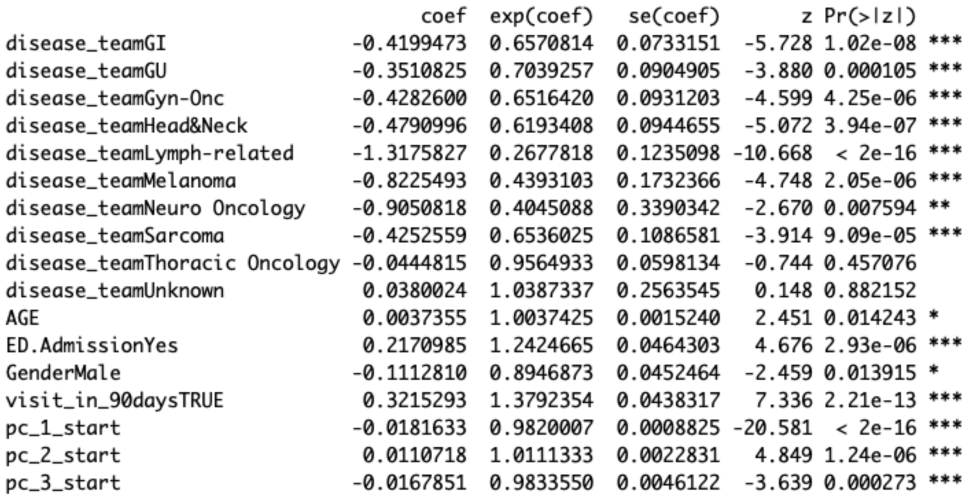
- PC2 represents a downward trend over time. Coefficient of its weight is negative and significant. This means trajectories with more resemblance to PC1 indicate higher likelihood of death.

- PC3 represents a downward followed by an upward trend, or recovery trend. Coefficient of its weight is negative and significant. This means trajectories with more resemblance to PC3 indicates less likelihood of death.

Based on estimated hazard rate, we can calculate the probability of surviving less than (or dying in) 30 days, 60 days, 90 days, and 180 days. These risk scores are the final output of our IMPAC tool.

2.3.4 Model Evaluation

We use Monte-Carlo Cross Validation [42] to create random training testing splits. In order to avoid cases where information from a patient's future encounters is used to train model



```
                                    coef exp(coef)  se(coef)       z Pr(>|z|)    
disease_teamGI                -0.4199473 0.6570814 0.0733151  -5.728 1.02e-08 ***
disease_teamGU                -0.3510825 0.7039257 0.0904905  -3.880 0.000105 ***
disease_teamGyn-Onc           -0.4282600 0.6516420 0.0931203  -4.599 4.25e-06 ***
disease_teamHead&Neck         -0.4790996 0.6193408 0.0944655  -5.072 3.94e-07 ***
disease_teamLymph-related     -1.3175827 0.2677818 0.1235098 -10.668  < 2e-16 ***
disease_teamMelanoma          -0.8225493 0.4393103 0.1732366  -4.748 2.05e-06 ***
disease_teamNeuro Oncology    -0.9050818 0.4045088 0.3390342  -2.670 0.007594 ** 
disease_teamSarcoma           -0.4252559 0.6536025 0.1086581  -3.914 9.09e-05 ***
disease_teamThoracic Oncology -0.0444815 0.9564933 0.0598134  -0.744 0.457076    
disease_teamUnknown            0.0380024 1.0387337 0.2563545   0.148 0.882152    
AGE                            0.0037355 1.0037425 0.0015240   2.451 0.014243 *  
ED.AdmissionYes                0.2170985 1.2424665 0.0464303   4.676 2.93e-06 ***
GenderMale                    -0.1112810 0.8946873 0.0452464  -2.459 0.013915 *  
visit_in_90daysTRUE            0.3215293 1.3792354 0.0438317   7.336 2.21e-13 ***
pc_1_start                    -0.0181633 0.9820007 0.0008825 -20.581  < 2e-16 ***
pc_2_start                     0.0110718 1.0111333 0.0022831   4.849 1.24e-06 ***
pc_3_start                    -0.0167851 0.9833550 0.0046122  -3.639 0.000273 ***
```

Figure 2.4: Variable Coefficients in Cox Proportional Hazard Model

to predict mortality for a previous visit, we randomly split the data set into training and testing set by patients so that approximately 70% of visits are in the training set and 30% of visits are in the testing set.

We evaluate IMPAC accuracy in three ways:

- Area Under ROC curve
- Area Under Precision-Recall curve
- Survival time distribution.

Precision-Recall Curve

The precision-recall curve plots and summarizes the trade-off between the true positive rate (sensitivity) and the positive predictive value for a classification model under different probability thresholds.

Although receiver operator characteristic curves are commonly used to present result for binary decision problems in machine learning, there are two situations where the precision-recall curve (PRC) works better as a performance metric:

- When dealing with highly imbalanced data sets, PR curves give a more informative picture of the algorithms performance [43, 44].

- When the cost of a false positive is much higher than the cost of a false negative classification.

We used PRC to evaluate IMPAC since it satisfies both conditions.

Suppose IMPAC generates mortality risk scores r that ranges from 0 to 1. One needs to set a threshold τ and classify patients with risk scores higher than the threshold τ to die soon. As this threshold increases, less patients get positive predictions and advised to shift to palliative care. But the overall precision of these positive prediction would also change. We use precision recall curves to measure the success of our mortality prediction algorithms. Suppose our definition of a patient being at end-of-life is being within T days of death. We use IMPAC to generate an estimate of the probability of T day mortality. For a given τ precision P,(or Positive Predictive Value) is defined as the proportion of the patients classified as at end of life, that actually died within T days.

$$P = \frac{TP}{TP + FP} = P(\text{Die in T days}|r \geq \tau)$$

Recall R, (or sensitivity) is defined as the proportion of patients who died within T days that were predicted to be dying in T days by IMPAC.

$$R = \frac{TP}{TP + FN} = P(r \geq \tau|\text{Die in T days})$$

Given a time horizon T, and recall value there is a corresponding probability threshold τ. We take T = 30, 60, 90 and 180 days. For each T, we do the following: For 20 iterations, we randomly split the raw data set by patient into training and testing set on a 7:3 ratio. And use linear interpolation to calculate precision levels corresponding to evenly spaced recalls values $(0, 0.01, 0.02, ..., 1)$. We calculate the average precision level over 20 data splits at each recall level. The average precision recall curve is shown in Figure 2.5.

For example, using $\tau = 50\%$ as a classification threshold for 90-day mortality, we achieve a corresponding recall of $R = 40\%$. At this recall IMPAC has a 60% average PPV. This means that of all the patients for which we estimate a likelihood of 90-day mortality greater than 50%, 40% will live longer than 90 days. I.e. we misclassified 40% of the patients and

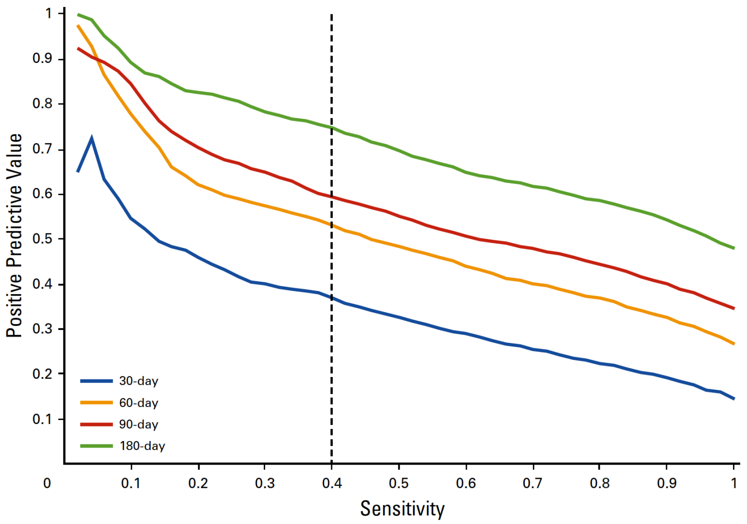


Figure 2.5: Average Precision Recall Curve for 30,60,90,180-day Mortality Predictions.

correctly classified 60%. What is the implication of the misclassification? From a medical perspective the concern is that a patient misclassified in this way will be directed toward palliative care prematurely, meaning that there was forgone opportunity for life extension from aggressive care. We therefore want to understand the classification accuracy and patient survival time distributions.

Survival time distribution

In this section we report on an analysis of the survival distribution for 90-day mortality predictions using a 50% probability of death as a classification threshold. In Figure 2.6 we display box plots of survival times for patients by classification accuracy. I.e. were the classifications True Positive, False Negative, True Negative, or False Positive. We can see that the True Positive and False Negative patients have a very narrow range of short survival times as they are bounded by 90 days. However, it is instructive to look at the patients who lived longer than 90 days. The median patient incorrectly classified as likely to die within 90 days (False Positive) had a median of 229 days. On the other hand the True-Negative patients lived a median of 526 days. This large difference indicates that even when "incorrect" that algorithm is identifying patients who will tend to die substantially sooner, if not exactly within 90 days.

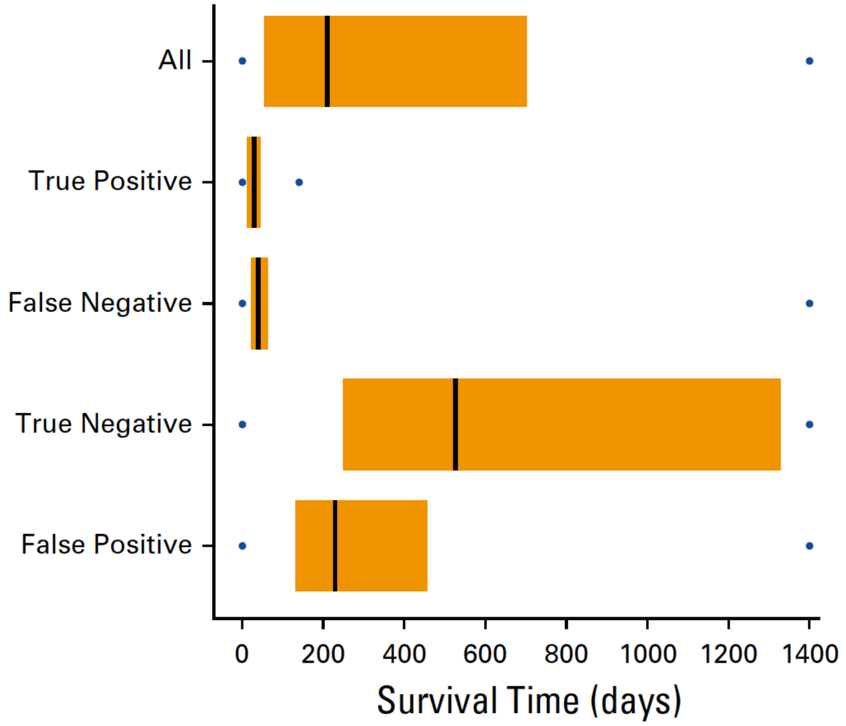


Figure 2.6: Survival Time Distribution for Different Groups of Patients.

2.3.5 Determination of Potential Avoidable Care

Reducing the amount of aggressive interventions a patient receives near the end-of-life has the potential to reduce their suffering and can allow them to spend their remaining time in an environment that is better suited to their needs and those of their families. This is one of the main purposes of palliative care. Reducing the intensity of medical interventions at end of life also reduce medical resource usage and overall waiting time. For example, it can reduce the load on critical care units in hospitals. It can also reduce medical costs brought by aggressive treatments, which are mostly expensive. In this section we attempt to calculate the potentially avoidable costs to the health system of the prediction algorithm we developed here.

To estimate the potentially avoidable cost of treatment, we selected one test set with the probability of death within 90 days threshold set at 50% and did a closer analysis of those patients. For this set of 309 inpatient encounters, 103 resulted in death within 90 days, and IMPAC correctly identified 41 of those encounters (38 unique patients). For the 41 encounters, the first hospitalization for each patient was deemed the index hospitalization.

We calculated the potentially avoidable cost using Yale New Haven Health System's cost accounting system. We classified all procedures and interventions between IMPAC's prediction and the time of actual death as potentially avoidable. We extracted direct cost data for products and services incurred at the hospital after the 48-hour patient assessment period of the index admission and during all subsequent inpatient and outpatient visits. Direct costs are patient care related (e.g. radiology, nursing, laboratory). Hospital indirect costs, physician services, and costs incurred outside of our health system were excluded.

In this sample, the average total direct cost incurred per patient during the index hospitalization after the initial 48-hour assessment period was $7,495. The average direct costs of subsequent hospitalizations and outpatient visits totaled $9,194 and $1,172, respectively, per patient. We assume the total of these costs ($17,861; 95% CI, $9,162 to $21,665) to be potentially avoidable.

We then compare these potentially avoidable costs with costs associated with hospice care for the same duration, based on the assumption that IMPAC would divert patients who are predicted to die soon into hospice care and thus avoid unnecessary interventions. We assume that these patients receive hospice care for the remainder of their lives and that survival time under hospice care is the same as in an acute care facility. Hospice costs were calculated on the basis of 2014 national hospice data showing the daily payment for four levels of care: routine home care, 93.8% at $156; general inpatient care, 4.8% at $694; continuous care, 1.0% at $91; and respite care, 0.4% at $161 daily rate. These were used to approximate an average daily hospice cost which, when multiplied by the total inpatient length of stay after IMPAC's prediction of death, approximated the cost of hospice care used for comparison.

Had these 38 patients been treated under hospice care after the initial 48- hour assessment, it would have prevented 491 days of inpatient acute care. The estimated corresponding cost of hospice care would be $2,448 per patient. This implies a savings in this sample of $15,413 (95% CI, $9,162 to $21,665), that is, the potential avoidable cost ($17,861) minus the cost of hospice ($2,448).

One patient in our cohort is illustrative. On her first admission, the IMPAC would have predicted death within 90 days. This patient was discharged home with services and was readmitted twice. She subsequently underwent surgery, made multiple trips to the ICU, and eventually died in the hospital. Cost of care for this one patient after IMPAC's prediction was $91,748 compared with $7,768 for equivalent days under hospice care after the index admission.

There are several limitations to the assumptions used in this cost-avoidance estimate. First, we cannot ensure that all patients flagged by IMPAC would indeed switch to hospice care, or that it would occur exactly at 48 hours into the admission, or that hospice care would last for the remainder of the patient's life. Second, we needed to set a time point after the 48-hour IMPAC score and assume that the inpatient care and procedures would have been prevented if the patient had transferred to hospice. It is likely that some of the procedures received after the 48-hour IMPAC reading may have been performed with the goal of palliation. However, the use of costly palliative procedures is limited under the hospice benefit and is thus unlikely to significantly alter the cost avoidance estimates. Although this cost avoidance analysis has limitations, it does provide an estimate of the financial implications of a more rational approach to end-of-life care.

2.3.6 Conclusion

In summary, we have developed a novel prognostic tool, IMPAC, that uses objective data to generate life expectancy probabilities automatically from EHR data in real time. Our novelty lies in using functional data analysis tools to extract temporal trends of cancer patient's health conditions to better predict imminent mortality.

Our data indicates heavy usage of aggressive treatments or delayed hospice care near end of life. There are a few possible explanations for the observed pattern of end-of-life care at our center. Patients may seek more aggressive care at quaternary facilities because they are not yet ready to relinquish the hope of improved survival, which they associate, often inaccurately, with further disease-modifying care. On the providers' side, expertise and resources available at major centers may lead to greater use of treatments that are not standard elsewhere even when physicians may recognize poor prognosis. The availability of an objective prognostic tool, such as IMPAC embedded within the EHR system, could assist oncologists and patients in designing a more realistic plan of care.

If integrated into the standard clinical workflow, the IMPAC will signal oncologists that goals-of-care conversations are imperative, helping to facilitate prognostic understanding and informed decisions regarding downstream health care interventions. Our financial analysis quantified the potential reduction in avoidable care that better mortality predictions could achieve.

The EOL project has the following limitations. First, IMPAC is trained only on data collected within a single hospital system. A multi-center study should be conducted to further validate the accuracy of IMPAC. Second, IMPAC uses RI, which is a proprietary commercial product, thus limiting applicability at hospitals that don't use RI. However, the RI is just one example of a high-frequency EHR-based patient health status index. When similar indices become more common in the future, we believe our approach couple be adapted to them as well. Third, we made the assumption in avoidable cost analysis that those captured by IMPAC could be immediately discharged to hospice. In reality, patients can't be immediately discharged until they meet certain kinds of discharge criteria. Patients might also insist on continuing to receive aggressive treatments, making our cost saving an over-estimation.

2.4 Using the Shapes of Clinical Data Trajectories to Predict Mortality in Intensive Care Units

2.4.1 Overview and Background

Because patients in an ICU are so closely monitored there is extremely rich time varying data associated with a patient stay in an ICU. The ability to automatically extract this data from EHR systems enables us to build dynamic mortality warning indicators for ICU patients. Existing illness severity scores like APACHE, MPM, SOFA and SAPS [4, 5, 6, 45] are used in ICUs for assessing patient acuity level, assigning individual nursing level, and prioritizing physician inspections. Although proven to be correlated with outcome metrics like 24-hour mortality, ICU mortality, hospital mortality and readmission, these prognostic scores only use static information taken at a fixed point in time (e.g., 24 hr after ICU admission) to identify patients at high risk of dying during an ICU visit. They ignore potentially valuable information conveyed by the changes in the measurements over time. Recognizing this limitation, a number of studies use summary statistics of the trajectories of the EHR measurements as inputs for mortality prediction.

This project has two aims. The first is to show how functional data analysis [17] can be used to incorporate the entire trajectory of a frequently measured variable into dynamic predictions of very near-term mortality (updated regularly during ICU stay). Second, we demonstrate that there is predictive value in incorporating this extra trajectory information: For periodically updated predictions of mortality over a 6-hour window in the ICU, our approach provides statistically significant improvements to both the area under the receiver operating characteristic (AUROC) curve and area under the precision-recall curve (AUPRC) over using trajectory features manually selected from the “same” dataset. In other words, this increase in accuracy comes at no extra cost (aside from negligible computational ones), and the same result is seen for 12- and 24-hour prediction windows as well. The functional data analysis method automatically extracts features from the trajectory. These features capture information about all aspects of the curve and can be fed as inputs into any predictive mortality model. Thus, the method can potentially be used to improve any risk score that uses some form of trajectory information.

The medical literature has many studies looking at mortality prediction using a variety of data sources and methods. The end of life prediction model we described in the previous section is an example. There are however very few studies of very short term models prediction models, where by short term we mean time scales of hours. There are many inherent challenges in making mortality predictions on such short time horizons. First and foremost, it requires predicting a very rare event. Even in a relatively high risk population such as ICU patients the in unit mortality rate is relatively low. In our data set it is approximately 11 percent. But the mortality rate in any 6 hour interval will of course be much lower. In our case only approximately 1.2 percent. Second, over short time periods patient clinical measures can fluctuate, even just from measurement error. On the other hand a near term prediction can alert physicians and nurses to patients that can be saved by timely intervention.

2.4.2 Data Description

Study Population

Our study was a retrospective review of de-identified patient records approved by the Yale University Institutional Review Board using data on 4,557 unique patients admitted to the medical ICU (MICU) of the Yale-New Haven Hospital between January 2013 and January 2015. A total of 5,505 hospitalization episodes and 6,113 MICU visits were recorded.

The mean age of patients was 63 years (sd, 17 yr), and 53% were male patients with mean weight of 179 lb (sd, 57 lb) and mean height of 5.5 ft. (sd, 0.4 ft). About 6.8% of patients had multiple ICU admissions, and 53% of ICU admissions occurred within 24 hours of being admitted to the hospital. After removing ICU stays shorter than 24 hours, the average length of ICU stay in our final dataset was 3.9 days. Among all observational units, 1.2% were followed by death within 6 hours (1.9% for 12 hr, 3.4% for 24 hr, and 17% were followed by death during the same ICU stay).

Data Fields

In addition to patient demographic data, the EHR also provided records for six different types of physiologic measures that were sampled periodically during the ICU stay. In addition, records on 26 types of laboratory values and 18 types of prescribed medications were also included. A summary is provided in Table 3. We also made use of a metric that is relatively unique to the Yale-New Haven Hospital, the Rothman Index (RI). This is an EHR-based measure of patient acuity that is continuously updated throughout an episode of hospitalization. The RI score is a composite measure updated regularly from the electronic medical record based on changes in 26 clinical measures including vital signs, nursing assessments, Braden score, cardiac rhythms, and laboratory test results.

Category	**Variable Names**
Rothman Index	Rothman Index
Physiology	Diastolic blood pressure, systolic blood pressure, temperature, Glasgow Coma Scale score, pulse (heart rate), and total respiratory rate
Laboratory values	Sodiumb, potassiumb, chlorideb, creatinineb, glucose, glucose meter, calcium, magnesium, phosphorus, WBC countb, hemoglobinb, hematocrit, international normalized ratio, lactate, bilirubin total, bilirubin direct, alanine transaminase, aspartate transaminase, alkaline phosphatase, albumin, prealbumin, troponin T, fibrinogen level, pH arterial, Po2 arterial, and Pco2 arterial
Medication	Vasopressors/inotropes, dobutamine, dopamine, epinephrine, norepinephrine, vasopressin, and phenylephrine Antibiotics/antivirals/antifungals: acyclovir, ceftriaxone, ciprofloxacin, doxycycline, ertapenem, fluconazole, gentamicin, moxifloxacin, vancomycin, valacyclovir, ampicillin-sulbactam, and piperacillin-tazobactam
Demographic	Age, gender, race, height, and weight
Chronic disease	Dialysis, chronic obstructive pulmonary disease, and HIV

Table 2.3: Raw Variables

Data Processing

Definition of Sample For each ICU visit, we generated (overlapping) observational units every 6 hours, where a unit is defined as a 24-hour window during the visit. For example, an ICU visit that lasted 39 hours generates the observational units (0, 24] hours, (6, 30] hours, and (12, 36] hours. We want to extract features from the time series data during these observational unit to predict 6-hour mortality within the end of the observational units. Figure 7 demonstrates how samples are taken and the targets to predict for each sample.

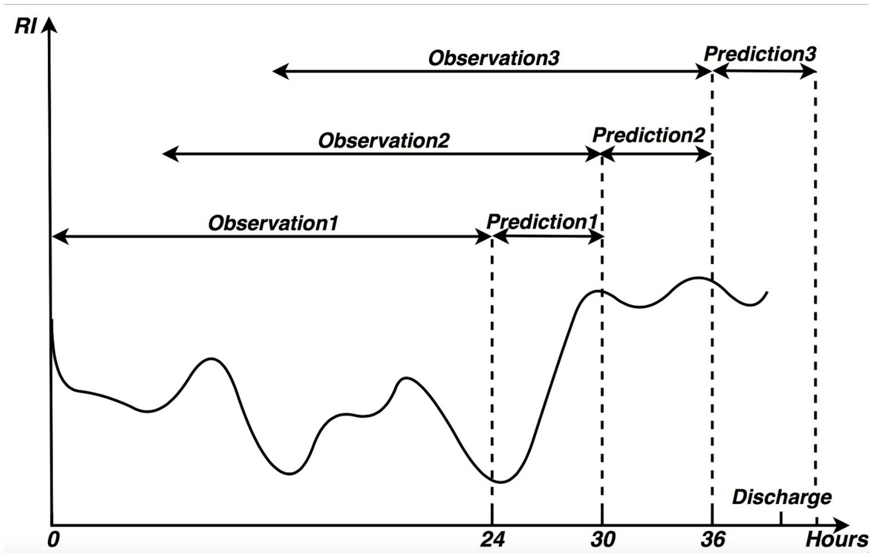


Figure 2.7: Observational Units and Prediction Targets.

Outcome of interest For each observational unit, we are interested in predicting 6-hour, 12-hour and 24-hour mortality, counting from the end of observational unit. If the patient died within the given time frame, the outcome is labeled 1. Otherwise, the outcome is labeled 0. On observational unit level, mortality within these time frames are rare events. Among all observational units, 1.2% were followed by death within 6 hours(1.9% for 12 hours, 3.4% for 24 hours, and 17% were followed by death during the same ICU stay).

Time Series Data Among the physiologic measures in Table 2.3,those that were sampled often enough to provide a usable time series during an ICU stay (updated hourly) include systolic blood pressure, diastolic blood pressure, heart rate, and the RI. To automatically extract features from each time series that characterizes their evolution over time, we fitted a continuous trajectory to each series in each observational unit. The basis functions that were used to interpolate the time series data points were cubic(order = 4) B-splines with 23 evenly spaced interior knots, and they were fitted using penalized least squares. This leads to 27 unique B-splines to represent each trajectory.

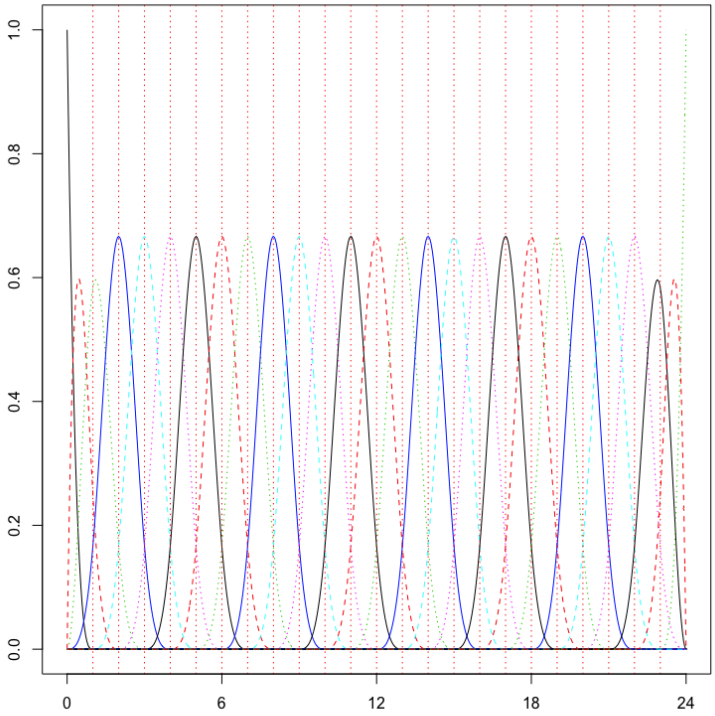


Figure 2.8: B-splines Used to Fit 24-hour Time Series for Each Observational Units.

Figure 8 displays an example of one such fit to RI time series data. the fitted coefficients capture information about the shape of the time series. We can then use these coefficients as features of the trajectory in our estimation model. To compare against the use of summary statistics that are manually selected from the trajectory in the literature, we also calculated the minimum, maximum, median, first value, last value, and the number of samples for each time series. We call these variables the "standard trajectory summaries."

Other Physiologic Measures. For the less frequently sampled physiologic measures, we used summary statistics for the measurements within each observational unit (maximum, minimum, and median) to represent their trajectories.

Laboratory Work. For infrequently performed laboratory tests, we used the latest

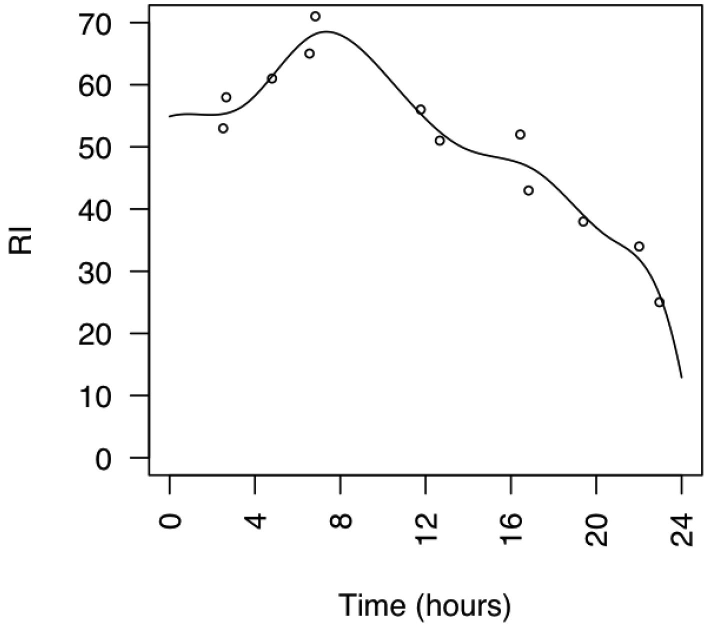


Figure 2.9: Fitted RI Curve Using Penalized B-splines.

measurement in the trailing 48 hours as the representative value for each observational unit. If a laboratory value was not available in the last 48 hours, it was treated as missing and was handled using the approach described below. The choice of 48 hours was taken from the protocol used in RI to handle laboratory values.

Medication Records. We consolidated patient medication records into two categories: a variable for the number of vasopressors and inotropes administered during the period, and an indicator variable that tracked whether a patient was on antibiotics, anti-fungals, or antivirals during the period. These variables serve as markers of shock and infection, respectively, and relate to disease processes.

Time in ICU. We also created a variable to record how long each patient had already spent in the ICU at the start of the observational unit.

Missing Values. For non time series variables, missing values were handled in the following way. Since the range for the non-missing values across all variables was substantially bounded away from −1,000, we encoded a missing value with this number to instruct our tree-based estimation algorithm to treat it differently. For time series data, an observational unit that had fewer than 12 measurements was considered to have a missing functional data point (the time series). Therefore, the coefficients for the 27-spline functions are all encoded as $-1{,}000$.

The prediction variables that were created above for each observational unit can be fed into any classifier to estimate the probability of death within the next 6 hours in the ICU (or indeed, any number of hours). We use the random forest classifier as our platform for investigating the use of trajectory data because it is a popular non-parametric method that consistently ranks as one of the top performing prediction tools.

2.4.3 Comparing different models

Employing this to estimate the probability of death, we assess the performance gains resulting from using our proposed trajectory variables in lieu of the standard trajectory summaries defined earlier. We do this by comparing four nested random forest models M1–M4 (Table 2.4 for details). In brief, M1 uses the most current values of all predictor variables aside from RI. M2 appends the additional statistics that make up the standard trajectory summaries used in the literature. M3 also adds the standard trajectory summaries for RI. Finally, M4 replaces all summaries with spline coefficients that capture the shape of the entire trajectory.

We used Monte Carlo cross validation to evaluate the performance of the four models: We performed 20 random splits of our set of unique patients into training (70%) and validation (30%) sets. Having 20 different splits of the data reduces the bias that could arise from any one particular split. For each split, we fit a random forest classifier to the observational units in the training set. The model was then used to predict the probability of death within the next 6 hours for each observational unit in the validation set.

Receiver operating characteristics (ROCs) plots are commonly used to assess the performance of binary classifiers. However, they can be misleading in situations where the outcome classes are highly imbalanced. Such is the case here since the number of observational units with a mortality event (1.2%) is much lower than the number of units without. For imbalanced outcomes, simulation studies (23–25) suggest that the precision-recall plot is more informative where “precision” refers to positive predictive value (PPV) and “recall”

Model	**Variables Used**
(M1) Snapshot of variables without RI	Most recent values for time series data except RI (diastolic blood pressure, systolic blood pressure, and heart rate), All other variables except RI
(M2) Trajectory summaries without RI	Add standard trajectory summaries of time series data (except RI) to model M1
(M3) Trajectory summaries with RI	Add standard trajectory summaries for RI to model M2
(M4) Full trajectory	Replace standard trajectory summaries in model M3 with the spline coefficients that capture complete trajectories

Table 2.4: Full Description of Model Variables.

refers to sensitivity. In light of this, we calculated both the AUROC curve and the AUPRC averaged over the 20 splits for each model, along with 95% CIs. We ran paired sample t tests to compare average AUROC and average AUPRC between different models.

We also applied the Hosmer-Lemeshow test to each of the 20 test sets to evaluate the calibration of the predictive model. Conceptually, if a well-calibrated model assigns (say) z% chance of a death event to each of 100 observational units, then about z out of the 100 should result in an actual death. The Hosmer-Lemeshow test measures the discrepancy between the expected and observed death rates for the observational units, with the null hypothesis being that the two quantities agree. In performing the test, we followed the guidelines in [42] for analyzing large datasets.

In addition to making mortality predictions over a 6-hour window, we also repeated the above mentioned analysis for 12- and 24-hour windows in Appendix. For 12-hour windows, predictions are made every 12 hours based on the trajectory of the measurements over the previous 24 hours. For 24-hour windows, predictions were made every 24 hours.

Model	AUROC	AUPR
(M1) Snapshot of variables without RI	0.883 (0.877–0.889)	0.342 (0.333–0.351)
(M2) Trajectory summaries without RI	0.887 (0.882–0.892)	0.350 (0.339–0.361)
(M3) Trajectory summaries with RI	0.896 (0.892–0.900)	0.366 (0.353–0.379)
(M4) Full trajectory	0.905 (0.900–0.910)	0.381 (0.368–0.394)

Table 2.5: Out of Sample Area Under ROC and Area Under Precision-Recall for 6-Hour Mortality Prediction(95% CI in Prentheses)

2.4.4 Results

Table 2.5 compares the AUROC and AUPRC for 6-hour mortality averaged over the 20 splits of the data for models M1–M4 described in Table 2.4. Figure 2.10 and 2.11 plots out the ROC and Precision-Recall Curve for M1-M4. The table encapsulates three findings, which also hold for the 12- and 24-hour prediction windows as well(Table 2.6)

	12 Hour		24 Hour	
Model	AUROC	AUPR	AUROC	AUPR
M1	0.876(0.871,0.881)	0.346(0.336,0.356)	0.870(0.866,0.874)	0.344(0.330,0.358)
M2	0.879(0.873,0.885)	0.353(0.341,0.365)	0.874(0.869,0.879)	0.354(0.341,0.367)
M3	0.888(0.881,0.895)	0.368(0.354,0.382)	0.886(0.881,0.891)	0.369(0.355,0.383)
M4	0.876(0.871,0.881)	0.346(0.336,0.356)	0.870(0.866,0.874)	0.344(0.330,0.358)

Table 2.6: Out of Sample Area Under ROC and Area Under Precision-Recall for 12-Hour and 24-hour Mortality Prediction(95% CI in Parentheses)

First, the AUROC and AUPRC for M4 (0.905 and 0.381, respectively) are both higher than those for M3 (0.896 and 0.366), and the differences are statistically significant (p = 0.004 and 0.017, respectively). In other words, the spline representation of the trajectories of time series data conveys additional predictive information that are not already captured by the standard trajectory summaries. Furthermore, we calculate total decrease in node impurities(measured by Gini Index[46]) from splitting on each variables, averaged over all trees. This gives the relative importance of variables(Figure 2.12). Figure 2.12 shows that

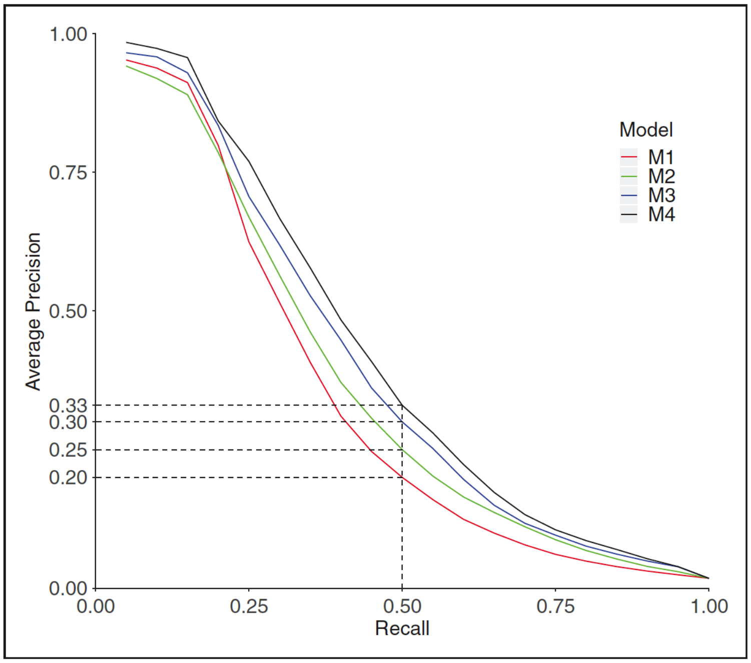


Figure 2.10: Average PR Curves for Model M1-M4.

more than half of the 20 most important variables are the spline coefficients for RI and pulse, particularly the ones describing the most recent evolution of the times series (e.g., “pulse27” is the coefficient for the last spline function in the 24-hour). This reflects the time decay in the predictive power of the time series data.

Second, the performance of M3 is in turn statistically significantly better than M2 with AUROC of 0.896 versus 0.887 ($p < 0.001$) and AUPRC of 0.366 versus 0.350 ($p = 0.002$). That is, the RI conveys additional predictive information over the other variables used in M2, including 11 of the 26 components used to calculate RI. This is reinforced by Figure 2.12, which shows that eight of the 20 most predictive variables are related to the trajectory of the RI.

Third, the difference in performance between M1 and M2 is small (0.883 vs 0.887 for AUROC and 0.342 vs 0.35 for AUPRC) and not statistically significant ($p = 0.363$ for AUROC and $p = 0.236$ for AUPRC). In other words, beyond the most recent measurement, the additional trajectory summaries employed in the literature do not add meaningful predictive power.

To make the performance measures more concrete, it is helpful to consider a single point

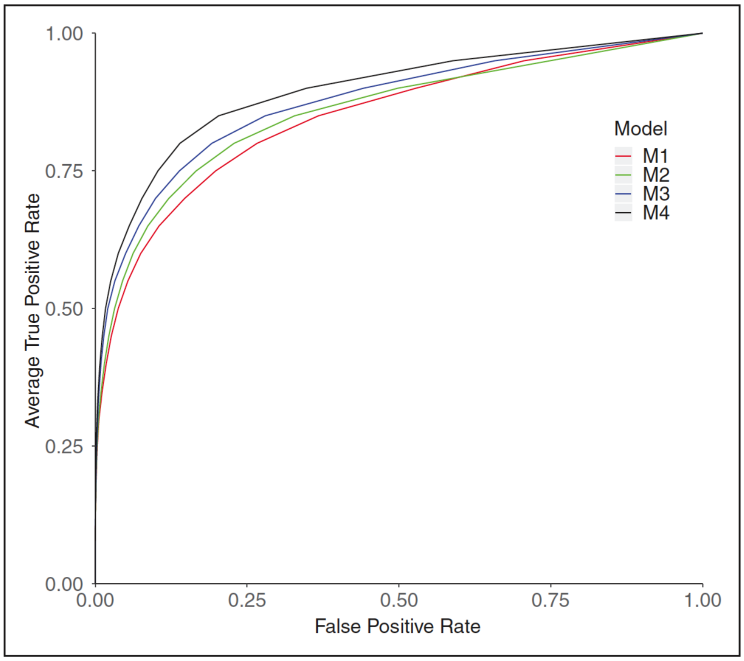


Figure 2.11: Average ROC Curves for Model M1-M4.

on the Precision-Recall curve. Figure 2.10 and 2.11 display the precision-recall curves and ROCs averaged over the 20 splits of the data for models M1–M4. At 50% recall (sensitivity), the average PPV for 6-hour mortality prediction is 33% (compared with 21% for M1, 25% for M2, and 30% for M3). In other words, if the observational units flagged by our algorithm are to include half of those that resulted in death within 6 hours, then one-third of all flagged cases will be correct. This seems like a low accuracy but compared with low rate of mortality in 6-hour windows it is informative. To elaborate, because 1.2% of the observational units were followed by death within 6 hours, this means that we can identify half of all potential deaths by focusing on just the top $1.2\% \times 50\%/33\% = 1.8\%$ of observational units with the highest predicted probabilities of death.

Finally, the results of the Hosmer-Lemeshow test (Table 2.7) showed that our final model M4 is well calibrated: For the 20 test sets, the p values for only three of them were less than 0.20 with the smallest one being 0.094. Thus, the null hypothesis was never rejected.

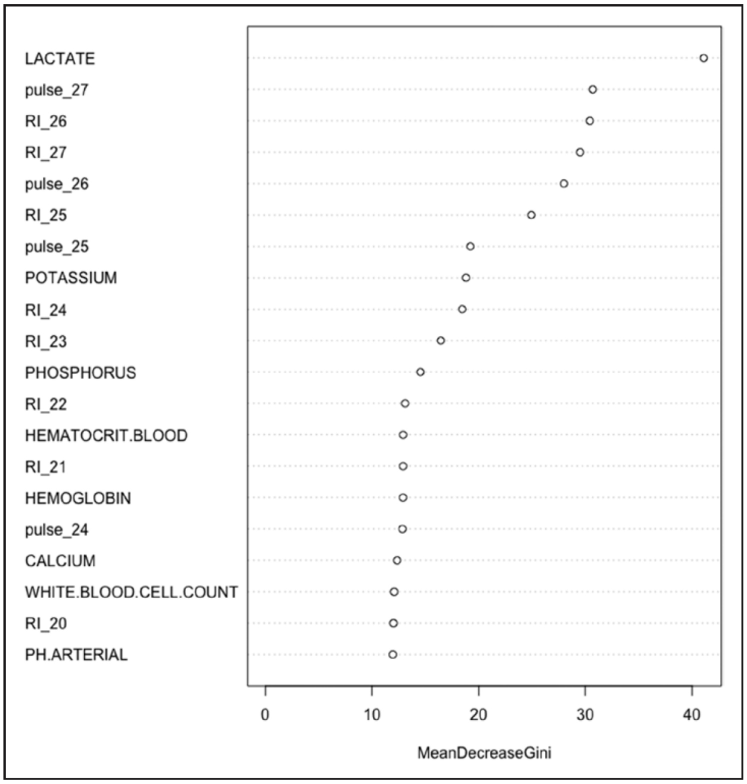


Figure 2.12: Variable Importance Plot for Model M4.

2.4.5 Conclusion

We have shown that using trajectory information of clinical data can improve the accuracy of mortality predictions. This benefit is apparent for both an approach that uses trajectory summary statistics and an approach that uses an algorithmically generated functional representation of the trajectory. Mortality over short time horizons is a very rare event even for acutely ill patients, which makes it difficult to predict. Our approach indicates how making fuller use of the available clinical data can help address this difficulty. One would expect that this approach would also benefit predictions of other outcomes for which trends in patient's health performance could be useful indicators such as readmission or response to specific treatments.

Any short-time horizon mortality prediction method has the potential to be the basis for a clinical early warning system. Nursing units and intensivists in ICU are limited resources. An accurate mortality prediction model have can potentially prioritizing medical resources to inspect the most needy patient could.

Seed	**P-value**	**Seed**	**P-value**
1	0.176	11	0.975
2	0.999	12	0.998
3	0.283	13	0.999
4	0.781	14	0.796
5	0.999	15	0.984
6	0.999	16	0.999
7	0.198	17	0.738
8	0.094	18	0.998
9	0.347	19	0.892
10	0.688	20	0.201

Table 2.7: P-value for Hosmer-Lemeshow Test of Each Split.

This model and analysis have a few limitations: First, like the End-of-Life mortality predictions in section 2.3, the model is only trained on single hospital's MICU. A multi-center study is needed to prove that our approach can be generalized to different kinds of medical units. Second, although our new prediction model shows a significant increase in both area under ROC and area under Precision-Recall, it's not clear how these increase would turn into actual benefit when working as decision support systems. To truly assess if a warning system is useful in this setting it is necessary to determine if the warning would identify high mortality risk cases that: 1) are not already known to the nurses and physicians and 2) could be aided by an intervention.

2.5 Discussion

In this chapter, we presented a novel approach for using functional data analysis to capture the temporal trends of EHR data. Compared to traditional methods like taking a snapshot and manually selecting summary statistics, this can significantly increase the prediction accuracy in two settings:

- EOL project: Predict 90-day(or similar magnitude like 60-day, 180-day) mortality for advanced cancer patients using 48-hour RI trajectory.
- MICU project: Predict 6-hour(or similar magnitude like 12-hour, 24-hour) mortality for Medical ICU patients using 6-hour RI and physiologic trajectory.

There are three key differences between the EOL and MICU settings.First, they have different time frame. In the EOL project, we focus on predicting 90-day or 180-day mortality. While in MICU project we do predict mortality within much shorter time frame like 6, 12 or 24 hours.Second, their purposes are different. The EOL project aims at identifying patients who win die within 90-days or 180-days given the disease runs its normal course. The goal is to minimize the aggressive treatments for those who are identified as close to death. The MICU project aims at identifying with high near-term mortality rate in order to avoid as much of these mortality as possible. Third, available data is different. Other than RI, the MICU project collects more patient level time series data including physiologic, lab results, medication and ventilation,

Despite these differences, we achieved significantly improved prediction accuracy in both projects by introducing functional data analysis techniques. As discussed in previous sections, one of the ultimate goals of designing these mortality prediction algorithms is to implement them as decision support systems. We recognize that more questions need to be answered before these tools can prove their efficacy as decision support.

For the EOL project, although IMPAC could potentially benefit the end-of-life decision process, it remains unclear (1) how this tool works compared to what the physicians do;(2) if physicians and IMPAC identify same group of patients; (3) what would be a good way for IMPAC and doctors to collaborate; (4) what's the potential benefit of this collaboration.

We will try to answer these questions by conducting a retrospective study using the same data in Chapter 3.

For the MICU project, while many similar works in recent years focus on building sophisticated machine learning models that brings benefits measured by area under ROC or Precision-Recall, it remains unclear how these increase would turn into actual benefits when working as an alert system in practice. In chapter 4, we will present a theoretical POMDP framework that evaluate the benefits of any given clinical warning system based on mortality predictions.

Chapter 3

Improving End of Life Care by Augmenting Physician Discharge Decision Process with Mortality Prediction Algorithms.

3.1 Introduction

There is great interest in developing tools to improve physician decision making. In particular, computer based statistical models that use sophisticated data analysis techniques and large clinical data sets are viewed as having great potential that is only increasing as computing power increases, data collection expands, and analysis methods advance. Before these tools become widely adopted, they need to, and should, overcome a number of hurdles. In this chapter we use the framework of mortality predictions for advanced cancer patients and the IMPAC model developed in Chapter 2 to explore some of the important issues related to assessing the use of prediction models in healthcare. While many of the topics we discuss here are generalizable to other settings, they will be clearer and more concrete within the context of a specific decision making setting.

Let's consider an oncologist treating a hospitalized patient knowing that the patient has an advanced cancer that is not curable. The physician must decide how aggressively to treat the patient and whether to refer the patient to palliative care. The physician's assessment of the patient can be summarized as the probability r_d that the patient will die within 90 days, i.e. the 90-day mortality risk of the patient. If this probability is greater than a threshold τ the physician will recommend palliative care and if $r_d < \tau$ the physician will recommend more treatment. The threshold τ to use in this setting will depend upon the potential quality of life benefits and life extension potential of additional treatment relative to palliative care. We do not explore that question here. However, given a threshold τ one can characterize the physician as classifying patients into one of two possible groups. Similarly, given τ, the role of a prediction algorithm is to classify a patient. A prediction algorithm would receive a variety of inputs (independent variables) describing the patient and would return a probability r_a of the patient dying within 90 days.

In Chapter 1 we developed a method for generating r_a and assessed its accuracy. In this chapter our goal is to assess the value of having an algorithm that can generate r_a in this setting. Apriori there are reasons to assume that such an algorithm would be useful. End-of-life care for patients with advanced cancer is known to be aggressive, costly, and often discordant with patients' wishes.[29, 30, 47, 48] Among Medicare decedents, 80% were hospitalized within 90 days of death, 27% were admitted to the intensive care unit (ICU) within 30 days, and 20% transitioned to hospice in the last 3 days.[31] Thirty percent of all spending for cancer occurs in the last year of life. [32, 33] Thus there is evidence that there is expensive care being given to patients at the end of life that is not benefiting them. There is also considerable evidence that physicians are not very good at make estimations of life expectancy near end-of-life. Oncologists, [34, 35, 37]and general physicians [36] have all been shown to be overly optimistic in their prognoses. In a study involving five outpatient hospice programs[36], 343 doctors provided survival estimates for 468 terminally ill patients at the time of hospice referral. Patients' estimated and actual survival. Median survival was 24 days. Only 20% (92/468) of predictions were accurate (within 33% of actual survival); 63% (295/468) were overoptimistic and 17% (81/468) were over-pessimistic. As a result, patients sent to hospice tend to be extremely close to their deaths. This study also showed that these estimations are significantly affected by individual characteristics such as gender and experience in years. This lack of accuracy and consistency are both likely to be alleviated with the help of a prediction algorithm.

To deliver value, a prediction algorithm must be adopted and there are significant hurdles to that. The algorithms must be widely tested prospectively to be viewed credibly by physicians. Algorithms will also face resistance if they are perceived to be black boxes because their methods and outputs are hard to interpret. Even if these hurdles are overcome it is still necessary to assess their value to decision makers. To assess the value of a prediction algorithm in this setting it must improve the classification performance of the physicians and do so in a way that can be actionable and reduce unnecessary care at end of life. It is actually remarkably rare, in the literature, for the performance of prediction algorithms to be compared to that of physicians(David 2017). Most studies that do compare performance between prediction approaches also do so in limited ways. The comparison of an algorithm to a physician is also a false comparison. It assumes that physicians and algorithms are substitutes when in reality if an algorithm is implemented it will be used to support physician decision making.

In this chapter we will compare the predictions of physicians and the IMPAC algorithm across multiple dimensions. We will show that there are important differences in the classifications they make and the thus potential for information gain by using them in conjunction. We will analyze different approaches to integrate the two and compare performance. In section 3.2 we describe the data set and review the prediction method. In section 3.3 we compare the performance of the prediction algorithm and physicians in terms of various measures of prediction accuracy. In section 3.4 we integrate the two prediction methods and analyze the prediction accuracy performance of the combination. We conclude in section 3.5 with an illustration of the potential benefits of better predictions to reduce aggressive treatments at end of life.

Oncologists are not capable of estimating, let alone recording in real time, probabilities or survival time. This causes two major limitations of our study. First, we are not able to observe the part of patients who declined doctor's offer to send them to hospice. Second, we are not able to identify the earliest time doctors realize that the patient needs to be sent to hospice. Both will make doctors mortality predictor in our data set more conservative than reality. However, we believe this is the most that we can do with our available data.

3.2 Data Description

3.2.1 Study Population

We examined the inpatient records of 2774 unique patients with advanced solid tumors identified by the Yale New Haven Hospital tumor registry with a hospital discharge(for any cause of admission) between October 1, 2013 and June 31, 2017. These patients, in total, generated 4778 inpatient encounters, during this time period.

3.2.2 Variable definitions

For the purposes of the analyses and discussion in this chapter we use 90-day mortality from the date of discharge as the dependent variable. Patients' survival status was collected from the institutional tumor registry as of June 31, 2019. For patients with more than one hospitalization during the study period, each visit (encounter) was treated as a separate observation. For each unique patient we have demographic data and a distinct data record during each encounter. For each encounter we have static variables and a time series of health status scores called the Rothman Index (described below). The independent variables are defined as follows:

- **ED Admission** If the patient got admitted to hospital through emergency department, this field is set to 1. Otherwise, this field is set to 0.
- **Previous Admission in Last 90 days** If the patient had at least one hospitalization within the past 90 days from current hospital admission, this field is set to 1. Otherwise, this field is set to 0.
- **Unique patient identifier**
- **Gender**
- **Age** Age of the patient (integer) at the point of current hospital admission.
- **Race** Race of the admitted patient.

- **Disease Team** The specialized cancer teams that primarily treats the patient. This is a proxy of the type of cancer the patient has.

- **Discharge Disposition** This field indicated where patient is discharged to. Possible values are: Home, Hospice, Nursing Facility, Self care, Rehabilitation Center, and Other Hospital Facilities.

Summary statistics for these variables are reported in Table 3.1.

Rothman Index

We used a metric available in the Electronic Health Record at Yale New Haven Hospital, the Rothman Index (RI; PeraHealth). RI is a real-time, EHR-based scalar measure of patient acuity that is continually calculated throughout the hospitalization and has been incorporated into most commercially available EHRs. It incorporates 26 clinical data elements, including vital signs, nursing assessments, and laboratory results, and has been shown in multiple settings to predict both mortality and readmission. Lower RI scores reflect a worse health status.To ensure that enough RI scores were spread across a sufficient period to inform the model, we excluded visits with less than 48 hours of RI monitoring (excluding 71 patients) and for which no RI was available between 36 and 48 hours (33 excluded). The final data set consisted of 2,774 unique patients with 4,575 inpatient encounters.

3.2.3 Prediction Model

We use the imminent mortality predictor for advanced cancer(IMPAC) introduced in Section 2.3 as the decision support tool to compare to physicians' predictions and briefly describe the methodologies used in IMPAC.

Since patients' physiologic and Rothman Index data are stored as time series by EMR system, it is natural to use curves as basic units of analysis, which is termed as functional data analysis (FDA). In comparison to traditional manual feature extraction, FDA has the advantage of fully extracting temporal trend of original time series data.

The original IMPAC used the RI trajectory during the first 48 hours of encounter to predict 30, 60, 90 and 180-day mortality starting from 48 hours after hospital admission.

Characteristics	Total Number	Mean	Standard Deviation	Percentage
Patient Specific	2774	-	-	-
Age, years	-	63	12	-
Gender, Males	-	-	-	46.0%
Visit Specific	4575	-	-	-
Entry Through ED	-	-	-	41.0%
Prior visit in last 90 days	-	-	-	35.0%
Length of Stays, days	-	6.74	6.84	-
Type of Cancer	-	-	-	-
Breast	-	-	-	7.3%
Endocrine	-	-	-	2.4%
GI	-	-	-	24.0%
Genitourinary	-	-	-	5.8%
Gynecologic	-	-	-	15.0%
Head and Neck	-	-	-	10.0%
Melanoma	-	-	-	3.5%
Neurologic	-	-	-	7.9%
Sarcoma	-	-	-	4.9%
Thoratic	-	-	-	18.0%
Undefined and Unknown	-	-	-	0.74%
Discharge Disposition	-	-	-	-
Hospice	-	-	-	7.2%
Home/Home Care	-	-	-	72.9%
Nursing Facility	-	-	-	16.8%
Rehabilitation Center	-	-	-	1.4%
Other Facilities or Unknown	-	-	-	1.3%

Table 3.1: Summary Statistics for Final Data Set

However, the aim of this new study is to compare and combine doctor's decision and IMPAC. In order to make fair comparison, we want to let them do the prediction at the same time. So it is more reasonable to train this model to predict mortality from the date of discharge.

Due to technical reasons, each patient's RI is not measured at synced times and number of observations during a fixed-length time window can vary significantly (Figure 3.1). Thus, we are not able to align these observations to generate a tabular data set. Instead, we used a system of K spline basis functions $\phi_k(t)$ in the linear combination to fit the trajectory of Rothman Index over 48 hours in the following form:

$$RI(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{3.1}$$

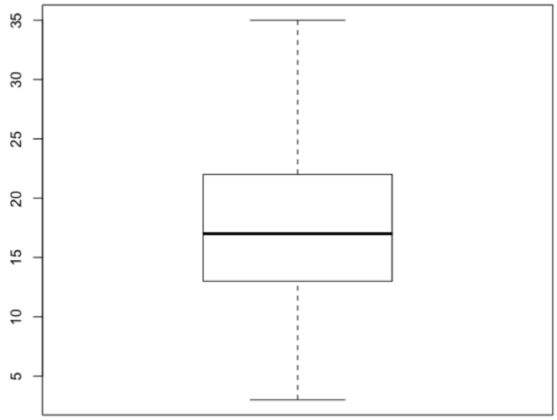


Figure 3.1: Boxplot: Number of RI Observations During 48-hour Time Window

$$\sum_{j}^{n} [y_j - RI(t_j)]^2 + \lambda \int [D^2 RI(t)]^2 dt \tag{3.2}$$

$$= \sum_{j}^{n} [y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2 + \lambda \int [D^2 (\sum_{k=1}^{K} c_k \phi_k(t))]^2 dt, \tag{3.3}$$

Suppose $\{(y_j, t_j)\}$ is a patient's discrete RI time series, a smoothed curve $RI(t)$ can be fit by minimizing the following penalized least squares criterion:

$$\sum_{j}^{n}[y_j - RI(t_j)]^2 + \lambda\int[D^2RI(t)]^2dt \tag{3.4}$$

$$=\sum_{j}^{n}[y_j - \sum_{k=1}^{K}c_k\phi_k(t_j)]^2 + \lambda\int[D^2(\sum_{k=1}^{K}c_k\phi_k(t))]^2dt, \tag{3.5}$$

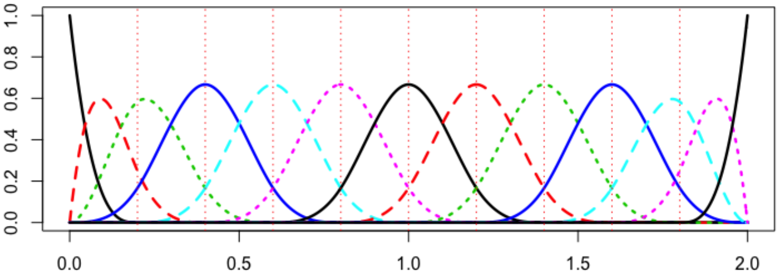


Figure 3.2: Set of 13 Basis Functions

where λ is penalty parameter. The larger λ is, the more $x(t)$ will be penalized for over fitting and $x(t)$ will be closer to a straight line.

Next, we utilized functional principal component analysis (FPCA), which is a dimension-reduction tool by identifying important modes of variation in high dimensional functional data. FPCA is defined as an orthogonal linear transformation that transforms the functional data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on (Jolliffe 2002). We will use the first three functional principal components, noted as PC1, PC2 and PC3 as features that represent temporal trends of Rothman Index during the last 48 hours of the encounter. Thus for each encounter the time series of RI scores is replaced by three variables that are the weights on the first three functional principal components. With this compression of the RI data complete we then train a Cox proportional hazard model in R([26] to estimate the survival time distribution of each patient.

In our case, parametric model based on exponential distribution may be written as:

$$\begin{aligned} log(h(t)) =& log(h_0(t)) + \beta_1 age + \beta_2 Gender + \beta_3 PC_1 + \beta_4 PC_2 + \beta_5 PC_3 + \\ & \beta_6 EntryThroughED + \beta_7 VisitIn90Days + \beta_8 Breast + \\ & \beta_9 Endocrine + \beta_{10} GI + \beta_{11} Genitourinary + \beta_{12} Gynecologic + \\ & \beta_{13} HeadAndNeck + \beta_{14} Melanoma + \beta_{15} Neurologic + \\ & \beta_{16} Sarcoma + \beta_{17} Thoratic \end{aligned} \tag{3.6}$$

Given fitted hazard rate, we can calculate the risk score of IMPAC, r_I, which represents probability of surviving less than 90 days, or equivalently, dying within 90 days after current hospital discharge.

3.3 Accuracy Comparison of IMPAC and Doctors

One of the challenges of comparing physician and algorithm predictions is that physician predictions are seldom recorded and thus retrospective studies cannot be used. At the same time prospective studies are difficult to execute. In this chapter we will use the recorded discharge dispositions of patients as a proxy for physician predictions about their life expectancy. If a physician sends a patient to hospice it implies that they believe the patient will die soon. The median survival time of patients sent to hospice in [36] was 24 days with the physician's median survival prediction 126 days. In our data the median survival time of patients discharged to hospice was only about 9 days. Based on conversations with oncologists at our study site we concluded that it was unlikely that an oncologist would discharge a patient to hospice if they felt they had significant chance of surviving more than 90 days. In fact 93% of the patients in our sample who were discharged to hospice died within 90-days.

The main limitation of using discharge dispositions as physicians' predictions is that we are only able to observe those patients who were sent to hospice. We are not able to observe when a doctor predicts that the patient will die within 90 days and recommends hospice but the patient or family declined to go to hospice. Thus we will underestimate the sensitivity of the real predictions made by doctors. Another limitation of our approach is that we aggregate the decision making of multiple physicians. Our data does not include information on individual physicians. There will be heterogeneity across physicians in terms of their risk assessments and conservatism regarding palliative care recommendations. Given that all the physicians are at the same institution and consult with each other we believe that the aggregate performance of the physicians is a good benchmark for physician prediction abilities.

3.3.1 Comparison Criteria

For a classification algorithm that generates mortality risk scores(ranging from 0 to 1), one needs to set a threshold and predict patients with risk scores higher than the threshold to die soon. As this threshold increase, less patients get positive predictions and advised to hospice. But the overall precision of these positive prediction would also change. We use precision recall curves to measure the success of our mortality prediction algorithms.

Precision(P), also known as positive predictive value, is defined as the proportion of the patients discharged to hospice that died within 90 days

$$P = \frac{TP}{TP + FP} \tag{3.7}$$

Recall(R), also known as sensitivity, is defined as the proportion of patients who died within 90 days that were sent to hospice.

$$R = \frac{TP}{TP + FN} \tag{3.8}$$

Compared to ROC, precision recall curve works better in the two scenarios. The first one is when outcome is imbalanced. The second one is when there's a huge difference between cost of False Positive and cost of False Negative. We will use precision recall curve as one of the major criteria to compare different risk predictors in our study for two reasons. First,in end-of-life settings, it is commonly believed that a false positive cost much more than a false negative. Second, outcome of 90-day mortality is imbalanced, with only 28% of all cases being positive.

Unlike classification algorithms, physicians, like other humans, find it difficult to consistently estimate probabilities and thus it makes sense to view physicians as operating with an unknown threshold τ_d. Associated with this threshold is a resulting aggregate precision and aggregate recall across all physicians.

$$P_d = P(\text{Die in 90 days}|r_d > \tau_d) \approx 0.93 \tag{3.9}$$

the recall achieved by the physicians is

$$R_d = P(r_d > \tau_d|\text{Die in 90 day}) \approx 0.23 \tag{3.10}$$

Although these classifications come with a high precision level, physicians identify only less than a quarter of those patients who die within 90 days from discharge. This indicates that physicians either (1) tend to overestimate the survival time of patients or (2) make very conservative discharge decisions. Of course as we mentioned above this estimate of physician recall is likely lower than reality because we do not observe the cases where hospice was recommended but either refused or unavailable. On the other hand it is unlikely that the unobserved cases would significantly increase precision.

3.3.2 Algorithm Performance

Following the same procedure in Chapter 2, we use Monte-Carlo cross validation [42] to create random training testing splits. For 20 iterations, we randomly split the raw data set into training and testing set on a 7:3 ratio by patient ID. This avoids using a patient's future information to predict the past. For each iteration, we use linear interpolation to calculate precision levels corresponding to 101 evenly spaced recalls between 0 and $1([0, 0.01, 0.02, ..., 1])$. Simple average of precision over 20 splits are calculated at each key recall level. Average precision and recalls for IMPAC prediction and doctor prediction are plotted(Figure 3.2). Similarly, we calculated average True Positive Rate and False Positive Rate and plot the average ROC in the appendix. We can see that precision recall and ROC curves almost goes through doctors' precision recall and tpr-fpr point respectively(Figure 3.3). This means we are not able to find any threshold such that corresponding decision will be strictly better than what the doctors achieve.

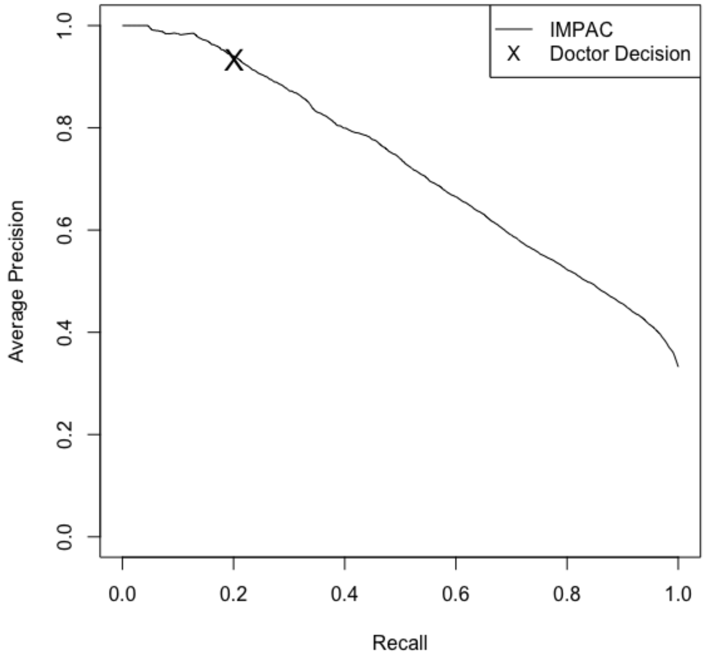


Figure 3.3: Average Precision Recall Plots of IMPAC and Doctors

This indicates that at current recall level(0.23), doctors can achieve approximately the same precision(0.93) by solely acting on their own opinion or solely following IMPAC. However, we should not conclude from this figure that IMPAC can't bring additional benefits to doctors' decision making process. It's likely that IMPAC and doctors are identifying different groups of patients. We will further discuss this in following sections.

3.3.3 Overlap between Two Predictors

In order to see whether IMPAC and doctor make highly overlapped positive predictions, we did two analysis.

1. We binned IMPAC scores into 10 buckets of length 0.1 and calculated the percentage of patients who were sent to hospice by doctors within each bucket(Table 3.2)

2. or each visit, we find 100 visits that has the smallest absolute difference in score. Percentage of 100 neighbors who were sent to hospice was then calculated and plotted with the IMPAC score as x-axis(Figure 3.3).

Both table 3.2 and Figure 3.3 shows a positive correlation between IMPAC score and probability of being sent to hospice. However, even at the very top of IMPAC's range, only 53.8% of positive cases were correctly identified by doctors. This indicates that there's a significant difference between sets of positives identified by IMPAC and doctors.

IMPAC score range	Percentage Sent to Hospice
[0,0.1]	0%
(0.1,0.2]	0%
(0.2,0.3]	0%
(0.3,0.4]	0.4%
(0.4,0.5]	0.8%
(0.5,0.6]	7.3%
(0.6,0.7]	12.5%
(0.7,0.8]	23.2%
(0.8,0.9]	41.9%
(0.9,1]	53.8%

Table 3.2: Percentage sent to hospice binned by IMPAC

3.4 Augmentation Oncologists' End-of-Life Decision with IMPAC

From the previous section, we drew two conclusions from data:

- Doctors' collective precision recall point almost exists on precision-recall frontier of IMPAC
- IMPAC's positive prediction get captured by doctors at no more than 53.8% chance.

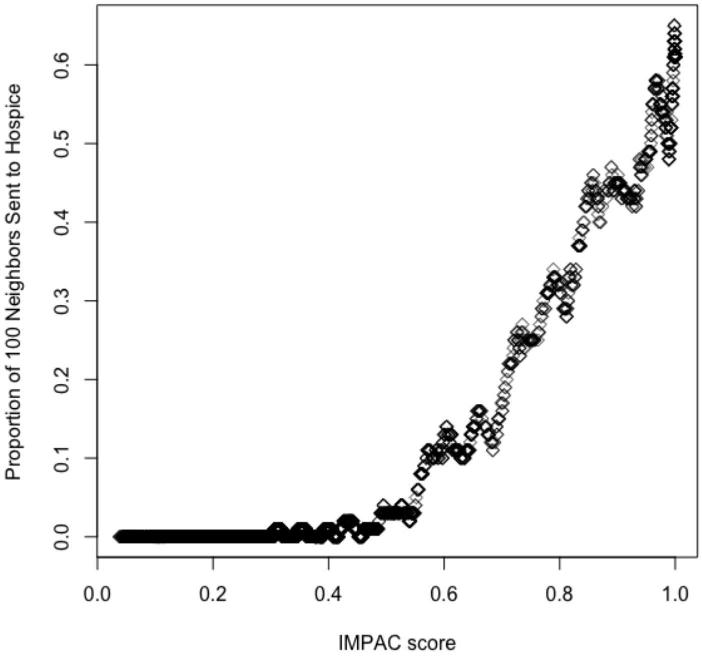


Figure 3.4: Proportion Sent to Hospice among 100 Neighbors with Closest IMPAC score

This incentivized us to further explore potential benefits from augmenting oncologists' end-of-life decision with IMPAC. We can assume both doctor and IMPAC make predictions at exactly the time of discharge. Although doctors may have identified these cases earlier than discharge date, there's relatively independent criteria for choosing discharge dates.Generally, patients need to finish their treatment in current episode to be discharged It's reasonable to assume that the observed discharge time is the earliest possible time of discharge. We assume that at the midnight before discharge date, the doctor made the prediction about if the patient will die within 90 days. We will also run IMPAC at the same time to make the fair comparison.

3.4.1 Pooling Method

Suppose IMPAC gives a score r_I. Given a risk score threshold c, we will classify those patients who are either (1) sent to hospice by physicians or (2) given a score that is no less than c as dying within 90 days. We call this method pooling because it pools positive cases in both opinions and send all of them into hospice. Since the positive cases given by pooling method contains all positive cases given by oncologists, recall level for this method can only range from R_D(when $c = 1$) to 1(when $c = 0$). We use $P_p(c)$ and $R_p(c)$ to denote precision and recall corresponding to cutoff level c when pooling the physician and algorithm classifications. We give the flow chart of pooling method in figure 3.4.

$$P_p(c) = P(\text{Die in 90 Days}|r_d > \tau_d \text{ or } r_I > c) \tag{3.11}$$

$$R_p(c) = P(r_d > \tau_d \text{ or } r_I > c|\text{Die in 90 days}) \tag{3.12}$$

3.4.2 Agreement Method

Given a risk score threshold c, we will only classify as dying within 90 days those patients who are both (1) sent to hospice by physicians and (2) given a risk score by the algorithm that is no less than c. We call this approach to integrating the classifications as the agreement method because it only send patients to hospice when both opinions agree to do so. Since the positive cases given by the agreement method is a subset of positive cases given by oncologists, recall level for this method can only range from 0(when $c = 0$) to R_D(when c = 1). We use $P_a(c)$ and $R_a(c)$ to denote precision and recall corresponding to an algorithm cutoff level c. We give the flow chart of agreement method in figure 3.5.

$$P_a(c) = P(\text{Die in 90 Days}|r_d > \tau_d \text{ and } r_I > c) \tag{3.13}$$

$$R_a(c) = P(r_d > \tau_d \text{ and } r_I > c|\text{Die in 90 days}) \tag{3.14}$$

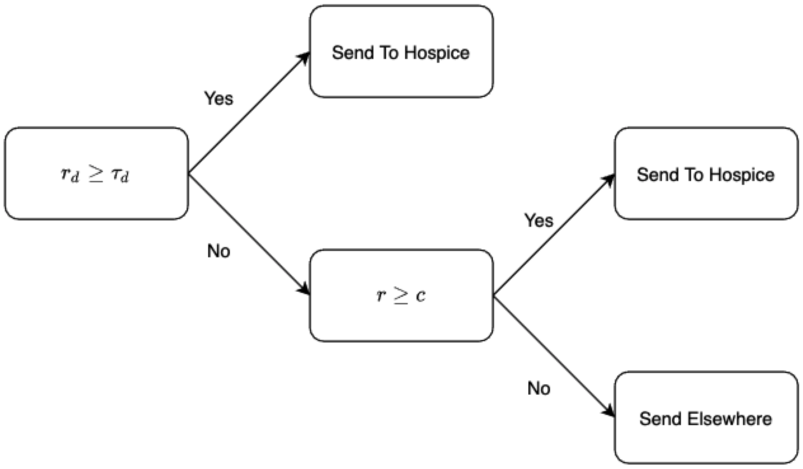


Figure 3.5: Flow Chart for Pooling Method.

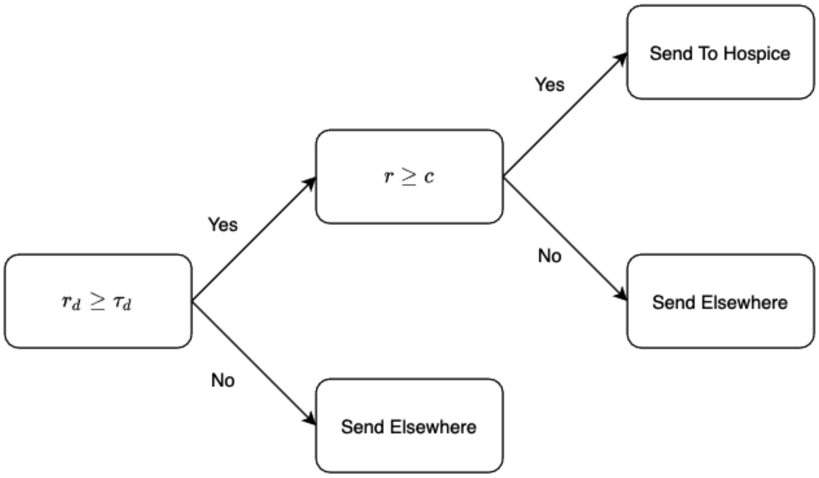


Figure 3.6: Flow Chart for Agreement Method.

3.4.3 Augmented Method

If we concatenate the precision recall curve of Pooling method and Agreement method, we can get a complete Precision-Recall curve with recall ranging from 0 to 1. This gives doctors the flexibility to be either more conservative or less conservative than they originally did. Since we are adjusting doctor's recall by changing c of IMPAC, we call this augmented method. Figure 4 demonstrates who the two pieces of augmented PR curve is connected by

PR point of doctor's prediction.

$$P_{aug}(c) = P_a(c)I(R_{aug}(c) <= R_d) + P_p(c)I(R_{aug}(c) > R_d) \tag{3.15}$$

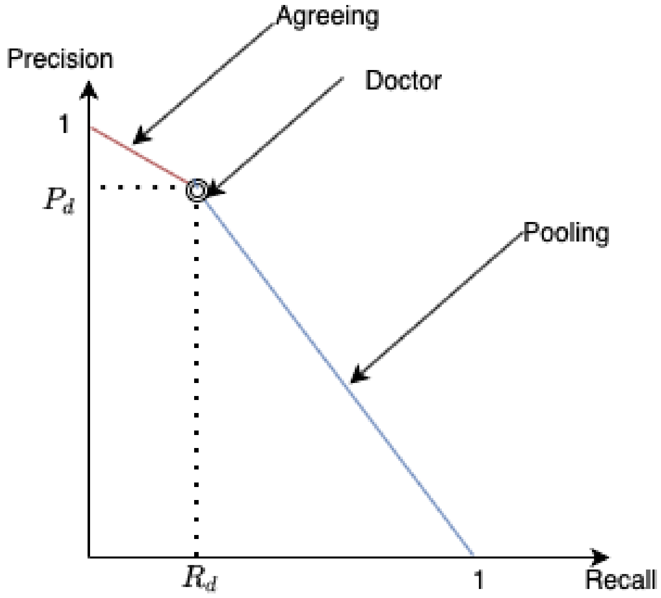


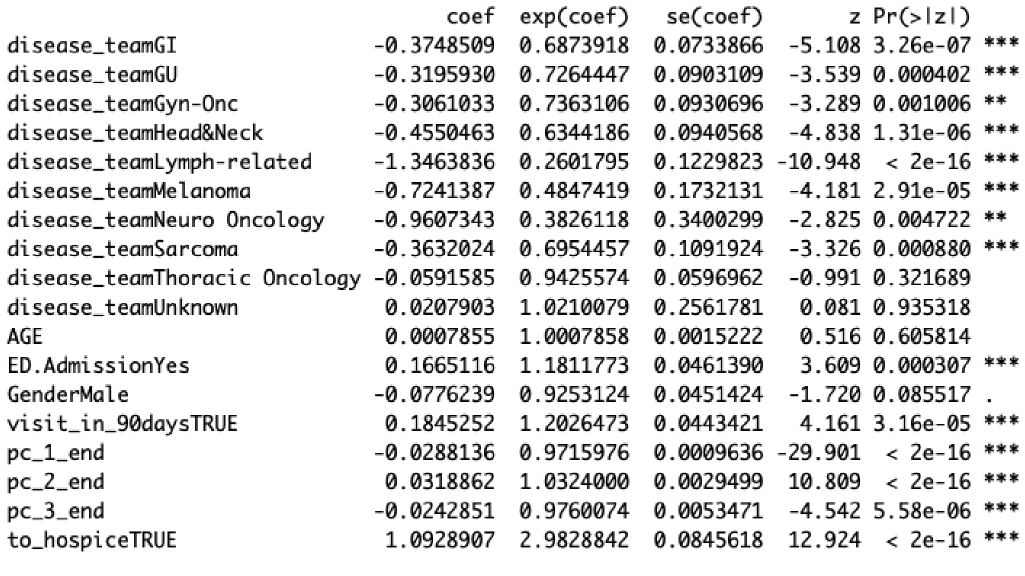
Figure 3.7: Demonstration of Precision-Recall Curve for Augmented Method.

3.4.4 Doctors' Discharge Decision as a new Covariate to IMPAC

From a different perspective, we can also add oncologists' discharge decision into the coxph model to provide IMPAC with additional information to see if it will help increase predictive power on survival. If we use $ToHospice$ to represent if the patient was discharged to hospice, the coxph model now becomes:

$$
\begin{aligned}
log(h(t)) = log(h_0(t)) &+ \beta_1 age + \beta_2 Gender + \beta_3 PC_1 + \beta_4 PC_2 + \beta_5 PC_3 + \\
&\beta_6 EntryThroughED + \beta_7 VisitIn90Days + \beta_8 Breast + \beta_9 Endocrine + \beta_{10} GI + \\
&\beta_{11} Genitourinary + \beta_{12} Gynecologic + \beta_{13} HeadAndNeck + \beta_{14} Melanoma + \\
&\beta_{15} Neurologic + \beta_{16} Sarcoma + \beta_{17} Thoratic \\
&+ \beta_{18} I_d
\end{aligned}
\tag{3.16}
$$

Since this method is essentially augmenting IMPAC with doctors' decision, we call this method Augmented IMPAC method. Figure 3.5 shows the coefficients and p values of new model parameters. We can tell that *to_hospice* has a positive(1.09) coefficient, which is significantly with p $< 2e - 16$. We used Aikaike Information Criteria(AIC) to dobackward variable selection: neither *to_hospice* nor any of the original variables was removed by the variable selection process. This indicates that doctors decision does add significantly more predictive power to the original IMPAC model.



```
                                 coef exp(coef)  se(coef)       z Pr(>|z|)    
disease_teamGI             -0.3748509 0.6873918 0.0733866  -5.108 3.26e-07 ***
disease_teamGU             -0.3195930 0.7264447 0.0903109  -3.539 0.000402 ***
disease_teamGyn-Onc        -0.3061033 0.7363106 0.0930696  -3.289 0.001006 ** 
disease_teamHead&Neck      -0.4550463 0.6344186 0.0940568  -4.838 1.31e-06 ***
disease_teamLymph-related  -1.3463836 0.2601795 0.1229823 -10.948  < 2e-16 ***
disease_teamMelanoma       -0.7241387 0.4847419 0.1732131  -4.181 2.91e-05 ***
disease_teamNeuro Oncology -0.9607343 0.3826118 0.3400299  -2.825 0.004722 ** 
disease_teamSarcoma        -0.3632024 0.6954457 0.1091924  -3.326 0.000880 ***
disease_teamThoracic Oncology -0.0591585 0.9425574 0.0596962  -0.991 0.321689    
disease_teamUnknown         0.0207903 1.0210079 0.2561781   0.081 0.935318    
AGE                         0.0007855 1.0007858 0.0015222   0.516 0.605814    
ED.AdmissionYes             0.1665116 1.1811773 0.0461390   3.609 0.000307 ***
GenderMale                 -0.0776239 0.9253124 0.0451424  -1.720 0.085517 .  
visit_in_90daysTRUE         0.1845252 1.2026473 0.0443421   4.161 3.16e-05 ***
pc_1_end                   -0.0288136 0.9715976 0.0009636 -29.901  < 2e-16 ***
pc_2_end                    0.0318862 1.0324000 0.0029499  10.809  < 2e-16 ***
pc_3_end                   -0.0242851 0.9760074 0.0053471  -4.542 5.58e-06 ***
to_hospiceTRUE              1.0928907 2.9828842 0.0845618  12.924  < 2e-16 ***
```

Figure 3.8: Coefficient and P-value for Augmented IMPAC

3.4.5 Evaluation Based on Precision Recall curve

We repeat the procedure in section 3.3 to create average precision recall curve for augmented method and augmented IMPAC(Figure 3.9).

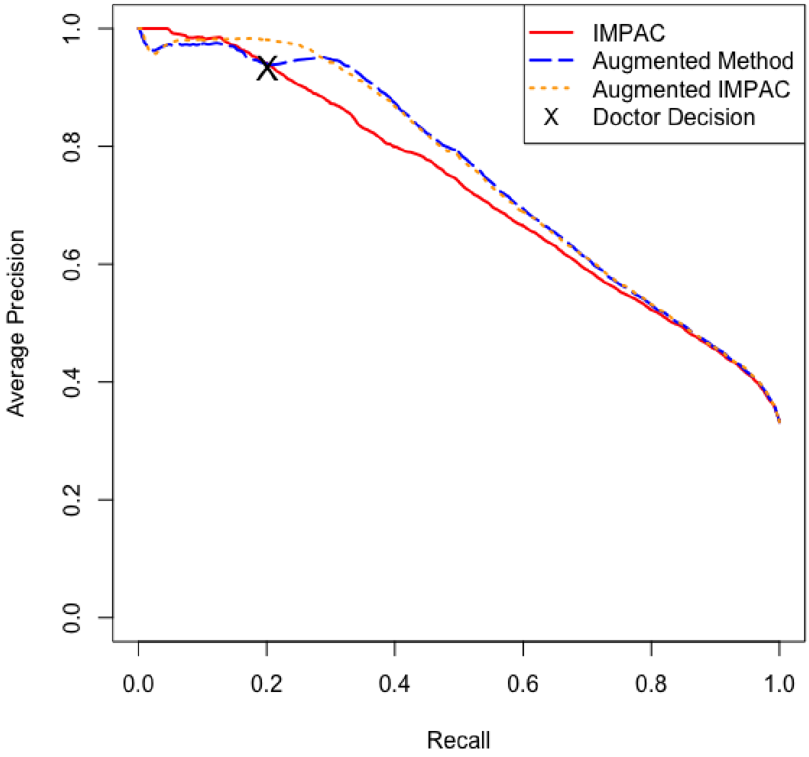


Figure 3.9: Average Precision-Recall Curves for 4 Classification Algorithms.

We refer a point on the Precision-Recall space (P_1, R_1) to be Pareto dominated by another point (P_2, R_2) if one of the following two conditions are satisfied

1. $P_1 \leq P_2$ and $R_1 < R_2$;

2. $P_1 < P_2$ and $R_1 \leq R_2$.

In Figure 3.9, there exists points on Precision-Recall curves of both Augmented Method and Augmented IMPAC that Pareto dominates doctors' decision.

As discussed in previous chapters, one major limitation of doctor's end-of-life decisions is being over conservative. This conservation may be caused by two factors:

- Doctors systematic tendency to overestimate survival time of patients;[36]

- Doctors intrinsic tendency to set a higher threshold to avoid false positives.

Assuming doctors are risk averse and their current precision level represent their bottom line, we want to calculate the most amount of positive cases that can be sent to hospice under this precision constraint. Equivalently, we need to solve the following optimization problem

$$R^* = \max_{\text{Precision} \geq P_d} \text{Recall} \tag{3.17}$$

Table 3.3 shows the maximum recall level achieved under the constraint that average precision does not exceed the doctor's current precision level($P_d = 0.93$). The last row gives the risk score cutoff τ picked that corresponds to Recall = R^*.

Compared to doctor as baseline($R_d = 0.23$, with the same precision level $P_d = 0.93$, Augmented method($R_a = 0.34$) and augmented IMPAC method($R_{AI} = 0.32$) can identify 45.5% and 37.7% positive cases respectively.

	IMPAC	Augmented Method	Augmented IMPAC	Doctor(Baseline)
R^*	0.22	0.34	0.32	0.23
Ratio	95.6%	145.5%	137.7%	100%
τ	0.64	0.65	0.67	-

Table 3.3: Maximum Achievable Recall with Guaranteed Precision

3.4.6 Survival time analysis

We have shown that under same precision, augmented method can achieve a significantly higher recall level. By definition, the true positives picked up by augmented method contains all true positives that were sent to hospice by doctors. It's worth exploring which part of patients, in terms of survival time, IMPAC helps doctor identify. We bin the patients who died within 90 days into 6 buckets with equal lengths and reported the percentage captured by doctors and augmented method within each bucket(Table 3.4). In the first row, percentage of each bucket that were sent to hospice was reported. In the second row, percentage of each bucket that were sent to hospice by augmented method that corresponds to (p_d, R^*) is reported. The third row reports the relative increment of percentage captured in each bucket.

Survival Time	(0,15]	(15,30]	(30,45]	(45,60]	(60, 75]	(75, 90]
Doctor	53.9%	13.8%	7.1%	5.6%	2.8%	4.0%
Augmented	61.9%	23.8%	15.5%	13.6%	11.0%	14.6%
Relative Increment	+14.8%	+72.5%	+118.3%	+142.0%	+192.8%	+265.0%

Table 3.4: Percentage Sent to Hospice by Doctor and Augmented Method, Binned by Survival Time

We can tell that doctors and IMPAC have comparative advantage in terms of identifying different groups of patients. Apparently, physicians identify a lot of positive cases that are close to death and perform badly at capturing positive cases who are close to 90 day survival time. Meanwhile, IMPAC will contribute a huge benefit in capturing positive cases who are relatively far from death.

Figure 3.9 shows the box plot of survival time distribution for True Positives and False Positives by setting precision and recall level at (P_d, R^*). We can tell that survival time of true positives by augmented are significantly bigger than those captured by doctors. The line in the middle shows the median of the distribution. The box in the Box Plot extends from the lower quartile to the upper quartile. The vertical lines, known as ”whiskers”, ranges from lower quartile - 1.5IQR to upper quartile + 1.5IQR.

3.5 Potential Avoidable Aggressive Treatments

So far, we have demonstrated that augmenting the doctor's end of life decision can increase recall by 45% while maintaining the doctor's previous precision level. Since the eventual goal of developing decision support system is to reduce aggressive treatments for patients close to death, we would like to quantify how many aggressive treatment can be correctly

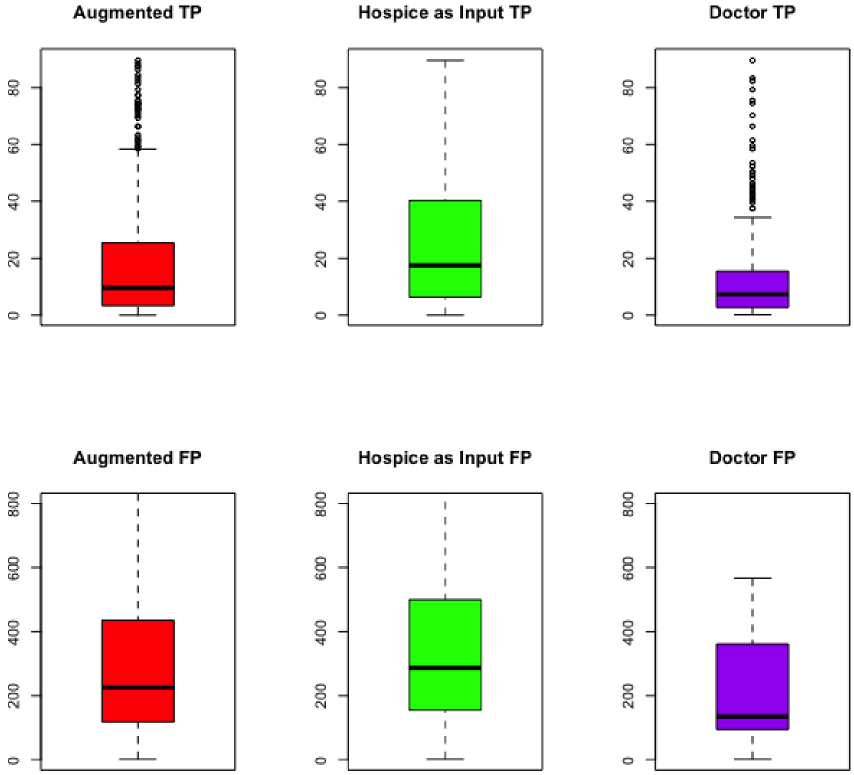


Figure 3.10: Survival Time Distribution in Box Plots

avoided by implementing augmented method.

Aggressive treatments for advanced cancer patients include chemotherapy, radiation, surgery, hospitalization and ICU admissions. Due to the limitations of our data, we only have access to the financial costs of a small subset of patients (n=669). This part of data indicates that end of life care does indeed involve many interventions that have little benefit. During the last 30 days of life, 27% of our patients were admitted to the ICU, with an average of 2.2 ICU admissions per patient. 38% presented to the ED, with an average of 2.3 ED presentations; 52% were admitted to an acute care inpatient service, with an average of 2.1 hospitalizations; 18% received chemotherapy; 13% received aggressive radiation; and 3% underwent an aggressive surgical procedure.

In section 2.2, we already did a financial analysis to estimate the potentially avoidable cost of treatment. To do that we selected one test set with the probability of death within

90 days threshold set at 50% and did a closer analysis of those patients whose financial cost data was available. For this set of 309 inpatient encounters, 103 resulted in death within 90 days, and IMPAC correctly identified 41 of those encounters (38 unique patients). For the 41 encounters, the first hospitalization for each patient was deemed the index hospitalization. Our estimation in chapter 2 was as follows. Had these 38 patients been treated under hospice care after the initial 48- hour assessment, it would have prevented 491 days of inpatient acute care. The estimated corresponding cost of hospice care would be $2,448 per patient. This implied a savings in this sample of $15,413 (95% CI, $9,162 to $21,665), that is, the potential avoidable cost ($17,861) minus the cost of hospice ($2,448).

We do not have detailed treatment and cost data for our full sample. Thus we cannot completely extend the above cost analysis to the augmented prediction context we examine here. Instead, we can quantify avoidable hospitalizations and lengths of hospital stay.

We randomly selected one test set with the probability of death within 90-days threshold set at 65% and did a closer analysis of what happened to the patients. Among the 1380 inpatient encounters, 379 (28.5%) actually died within 90 days from the discharge date of current inpatient visit. The patients in these 379 encounters would have been good candidates for hospice. However, only 97 (25.6%) were actually discharged to hospice. If IMPAC were used to augment the physician's decisions, the number correctly sent to hospice could have increased to 138 (36.5%). That is, IMPAC identified an additional 41 of positive cases to send to recommend to hospice.

From a different perspective, we can see that the average number of hospitalizations that can be avoided by implementing the augmented method. Among the 931 unique patients who had at least one inpatient encounter during their last 180 days of life, 281 of them were actually sent to hospice by the doctors before they died. If doctors had used IMPAC as their decision support, they would have sent 380 of them to hospice before they die, which is an increase of 35%.

Let $T_{d,i}$ be the date when patient i was discharged to hospice by the doctor and $T_{a,i}$ be the earliest when patient i could have been sent to hospice by the augmented method given $\tau = 0.65$. We know that $T_{d,i} \geq T_{a,i}$. Figure 3.11 shows the empirical cumulative distribution function of $T_{d,i}$(red) and $T_{a,i}$(blue).

We can treat the number of hospitalizations and corresponding length of stays(LOS) that happened during $(T_{a,i}, T_{d,i}]$ as a measure of potential savings of aggressive treatments. For any visit (LOS) that satisfy this condition, if the visit (LOS) happened within 90 days from patients death, we call this a right avoidance. If a visit (LOS) happened between 90 and 180 days from patient death, we call this a neutral avoidance. If the visit (LOS) happened before 180 days, we call this a problematic avoidance.

Among the 931 patients, augmented method saved 88 inpatient encounters that added up to 689 days of LOS. Among them, 61 encounters (476 days LOS)are right; 16 encounters (116 days LOS) are neutral; 11 encounters (97 days LOS) are problematic. Since 1209 encounters actually happened within 90 days from patient mortality, implementing IMPAC to assist physicians would avoid 5.0% of these encounters.

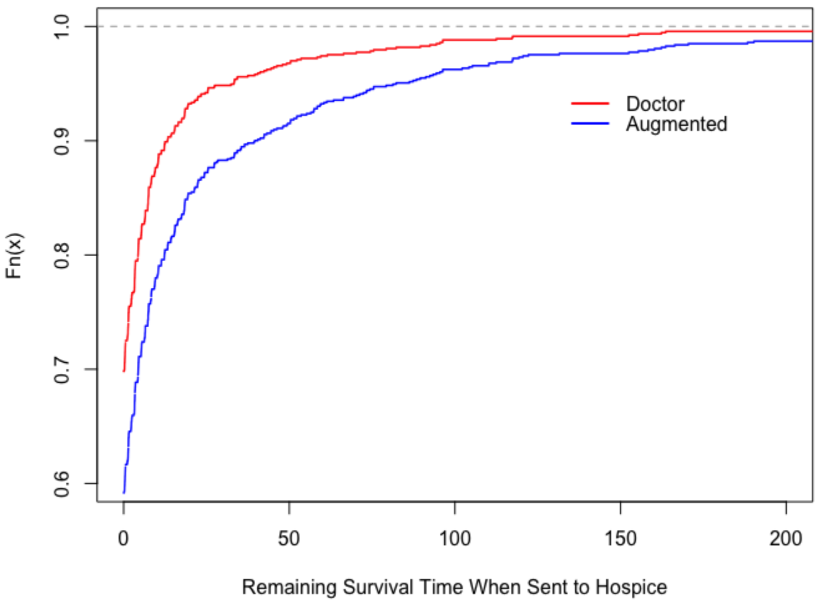


Figure 3.11: Empirical c.d.f. of Remaining Survival Time When Patient was Sent to Hospice for the First Time

3.6 Discussion

Although mortality predictors for end-of-life cancer patients have been widely studied in cohort studies, the literature mainly focused on improving machine learning models to achieve better prediction accuracy. Since the ultimate goal of these predictors is to serve as decision support tools, it's worth exploring how to incorporate these predictors into doctor's end-of-life decision process. Unfortunately, very few analysis have been done according to our literature review.

In this paper, we designed a retrospective study to compare discharge decisions made by doctors and IMPAC. We use the doctor's discharge decision as a proxy of what they predict the patient as dying in 90 days and showed that doctor's predictions are indeed conservative(precision=0.93,recall=0.23). Although at the same recall level, IMPAC has approximately the same average precision, we found that IMPAC and doctors identify significantly different group of patients who died within 90 days. The majority of true positives by doctors had a survival time below 30 days, while true positives by IMPAC were more widely spread across the [0,90] day interval. This motivated us two intuitive ways of augmenting doctor's decisions with IMPAC scores. Both methods identifies significantly more patients close to death at the same precision level as the doctor's currently achieve. In the last section, we estimated potential hospitalizations and hospital length of stays avoided if the doctors use augmented procedure instead of acting on their own beliefs. By augmenting doctor's decision with IMPAC, more than 5% of the current visits that happened within the last 90 days of patient life in our data could be avoided.

There are a few limitations in our analysis. First, we used the discharge disposition as a proxy of 90-day mortality prediction made by doctors. However, we can't observe cases when doctors initiated hospice care conversation but got declined by patient. Second, we made the assumption that discharge date can't be made earlier. Although this is a reasonable assumption since patients need to finish treatments in the current episode, in future studies it would be interesting to estimate the earliest time physicians pick up the hospice candidates and compare it with timeliness of IMPAC predictions.

Chapter 4

A POMDP Framework to Evaluate Decision Support System's Contribution to ICU patient Monitoring

4.1 Introduction

Intensive Care Units (ICUs), with the highest mortality rate reported (8%-19%) among all hospital units, account for around 20% of US hospital costs [49]. ICU staff carry the mission of closely monitoring patients who face life threatening health conditions. Timely detection of potentially fatal clinical conditions and intervention is crucial to improving ICU patient outcomes [50]. As a result, ICUs strive to maintain a very high nurse to patient ratio to guarantee continuous monitoring of patients to pick up any deterioration of patient condition before it becomes fatal. High nursing to patient ratios do not automatically reduce all risks to patients. These ratios are difficult to maintain consistently because of random fluctuations in the number of patients and supply of nursing. Also, nurses vary in their experience and skill at detecting patient distress.[51] Physicians are more scarce than nurses and must rely on the nurses to alert them of emergencies. There is therefore an interest in developing patient monitoring technology to assist nurses and physicians.

With the fast development and wide implementation of biomedical sensors, ICU patients' physiologic data can be collected and stored by the Electronic Medical Record (EMR) systems at high frequency. This provides both opportunities and challenges for improving ICU patients chance of survival. On one hand, larger volumes of clinical data contain more signals that can serve as early alerts of patients' deterioration for the physicians. On the other hand, such high dimensional, high frequency data is extremely hard for humans to process in a timely and consistent way to utilize in a way that would improve decisions.

Early patient monitoring scores like SAPS II (Le Gall et al. 1993), APACHE II (Knaus et al. 1985) and SOFA (Vincent et al. 1996) use manually selected commonly measured clinical predictors that are associated with a particular outcome that leads to patient mortality. These scores are only calculated once and can't be used as dynamic scores to keep track of patient health status.

A lot of recent research have focused on developing dynamic mortality prediction algorithms that can be updated in real time. With the help of these predictors, physicians can keep track of patient health status by observing only a single dynamic score. This is extremely helpful when the units are busy. Physicians can choose to only receive alerts when any patient's risk score jumps above a certain threshold.

In this sense, predicting mortality of ICU patients can improve the allocation of precious resources including physician time, medication and equipment. Most recent works have tried to implement a broader range of machine learning algorithms with main focus falling into one of two categories.

The first category tries to capture temporal trends of physiologic measurements. [52] used penalized spline fitting to extract temporal trend of physiologic data and showed that the shapes of the clinical data trajectories convey information about ICU mortality risk beyond what is already captured by the summary statistics currently used in the literature. [53] used a set of manually defined trend patterns to predict patient ICU mortality. [54] used long-short term short memory recurrent neural network to monitor ICU mortality risk.

Methods in the second category try to improve mortality predictions by adding features that are not used before. [55] developed a supervised learning approach for ICU mortality prediction based on unstructured electrocardiogram text reports. [56] improved ICU mortality prediction by integrating physiologic time series and clinical notes with deep neural net.

Despite diversified methodologies, the ultimate goal of these ICU mortality prediction algorithms is to build a decision support system that guides ICU physicians and nurses to save as many lives as possible given resource constraints on physicians' time and medical devices. So far, the literature on ICU mortality prediction uses discrimination and calibration to evaluate these prediction algorithms. Discrimination refers to the models ability to differentiate those patients facing high risks of mortality from those with low risks. Area under the receiver operating characteristic (ROC) curve has been widely used as a standard to compare discrimination across different models [57]. The ROC curve plots a classifier's true positive rate versus its false positive rate, corresponding to all possible classification thresholds. The area under ROC (AUROC), also called c-statistic, is equivalent to the probability that the model ranks a random positive sample more highly than a random negative sample. However, in the ICU setting the AUROC has two major limitations: (1) it tends to fail as a useful metric for discrimination with imbalanced outcomes; (2) false positive costs much less than a false negative. In this setting a false positive would cause nurses and physicians to check on a patient more frequently than necessary. A false negative could cause a patient in distress to not be checked on in time to save their life. Consequently, some argue that the area under the precision recall (AUPRC) is a better evaluation metric.

However, although it is realized that consistent monitoring of patients is challenging and that there are large disparities in the cost of false negatives vs false positives, it's still unclear how much a marginal improvement of prediction accuracy would translate to chance of survival per ICU stay, regardless of the prediction accuracy measured used. There hasn't been any research to build a framework to evaluate by how much the patient monitoring process will benefit from marginal increase of the AUROC or AUPRC. To study the potential benefit of these prediction algorithms we have to model their operations.

The predictions are providing information to decision makers. The decision makers here are the care providers staffing the ICU who must allocate their time to monitoring and caring for the patients on the unit. Medical devices like heart monitors are serving as warning systems to alert staff to when a patient is in distress. By the time the staff responds to such an alert the patient may be past the point of recovery. The purpose of a sophisticated prediction algorithm is to give an alert before the patient is in such distress that they cannot be saved. E.g. an alert anticipating an upcoming heart monitor alert. Operationally, such a signal can enable staff to spend more time closely monitoring the patients who are at higher risk of death. The implication is that for the same amount of staff time the patient mortality rate can be reduced.

To make the above more concrete we can consider the prediction algorithm for MICU patient mortality from Chapter 1. If we took a sample of 1000 six-hour blocks of individual patient time in the MICU, according to the data from the setting we studied, there would be on average 12 deaths (i.e. a 1.2 percent rate). An algorithm with 50 percent recall and 33 percent PPV would alert staff 18 times of which six of those times they would be checking on a patient that was going to die in the next six hours. In other words those 18 alerts would give the staff a chance to intervene in half of the cases. It would help them focus their attention on a much smaller group of patient-time intervals. From this example we can also see that the algorithm is not accurate enough to be a replacement for other monitoring efforts. But, it has potential to serve as a layer of security on top of the standard monitoring efforts. In the following we construct a theoretical model of exactly this situation to create a framework for exploring more precisely how an decision support system can augment the patient monitoring process in critical care.

There are of course additional limitations to prediction algorithms that need to be addressed. Because most mortality prediction models are black boxes for deaths of all causes, they are not able to directly give guidance of interventions. In this sense, we can assume that nurses are not able to use it as guidance. It's also possible that physicians could inspect a patient with extremely high risk scores but still not able to figure out what the potential cause of death would be. In this case, they might need to come back to the patient later to see if they are able to identify any issues. It is yet to be studied if the alerts provided by prediction systems can actually lead to interventions that will reduce mortality. Despite all these limitations, we believe that is only a matter of time before technology improves enough so that the actionable predictions are being provided and thus the framework and model presented here will become increasingly relevant.

In this chapter, we propose a Partially Observable MDP (POMDP) framework, with the physician as decision maker, that addresses above characteristics of mortality predictors. The rest of this chapter will be organized in as followed.

In section 2, we introduce the study setting, model assumptions and basic elements of our POMDP. In section 3, we solve a baseline POMDP where physicians make patient monitoring decisions based upon only their own observations. In section 4, we solve a more sophisticated POMDP with a dynamic mortality prediction algorithm providing decision support to physicians. In this model, the DSS will trigger an alert when the real time mortality risk exceeds a certain threshold. We will model two different approaches to how the physician uses the information provided by the DSS. In one approach the physician reacts to the alerts and in a sense allows the DSS to completely determine the monitoring decision. In the second approach the physician uses the alert from the DSS to update their own beliefs about the patient health status. In section 4, we use the modeling framework to analyze how different prediction algorithm accuracy levels can impact ICU mortality rates when overall monitoring effort is held constant.

4.2 POMDP Model Setup

We use a discrete time, infinite horizon partially observable Markov decision process (POMDP) to model individual patient ICU stays. A POMDP models an agent decision process in which it is assumed that the system dynamics are determined by an MDP, but the agent cannot directly observe the underlying state (Lovejoy 1991, Aoki 1965). Formally, a POMDP is a 7-tuple $(S, A, T, R, \Omega, O, \gamma)$, where

- S is a set of states;
- A is a set of actions;
- T is the conditional transition rules between states, given actions;
- R: $S \rightarrow R$ is the reward function;

- Ω is a set of observations;
- O is a set of conditional observation probabilities
- $\gamma \in [0, 1]$ is the discount factor.

Modeling the unit from the perspective of a single patient simplifies the exposition and computation. In the analysis of the model and numerical experiments we will hold constant the physician's rate of inspection to capture the fact that there are other patients in the unit. Thus our observations about a single patient will carry over to all patients. We also model the unit as having a single physician. This is often the case in medium and small ICUs. In large ICUs a physician will have a subset of patients they are responsible for. As a result we do not believe that the single physician assumption is very limiting.

4.2.1 ICU setting

The ICUs we study mainly has two types of staff who are responsible of monitoring patients:

1. **Physician(Intensivist)**: Each unit typically only carries one specialized intensivist who's competent at detecting a broad spectrum of conditions among critically ill patients and determining appropriate interventions. As part of daily routines, these physicians spend a part of their time inspecting patients to try to identify potential hazards and come up with treatment plans in a timely manner. In our model, we assume physicians can perform proactive inspections on patients ahead of time. After spotting a potential hazard, they are able to come up with a treatment plan that decreases the chance of critical conditions.

2. **ICU Nurses**: ICU nurses monitor patients, administer medications and responds to emergencies. ICUs usually have a high nurse to patient ratio. For example, California has legally required ICU nurse to patient ratio to be no lower than 1:2. However, compared to physicians, nurses are in general less trained and less capable of detecting possible life-threatening conditions. In our model, nurses are not explicitly modeled. They are assumed to be monitoring patients and implementing physician directed treatment plans.

ICU patients in our model switch among different states representing different hazard rate of death or severity of disease. There are three stages based on observability of incoming life-threatening conditions. We will give more details in Section 2.2. Physicians can instantaneously reveal the current state of each patient by inspecting them. However, after the inspection is finished, physicians are not able to keep track of the real state of patient before the next inspection. Instead, they form an unbiased belief of which state the patient might be at over time. Physicians need to determine an optimal patient inspection policy. Their objective is to minimize the expected number of patient deaths per inspection visit, or equivalently maximize the patient chance of survival per ICU stay.

We recognize that health conditions in ICU are also time dependent. To simplify the model, we don't track length of time patient is in the unit as a state variable. But that is to some degree captured in the Bayesian updating of b. As time goes by, belief b on the patient would go up and motivate the physicians to put some priority on the patients whom they have not inspected for longer time.

In this section, we will define a modified version of POMDP to model this process. This is a baseline model that does not include any decision support systems. In other words, there's no Ω and O included yet in the this baseline model.

4.2.2 State Space

States

We first define the true state space of the models. Based on patient's health condition, they switch among following true states during their ICU visits:

- **Stable(S)**: Stable patients have no human-foreseeable, life-threatening health issues. Physician (or nurse) inspections are not able to spot any near term hazard or come up with any change of treatment for this type of patients.

- **Pre-critical(P)**: Pre-critical is an intermediate state between Stable and Critical. Patients at this stage don't have any immediate life threatening conditions, but shows human-detectable physiologic precursors that indicates entrance into Critical State in the near future. State S and P belong to a partially observable state that can not be distinguished by ICU nurses. However, these precursors can be picked up by physician through inspections. In our model one distinction between nurses and physicians is that a physician can distinguish between the pre-critical state and stable state while a nurse cannot.

- **Critical(C)**: Observable state when life threatening health issues occurs and immediately triggers an alert to draw attention from ICU staffs. Once a patient enters this stage, a physiologic alert will immediately be picked by nurses. The nurses will inform the physicians, who immediately inspects the patient. The intervention at this stage is considered "last-minute", so the probability of treatment failing is higher than in the pre-critical state. Also if treatment fails at this stage it leads to immediate patient mortality within the same epoch.

A patient's ICU stay ends after hitting one of two terminal states:

- **Mortality(M)**: patient dies within ICU;
- **Discharged(D)**: patient leaves the ICU in healthy condition.

Partially Observable States and Belief Space

We define partially observable states from the physicians' perspective. Without careful inspection, ICU physicians are unable to differentiate Stable patients from Pre-Critical ones. So we define $PO = \{S, P\}$ to be partially observable state. Upon inspection, physicians instantaneously reveal the true state of the patient.

We assume patient's true state transitions based upon a MDP. Physicians understand how the patient state transition over time, and thus can maintain an unbiased belief on true states through Bayesian Update. When patient is in PO, we use $b \in [0, 1]$ to denote the patient's probability of being at state P.

Post Treatment State(PT)

To simplify notations, we define this post treatment state, which is essentially a probability distribution over belief space. A patient enters post treatment state in one of the following two scenarios:

- Patient enters Critical and does not die;
- Patient is in the Pre-Critical state and gets an inspection.

Once the patient enters PT, with probability q, the physician know that the patient is back to stable, so b is reset to 0. With probability $1-q$, physician is not certain about the patient post treatment state, and belief state becomes f. q captures how capable physicians are at telling if the treatment worked for certain patients right away. $1-(1-q)f$ represents the effectiveness of treatments.

4.2.3 Decision Variable

At the beginning of epoch, the physician needs to decide whether or not to inspect the patient. We use $a \in \{I, N\}$ to denote their actions, where I means the physician inspects the patient and N means the physician does not inspect the patient. A belief based policy π maps beliefs about the patient state at the time of making decisions to an action:

$$\pi : b \to a.$$

4.2.4 State Transition Rule

Critical

At the beginning of an epoch, a critical patient is immediately spotted by ICU nurses, who will trigger an immediate inspection and intervention by physicians. This is a realistic assumption since we are modeling the event of a patient being in severe enough distress to be obvious to nurses by direct observation of the patient or by a physiologic alert is triggered by a medical device. When a patient enters the critical state two things can happen:

- With probability m, the intervention fails and patient dies immediately,
- With probability $1-m$, the patient survives. Patient enters the aggregate state PT defined as follows: With probability q, the treatment works immediately, and the physician knows that the patient is back to Stable. In this case, $s = S$ and $b = 0$.

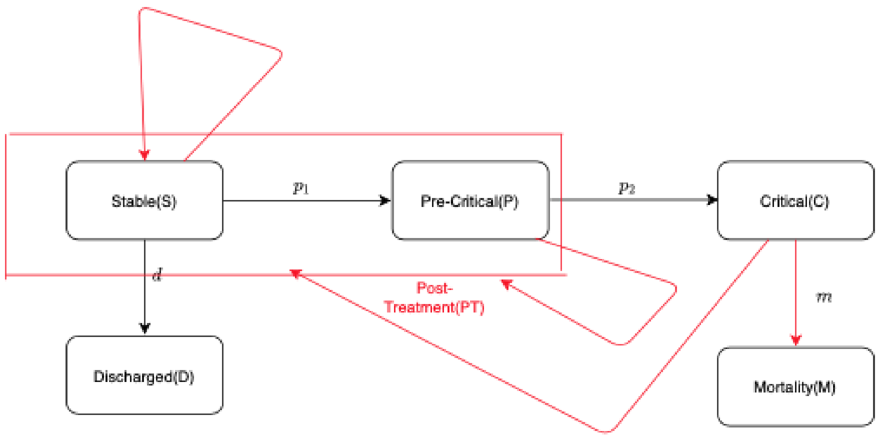


Figure 4.1: State Transition Diagram.

With probability $1 - q$, the physician is not sure about the patient post treatment state, and we assume the physician's belief becomes $b = f$.

Stable

Regardless of inspections, Stable patients turn Pre-critical with probability p_1, get discharged with probability d and stay Stable with probability $1 - p_1 - d$. We assume that those patients who are ready for discharge will be identified by nurses during routine inspections and won't require any additional inspections from the physician.

Figure 1 shows how true state transits from epoch to epoch. The lines in black show the natural transitions, without any inspections from physicians. The red line shows what happens once the physician inspects the patient.

Pre-Critical

Without intervention, Pre-Critical patients enter Critical at the end of the epoch with probability p_2 and stay Pre-Critical with probability $1 - p_2$. With inspection, which is assumed to be instantaneously completed at the beginning of epoch, the physician immediately gets full knowledge about the patient state. If the patient is Stable, the post-treatment (there is no treatment happening here but to make it consistent we use this term) belief state b

will be set back to 0. Otherwise, a treatment will be given to the Pre-Critical patient. Post inspection state becomes Stable with probability $q + (1 - q)(1 - f)$ and stay Pre-Critical with probability $(1 - q)f$. Post inspection belief state of a Pre-Critical patient becomes 0 with probability q and f with probability $1 - q$.

4.2.5 Objective

The objective of this POMDP model is to minimize mortality rate per ICU visit. The cost in our model should be measured by the number of deaths. There are two ways of incurring a cost during a patient's ICU stay:

- Discharge(Terminal) cost: If a patient is discharged through M, a terminal cost of $c_m = 1$ is incurred. If a patient is discharged as healthy, no terminal cost will be incurred.

- Inspection cost: Each physician inspection incurs a fixed cost of c_I, which represents opportunity costs of physicians efforts that could be otherwise used to save other patients in the same unit. Both proactive inspection (at partially observable states) and reactive inspection (at critical state) incur the same cost c_I.

The inspection cost c_I can be viewed as representing the degree to which the physician is time constrained. Later we remove this cost and place a constraint on the physician inspection rate.

4.3 Baseline Model

4.3.1 Event timeline

During an epoch in our baseline model, the physician first picks an action based on their initial belief state for the patient. If the physician does not inspect the patient ($a = N$), nothing happens to the real state and belief state during the epoch. Post-action, the belief state will have a Bayesian update. The patient's real state will go through the natural transition process during this epoch (Black lines in Figure 1).

If $A = I$, physician instantly reveals the true state of the patient. If the patient is revealed at Stable, post-action belief is set to be 0. and the Stable patient go through the natural transition during the epoch. Else, if the patient is revealed to be Pre-Critical, the physician will give a treatment. Post inspection, the true state becomes Stable with probability $q+(1-q)(1-f)$ and Pre-Critical with probability $(1-q)f$. Post inspection, the belief state of a Pre-Critical patient becomes 0 with probability q and f with probability $1-q$. Figure 2 shows how true state and belief state evolve during an epoch.

4.3.2 Bellman Equation for Baseline Model

First we introduce a few notations to help set up Bellman Equation:

- $J(b)$: cost to go function for any patient in belief state b;
- $Q(b,a)$: the cost to go function given belief b, immediate action a and implementation of optimal policy afterwards.

By definition, minimum cost to go function $J(b)$ satisfies

$$J(b) = \min(Q(b,N), Q(b,I)) \tag{4.1}$$

If a patient with belief state b is not given an inspection: (1) no cost of inspection is incurred; (2) with probability bp_2, patient is Pre-critical and will transit to Critical in the next epoch, cost to go becomes J_C; (3) with probability $(1-b)d$, is discharged, cost to go becomes 0; (4) with probability $b(1-p_2)+(1-b)(1-d)$, patient remains in partially observable states. Belief at the start of new epoch becomes $\frac{b(1-p_2)+(1-b)p_1}{b(1-p_2)+(1-b)(1-d)}$, cost to go is $J(\frac{b(1-p_2)+(1-b)p_1}{b(1-p_2)+(1-b)(1-d)})$. So

$$Q(b,N) = bp_2J_C + [b(1-p_2)+(1-b)(1-d)]J(\frac{b(1-p_2)+(1-b)p_1}{b(1-p_2)+(1-b)(1-d)}) \tag{4.2}$$

If a patient with belief state b is given an inspection: (1) a cost of inspection c_I is incurred; (2) with probability $b(1-q)$, the patient post inspection belief becomes f. Since physicians won't be able to inspection again, cost to go becomes Q(f,N). (3) with probability $1-b(1-q)$, the patient post inspection belief becomes 0. Since physicians won't be able to inspection again, cost to go becomes Q(0,N).

$$Q(b,I) = c_I + (1-b+bq)Q(0,N) + (b-bq)Q(f,N) \tag{4.3}$$

If a patient turns Critical, regardless of belief state, patient incurs an inspection with cost c_I. With probability m, patient dies and incurs cost $c_m = 1$. With probability $(1-m)q$, patient post treatment belief becomes 0, cost to go becomes $Q(0,N)$. With probability $(1-m)(1-q)$, post treatment belief becomes f, cost to go equals $Q(f,N)$.

$$J_C = c_I + m + (1-m)qQ(0,N) + (1-m)(1-q)Q(f,N) \tag{4.4}$$

4.3.3 Solving the Baseline POMDP

State Discretization

We start with discretizing belief state space. For each belief b, we round it to the nearest hundredth. Even though this will lead to a slightly different optimal policy, we believe that state discretization is good enough because in practice, rounding the belief to 0.01 is much more precise than what the physicians can keep track of and analyze. The continuous belief space $B = \{b : b \in [0,1]\}$ is now discretized into $B^{'} = \{0, 0.01, 0.02, .., 1\}$.

Policy Iteration

We use value iteration (Howard et al, 1960) to calculate $Q(b,a)$ for all $b \in B^{'}$ and $a \in \{I, A\}$.

Initialization

- For all $b \in \{0, 0.01, ..., 1\}$ and $a \in \{I, A\}, Q(b,a) = V(b) \leftarrow q_0, q_0$ is an arbitrary number;
- Stopping criteria: $\delta \leftarrow 10^{-5}$;

- t ← 0.

Value Iteration

```
While $t = 0$ or $\max_b |J_{t-1}(b) - J_t(b)| \geq \delta$:
[5em]2em For $b \in \{0, 0.01, 0.02, ...1\}$:
        $J_C \leftarrow c_I + m + (1-m)qQ(0, N) + (1-m)(1-q)Q(f, N)$
        $Q(b, I) \leftarrow c_I + (1 - b + bq)Q(0, N) + (b - bq)Q(f, N)$
        $J_{t+1}(b) \leftarrow \min(Q_{t+1}(b, N), Q_{t+1}(b, I))$
$t \leftarrow t + 1$
```

Value Function and Policy Function

$$J^*(b) = J_t(b), \text{for all b}$$

$$\pi^*(b) = argmax_{a \in \{I,N\}} Q(b, a)$$

4.3.4 Numerical Experiment

Our goal is to understand the potential real impact of short time horizon prediction algorithms on physician care for critically ill patients. To do that we need to use realistic parameter values in the analysis of the POMDP model. The key parameters of the model are (p_1, p_2, m, c_m, d). There have been no randomized rigorous studies observational studies of critical care units that we are aware of that could provide definitive values for these parameters even for a single specific ICU. However, based on a few metrics like Rothman

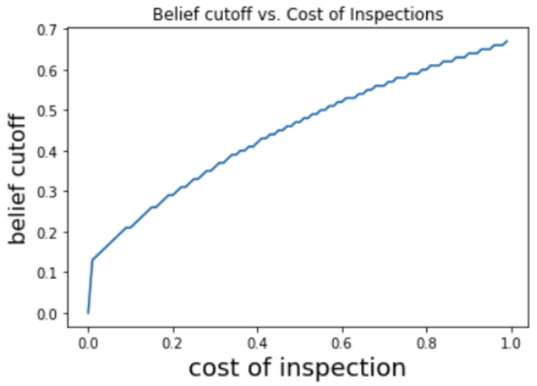


Figure 4.2: Belief Cutoff $\bar{b}$ vs Cost of Inspection C_I

Index, mortality rate and discharge rate, we are believe that it is possible to estimate values of these parameters that can be used to calculate performance measures of the right order of magnitude. Our approach to estimating these model parameters is described in detail in section 3.5. For the numerical examples in this chapter we will use the following values derived from the ICU described in Chapter 1: $(p_1, p_2, m, c_m, d) = (0.0027, 0.04, 0.2, 1, 0.0064)$. We do not claim that these are the best or only values for these parameters. Rather, we are only claiming that these values are of the right relative magnitude to be able to inform our understanding of how prediction algorithms can impact patient outcomes in an ICU.

Given this set of estimators, we can range cost of inspection c_I from 0 to 1 and solve for the optimal policy for each belief level. We have the following observations:

- **Observation 1**: Optimal policy for baseline model is a threshold policy. Given all model parameters, there exists $\bar{b}$, such that $A^* = I$ if and only if $b \geq \bar{b}$.
- **Observation 2**: J(b) is non-decreasing in b. Under optimal policy, the patient is better off when his/her chance of being Pre-critical gets lower.

We also ranged the model parameters and Observation 1 always hold. Observation 2 is demonstrated in Figure 3. As we can see, as the cost of inspection gets larger, physicians tend to only inspect patients with higher belief.

Next we vary $f \in \{0, 0.1, 0.2, 0.4, 0.5\}$ and fixed all other parameters and solve the baseline POMDP. Recall that f is the failure rate per inspection.

- **Observation 3**: $\bar{b}$ is non-decreasing in f.

Intuitively, an increase in f indicates decreased benefit from inspecting a patient in state P. When treatment becomes less effective, physicians will tend to inspect only those they believe have a higher likelihood of being in the pre-critical state.

Similarly, we vary m within range of $\{0, 0.1, 0.3, 0.5, 1\}$ and fix other parameters.

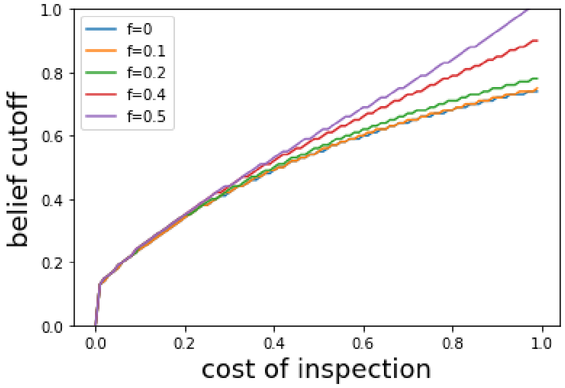


Figure 4.3: Belief Cutoff vs Cost of Inspections by Varying f

Observation 4: $\bar{b}$ is decreasing in m.

In our model, since the only difference between (1) the event in which a state P patient gets inspection and (2) a state C patient incurs inspection is that m proportion of (2) will die before going through the same process as (1). In this sense, m measures the cost of waiting until a state P patient turns Critical to inspect (Figure 5). With a fixed c_I, bigger m would make physicians tend to inspect more patients.

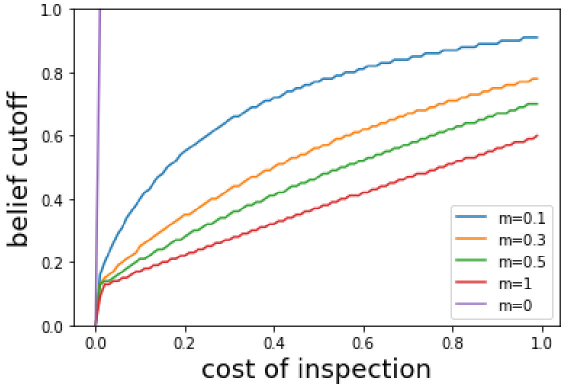


Figure 4.4: Belief cutoff vs Cost of Inspections by Varying m

Observation 5: $\bar{b}$ is decreasing in p_2.

p_2 represents the rate at which a patient transitions from Pre-Critical to Critical without any interventions. Intuitively, as p_2 goes up, the expected cost of physician not inspecting a Pre-critical patient in the current epoch increases because the patient will move to Critical more quickly. We plot the $\bar{b}$ vs. c_I for different levels of p_2 in Figure 5.

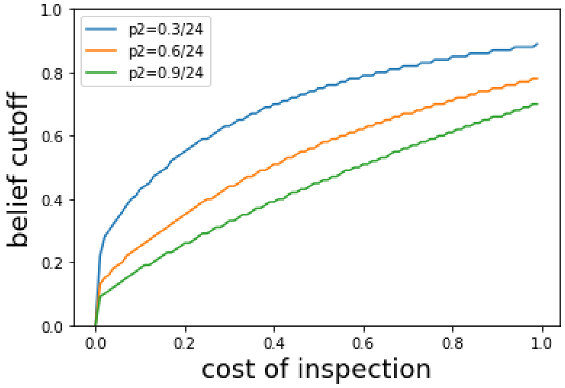


Figure 4.5: Belief Cutoff vs Cost of Inspections by Varying p_2

4.3.5 Estimation of Model Parameters

In this section, we describe how we derived estimates for a reasonable range of model parameters. The list of parameters included in the baseline model are:

- p_1: Probability a stable patient turns pre-critical in the next epoch;
- p_2: Probability a pre-critical patient turns critical in the next epoch;
- m: Probability of a critical patient dying within same epoch given last-minute intervention;
- c_m: Cost incurred per mortality, set to be 1;
- c_I: Opportunity cost incurred per physician inspection;
- d: Probability of discharge for stable patients;

To estimate model parameters, we used data in our previous work [52] based upon 4,557 unique patients admitted to the medical ICU (MICU) of the Yale-New Haven Hospital between January 2013 and January 20. A total of 5505 hospitalization episodes and 6113

MICU visits were recorded. Overall ICU mortality rate is 11% and average length of stay is 3.73 days. Yale-New Haven Hospital records a real time severity score, Rothman Index(RI), in its EMR system. The RI score is a composite measure updated regularly from the electronic medical record based on changes in 26 clinical measures including vital signs, nursing assessments, Braden score, cardiac rhythms, and laboratory test results. This score is independent of diagnosis, and it was developed to be used for any inpatient, i.e. medical or surgical patients including critical care patients. With a theoretical range from –91 to 100, the majority of patients on a general medical or surgical unit fall within the range from 0 to 100. The RI has previously been shown to have predictive power in forecasting 24-hour in-ICU mortality.

We will use the RI to classify patients into the Stable and Pre-critical states. To do that we take into account that the Stable state is defined so that a patient in that state can not transition to death and that we will use a one-hour duration for epochs. In our data set for each ICU visit, we take samples once every hour after admission, until the visit ends. This gives us 74,067 unique observation units. Figure 7 and Figure 8 shows the RI distributions for people who will not die within 1 hour and those who will die within 1 hour respectively. Since almost all patients who died within an hour had a below 30 Rothman Index, we use $RI > 30$ as a proxy of being in state S and $RI \leq 30$ as a proxy of state P. Under this criteria, 17% of these 1-hour samples are pre-Critical and 83% are Stable. We assume that the fraction of time spent in state C is negligible compared to the others by definition.

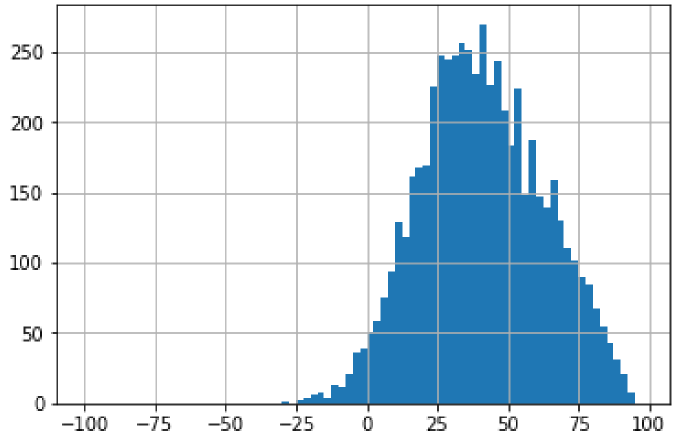


Figure 4.6: RI Distribution for Those not Dying in 1 Hour

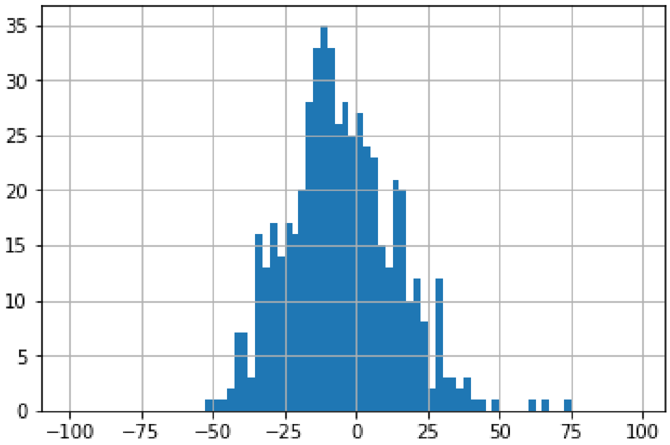


Figure 4.7: RI Distribution for Those Dying in 1 Hour

We also look at the initial Rothman Index for each ICU stay. Among all ICU encounters, 91% start with patients in Stable and 9% start with patients in Pre-Critical. We will use this number to set up initial belief state at admission in our simulation. Using these observations, and data on discharge rates and mortality rates we can estimate some of the key model parameters as follows:

$$p_1 \approx P(RI_t \leq 30|RI_{t-1} > 30) = 0.0027$$

$$p_2 \times m \approx P(\text{Dead at t}|RI_{t-1} \leq 30) = 0.045$$

$$d \approx P(\text{discharge in 1h}|S) = 0.0064$$

To estimate values for the parameters f and q we take a calibration approach. We use the baseline under the assumption that physicians inspect patients once a day on average. We then simulate the performance for different values of these parameters to identify a set for which the overall mortality rates and residence time in each state match the data. We find that by using $f = 0.5, q = 0.3$ and $m = 0.6$ the overall ICU mortality rate is very close to what we observe from MICU data set(11%).

4.4 Model with Binary Decision Support System

Now we add a decision support system (DSS) to assist ICU physicians on making their inspection decisions. This decision support system can be based on any quantity that has relatively strong correlation with patient state. We use the real time 24-hour ICU mortality predictor introduced in Chapter 1 as an example. At the beginning of each epoch(hour), the system automatically run the algorithm to generate a risk score $r_{24} \in [0, 1]$. Suppose we use a cutoff c such that an alert is triggered if and only if $r \geq c$.

We assume the decision support system generates binary signals at the beginning of epoch. This signal follows a distribution that only depends on the patients current state:

- Given $s = P$, the alert $(r >= c)$ is triggered with probability α and not triggered$(r < c)$ with probability $1 - \alpha$.
- Given $s = S$, the alert $(r >= c)$ is triggered with probability β and not triggered$(r < c)$ with probability $1 - \beta$.

The alert is valuable in the sense that it can help physician perform an additional layer of Bayesian update and potentially update the belief in the direction towards true state.

4.4.1 Policy

For a given set of (α, β), we will study two policies:

1. **Belief Based Policy(BBP)**. Physician update belief on patient type based on alert at the beginning of epoch. This post-alert belief will be used to decide whether the patient would be checked or not.
2. **Reactive Policy(RP)**. Physician checks the patient if and only if an alert is triggered. This does not require solving POMDP, but provides us a baseline model of how DSS work when physicians totally rely on decision support to inspect patients.

4.4.2 Event Timeline for Belief Base Policy Model

A Belief Based Policy is a mapping from post alert belief $b \in [0, 1]$ into action $a \in \{I, N\}$. For BBP model, there are three major stages during an epoch:

- **Alert** When a new epoch start, a patient has an initial true state and belief state. Then the alert would add a layer of Bayesian update of belief, in the following way: Suppose a patient has initial belief b. With probability $b\alpha + (1-b)\beta$, alert would be triggered and post-alert belief becomes $b_{PA} = \frac{b\alpha}{b\alpha+(1-b)\beta}$. With probability $b(1-\alpha) + (1-b)\beta$, alert will not be triggered and post-alert belief becomes $b_{PA} = \frac{b(1-\alpha)}{b(1-\alpha)+(1-b)(1-\beta)}$.

- **Action** BBP policy is a mapping from post alert belief space into action space. Physicians observe the post-alert belief of the patient, and based on BBP policy set in advance, picks the action. Action brings instantaneous transition of true state and belief state in the same way as described by Section 3.4.

- **Transition** We use $T(b) = \frac{b(1-p_2)+(1-b)p_1}{b(1-p_2)+(1-b)(1-d)}$ to denote the transition of belief state from post action till the end of epoch. The true state transition follows the black arrows given in Figure 1.

4.4.3 Bellman Equation for Belief Based Policy Model

We use b to denote belief state of the patient right before physician takes actions in each epoch. Given that the patients' pre-action belief state is b and physician inspects the patient, with probability $b(1-q)$, the patient's post treatment belief becomes f. With probability $1-b(1-q)$, the patient's post treatment belief becomes 0. After that, the patient will make transition as if the patient started with post-action belief and no inspection was given.

$$Q(b, I) = c_I + (1-b+bq)Q(0, N) + (b-bq)Q(f, N) \tag{4.5}$$

If the patient is at pre-Critical, with cost to go denoted by J_C, the patient will automatically incur an inspection cost c_I. With probability m, the patient will die with terminal cost 1. With probability 1-m, the patient will survive, with cost to go function denoted as

J_PT.

$$J_C = c_I + m + (1-m)J_{PT} \tag{4.6}$$

After treatment, patient belief becomes f with probability (1-q) and 0 with probability q. Then patient goes through the natural state transition based on figure 1. To simplify notations, we use

$$T(x) = \frac{x(1-p_2)+(1-x)p_1}{x(1-p_2)+(1-x)(1-d)} \tag{4.7}$$

to denote mapping from post-action belief to end-of-epoch belief.

Given that patient enters post-treatment state(PT), with probability $f(1-q)p_2$, patient will enter Critical again when epoch ends and cost to go equals J_C. With probability $(1-f)(1-q)d+qd$, patient gets discharged with cost to go 0.

With probability $(1-q)(1-fp_2-(1-f)d)$, the patient post treatment belief becomes f and after state transition, patient remains in state S or P. At the end of epoch, belief state becomes $T(f)$. After that:

- With probability $T(f)\alpha+(1-T(f))\beta$, this patient with belief $T(f)$ will trigger an alert. And the post-alert belief(or pre-action belief) becomes $\frac{T(f)\alpha}{T(f)\alpha+(1-T(f))\beta}$.
- With probability $T(f)(1-\alpha)+(1-T(f))(1-\beta)$, this patient with belief $T(f)$ will not trigger any alert. And the post-alert belief(or pre-action belief) becomes $\frac{T(f)(1-\alpha)}{T(f)(1-\alpha)+(1-T(f))(1-\beta)}$.

With probability $q(1-d)$, the patient's post treatment belief becomes 0 and stays in S or P at the end of epoch. Post transition belief becomes $T(0)$. After that:

- With probability $T(0)\alpha+(1-T(0))\beta$, this patient with belief $T(0)$ will trigger an alert. And the post-alert belief(or pre-action belief) becomes $\frac{T(0)\alpha}{T(0)\alpha+(1-T(0))\beta}$.
- With probability $T(0)(1-\alpha)+(1-T(0))(1-\beta)$, this patient with belief $T(0)$

will not trigger any alert. And the post-alert belief(or pre-action belief) becomes $\frac{T(0)(1-\alpha)}{T(0)(1-\alpha)+(1-T(0))(1-\beta)}$.

We define J_{PT} as cost to go right after treatment is done at Critical or Pre-Critical. J_{PT} can be written as:

$$\begin{aligned}
&J_{PT} = (1-q)(1-fp_2-(1-f)d)[T(f)\alpha+(1-T(f))\beta]J(\frac{T(f)\alpha}{T(f)\alpha+(1-T(f))\beta})+ \\
&q(1-d)[T(0)\alpha+(1-T(0))\beta]J(\frac{T(0)\alpha}{T(0)\alpha+(1-T(0))\beta})+ \\
&(1-q)(1-fp_2-(1-f)d)[T(f)(1-\alpha)+(1-T(f))(1-\beta)]J(\frac{T(f)(1-\alpha)}{T(f)(1-\alpha)+(1-T(f))(1-\beta)})+ \\
&q(1-d)[T(0)(1-\alpha)+(1-T(0))(1-\beta)]J(\frac{T(0)(1-\alpha)}{T(0)(1-\alpha)+(1-T(0))(1-\beta)})+f(1-q)p_2J_C
\end{aligned} \tag{4.8}$$

If a patient in belief state b does not get an inspection from the physician:

- With probability $d(1-b)$ patient gets discharged and cost to go equals 0;
- With probability bp_2 patient gets to critical and cost to go equals J_C;
- With probability $b(1-p_2)+(1-b)(1-d)$, patient stays in either S or P at the end of epoch and physician can't tell them apart. The belief will be updated to $T(b)$. Then either (1)an alert is triggered with a probability $T(b)\alpha+(1-T(b))\beta$ and belief will be updated to $\frac{T(b)\alpha}{T(b)\alpha+(1-T(b))\beta}$ or (2) alert is not triggered with probability $T(b)(1-\alpha)+(1-T(b))(1-\beta)$ and belief will be updated to $\frac{T(b)(1-\alpha)}{T(b)(1-\alpha)+(1-T(b))(1-\beta)}$.

$$\begin{aligned}
&Q(b,N) = bp_2[c_I+m+(1-m)J_{PT}]+ \\
&(b(1-p_2)+(1-b)(1-d))[T(b)\alpha+(1-T(b))\beta]J(\frac{T(b)\alpha}{T(b)\alpha+(1-T(b))\beta})+ \\
&(b(1-p_2)+(1-b)(1-d))[T(b)(1-\alpha)+(1-T(b))(1-\beta)]J(\frac{T(b)(1-\alpha)}{T(b)(1-\alpha)+(1-T(b))(1-\beta)})
\end{aligned} \tag{4.9}$$

Also by definition,

$$J(b) = \min(Q(b, N), Q(b, I)) \tag{4.10}$$

4.4.4 Policy Iteration

Like the baseline model, we discretize the belief state by rounding all beliefs to the nearing hundredth. We use the following policy iteration algorithm to solve the optimal policy for the discretized belief state MDP.

Initialization For all $b \in \{0, 0.01, ...1\}$ and $a \in \{I, N\}$ Initialize $Q(b, a)$ with arbitrary number q_0. $t \leftarrow 0$.

While $t = 0$ or $\max_s |V^t(s) - V^{t-1}(s)| \geq \delta$, do:

Value Iteration: For each $b \in \{0, 0.01, ..., 1\}$ and $a \in \{I, N\}$, update $J(b)$, J_C and J_{PT}, do

$$\begin{aligned}
J_C \longleftarrow & c_I + m + (1-m)J_{PT} \\
J_{PT} \longleftarrow & (1-q)(1-fp_2-(1-f)d)[T(f)\alpha + (1-T(f))\beta]J_t(\frac{T(f)\alpha}{T(f)\alpha + (1-T(f))\beta})+ \\
& q(1-d)[T(0)\alpha + (1-T(0))\beta]J_t(\frac{T(0)\alpha}{T(0)\alpha + (1-T(0))\beta})+ \\
& (1-q)(1-fp_2-(1-f)d)[T(f)(1-\alpha) + (1-T(f))(1-\beta)]J_t(\frac{T(f)(1-\alpha)}{T(f)(1-\alpha) + (1-T(f))(1-\beta)})+ \\
& q(1-d)[T(0)(1-\alpha) + (1-T(0))(1-\beta)]J_t(\frac{T(0)(1-\alpha)}{T(0)(1-\alpha) + (1-T(0))(1-\beta)}) + f(1-q)p_2J_C \\
Q_{t+1}(b, I) \longleftarrow & c_I + (1-b+bq)Q_t(0, N) + (b-bq)Q_t(f, N) \\
Q_{t+1}(b, N) \longleftarrow & bp_2[c_I + m + (1-m)J_{PT}]+ \\
& (b(1-p_2) + (1-b)(1-d))[T(b)\alpha + (1-T(b))\beta]J_t(\frac{T(b)\alpha}{T(b)\alpha + (1-T(b))\beta})+ \\
& (b(1-p_2) + (1-b)(1-d))[T(b)(1-\alpha) + (1-T(b))(1-\beta)]J_t(\frac{T(b)(1-\alpha)}{T(b)(1-\alpha) + (1-T(b))(1-\beta)}) \\
J_{t+1}(b) \longleftarrow & \min(Q_{t+1}(b, I), Q_{t+1}(b, N)) \\
t \leftarrow & t+1
\end{aligned} \tag{4.11}$$

After solving the optimal belief based policy, we have the following observations:

- **Observation 6** Like the baseline POMDP, BBP is a threshold policy, which means there exist a cutoff $\bar{b}$, such that $A^*(b) = I$ if and only if $b \geq \bar{b}$.

- **Observation 7** Given fixed β, belief cutoff $\bar{b}$ is increasing in α. (Figure 4.9)

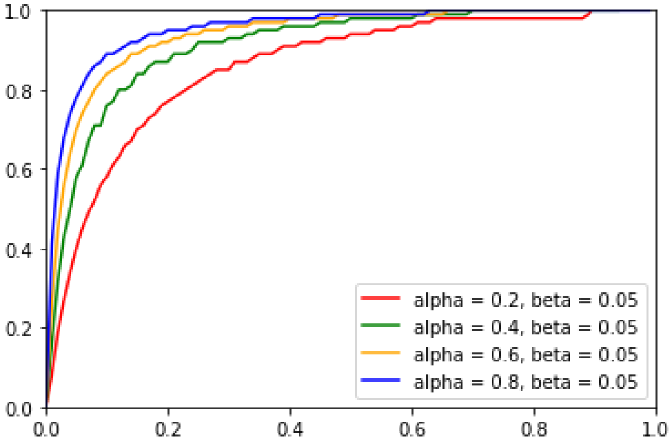


Figure 4.8: c_I vs. $\bar{b}$ Plot

- **Observation 8** Given fixed α, belief cutoff $\bar{b}$ is decreasing in β.

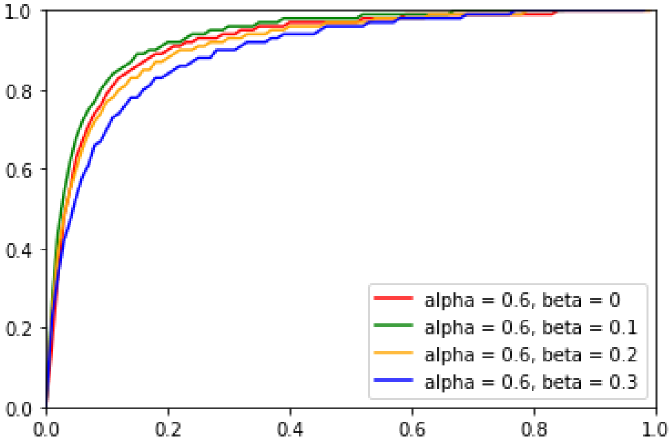


Figure 4.9: c_I vs. $\bar{b}$ Plot

4.4.5 Minimum Mortality Rate under Inspection Rate Constraints

As observed in previous sections, the optimal belief based policy is a threshold policy given estimated model parameters. We want to run simulation of an ICU in which physicians update belief on patient type based on the DSS model we described and inspect only patients with belief state higher than a pre-specified threshold $\bar{b}$.

For each $(\alpha, \beta, \bar{b})$ combination that satisfies $\alpha \in [0,1], \beta \in [0,\alpha], \bar{b} \in [0,1]$, we run simulation for 20000 ICU stays and calculate inspection rates (IR) per epoch and their corresponding average number of deaths per ICU visit (MR).

Figure 4.10 plots mortality rate versus inspection rate with α fixed at 0.9 and β varying. We can observe from it that under fixed α and β, lower threshold $\bar{b}$ leads to higher inspection rate and lower mortality rate. Figure 4.11 plots mortality rate versus inspection rate with α fixed at 0.9 and β varying. Given fixed α and inspection rate constraint, higher β leads to higher mortality rate.

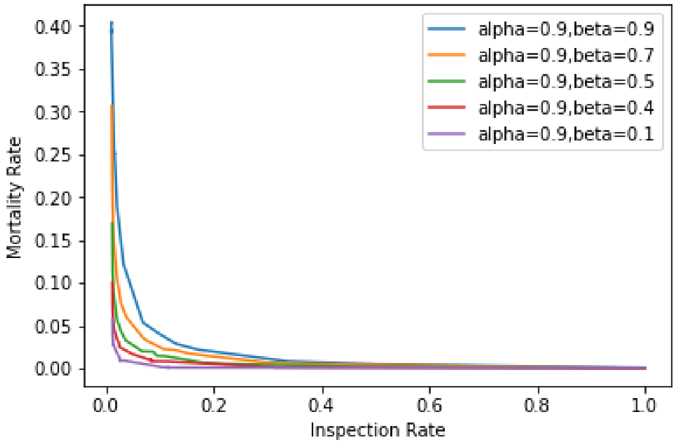


Figure 4.10: Inspection Rate vs Mortality Rate for combinations of α, β

Figure 4.11 and 4.12 plots the MR vs IR given a fixed α to β ratio. Given fixed α to β ratio and $\bar{b}$, higher α leads to lower mortality rate.

4.5 Minimize mortality under inspection rate constraint over ROC

Suppose we have a classification algorithm like the one we introduced in 2.4 to predict the probability of a patient dying in the following 24 hours. When an epoch starts, a machine learning model uses all available information of the patient stored in the EMR and outputs a risk score, r, between 0 and 1. If r is higher than a given threshold, c, then the model

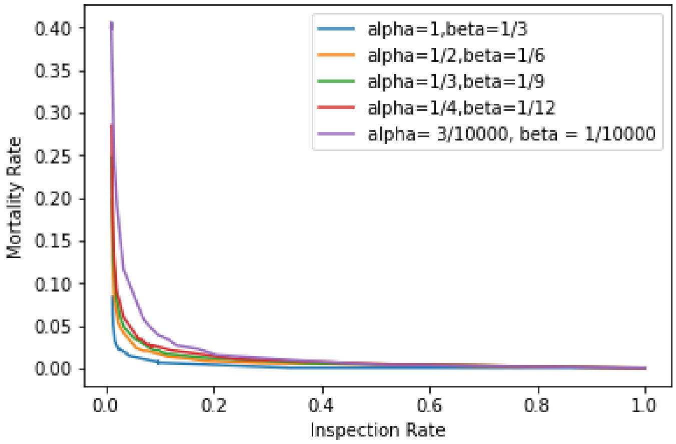


Figure 4.11: Inspection Rate vs Mortality Rate for Combinations of α, β

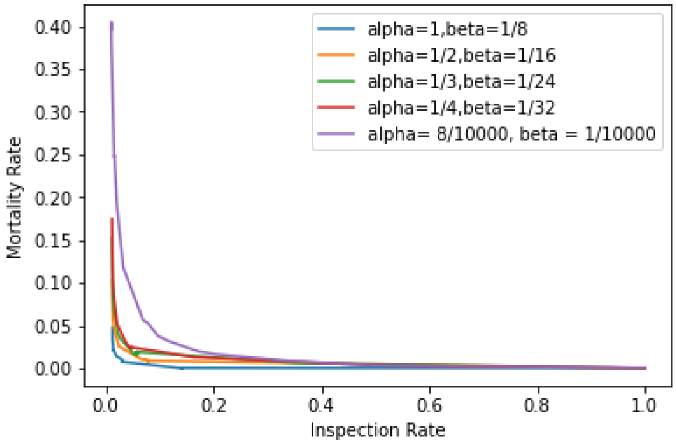


Figure 4.12: Inspection Rate vs Mortality Rate for Combinations of α, β

predicts that the patient is at risk of dying and sends out an alert.

4.5.1 Receiver Operating Characteristic(ROC)

ROC curve plots true positive rate(tpr) against false positive rate(fpr) corresponding to all cutoff $c \in [0, 1]$.

- True Positive Rate (tpr): the proportion of positives(patients who die in 24 hours) that correctly triggered alert;

$$tpr(c) = P(r \geq c|\text{Die in 24 hours}) \tag{4.12}$$

- False Positive Rate (fpr): the proportion of negatives(patients who survived 24 hours) that are incorrectly triggered alert.

$$fpr(c) = P(r \geq c|\text{Not Die in 24 hours}) \tag{4.13}$$

4.5.2 Get (α, β) plot based on ROC

In this section, we will build a bridge between ROC and Pareto frontier of (α, β) set.

By definition:

$$\alpha = P(alert|s = P)$$

$$\beta = P(alert|s = S)$$

$$\text{tpr} = P(alert|\text{die in 24h})$$

$$\text{fpr} = P(alert|\text{die in 24h})$$

$$\begin{aligned}
\text{tpr} &= P(alert|\text{die in 24h}) \\
&= \frac{P(alert, \text{die in 24h})}{P(\text{die in 24h})} \\
&= \frac{P(S)P(alert|S)P(\text{die in 24h}|S) + P(P)P(alert|P)P(\text{die in 24h}|P)}{P(\text{die in 24h}|S)P(S) + P(\text{die in 24h}|P)P(P)} \\
&= \frac{0.83\beta \times 0.004 + 0.17\alpha \times 0.05}{0.004 \times 0.83 + 0.05 \times 0.17} \\
&\approx 0.037\beta + 0.963\alpha
\end{aligned} \tag{4.14}$$

$$
\begin{aligned}
\text{fpr} &= P(alert|\text{not die in 24h}) \\
&= \frac{P(alert, \text{not die in 24h})}{P(\text{not die in 24h})} \\
&= \frac{P(S)P(alert|S)P(\text{not die in 24h}|S) + P(P)P(alert|P)P(\text{not die in 24h}|P)}{P(\text{not die in 24h}|S)P(S) + P(\text{not die in 24h}|P)P(P)} \\
&= \frac{0.83\beta \times 0.996 + 0.17\alpha \times 0.95}{0.996 \times 0.83 + 0.95 \times 0.17} \\
&\approx 0.84\beta + 0.16\alpha
\end{aligned}
\tag{4.15}
$$

By Equation 4.14 and 4.15, (α, β) can also be written as linear functions of tpr, fpr. We can also tell that true positive rate is highly correlated with α while false positive rate is more correlated with β.

4.5.3 Three ROC examples

In this section,we create the following three example ROC curves:

- The first ROC curve is a unit circle in the first quadrant that satisfies

$$tpr^2 + (1 - fpr)^2 = 1 \tag{4.16}$$

  and we can calculate $AUC = \frac{\pi}{4} = 0.78$.

- The second ROC curve is generated by decreasing the tpr in the unit circle ROC by 0.05. If the new tpr is less than 0, we will assign 0 to it. AUC ≈ 0.75

- The third ROC curve is generated by increasing the tpr in the unit circle ROC by 0.05. If the new tpr is bigger than 1, we will assign 1 to it. AUC ≈ 0.81

We use predictor 1 to represent the one with the lowest ROC(0.75), predictor 2 to represent the one with ROC in the middle(0.78), and predictor 3 to represent the one with the highest ROC(0.81). The ROCs and (α, β) frontiers are plotted in Figure 4.13 and Figure 4.14 respectively.

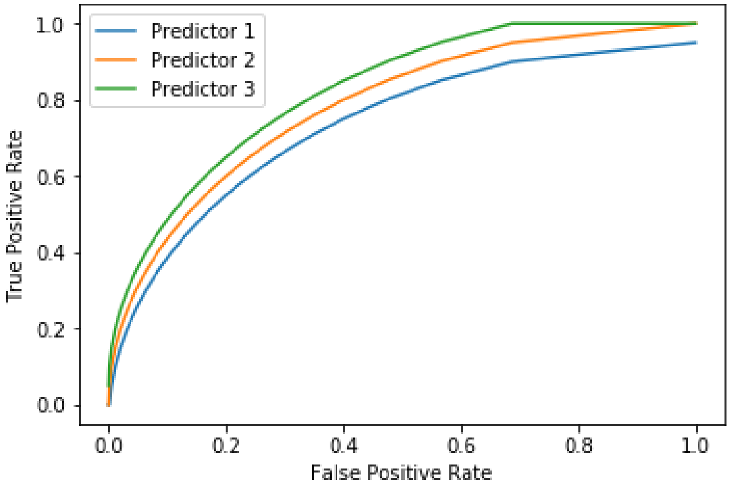


Figure 4.13: ROC Curves

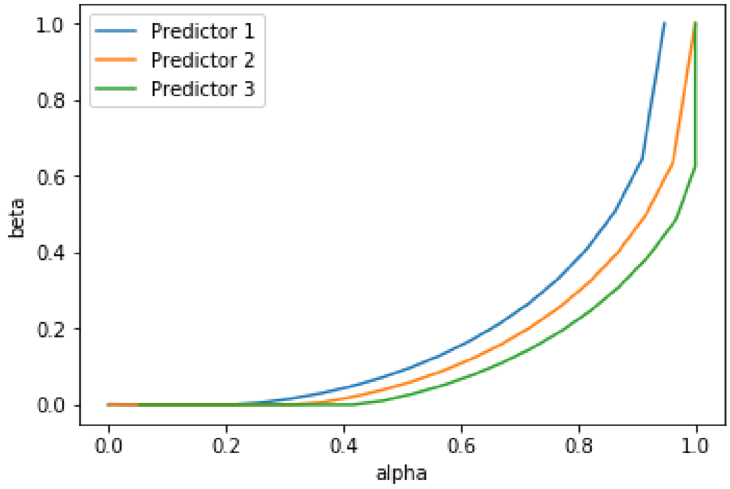


Figure 4.14: α, β Curves

4.5.4 Minimum Mortality Rate under Inspection Rate Constraints

With a fixed inspection rate constraint, we can compute minimum mortality rate corresponding to all (α, β) combinations. Based on a previous study [58] the patient-intensivist ratio (PIR) in ICUs in the United States ranges from 1.0-23.5 with IQR 6.9 - 10.8. This means more than half of the time each physician needs to monitor 6 to 11 patients. Also, most physicians inspect patients on a daily basis. Each epoch in our model represents one hour. If on average each patient gets inspection once per day, the average inspection rate hour is about 4.2%.

For each of the following policies, we are interested in the minimum mortality rate that can be achieved given certain inspection rates. We simulate the Belief-Based Policy (BBP) and the Alert Triggered Policy (ATP) with support from DSS based on Predictor 1, 2 and 3. For each (tpr, fpr) combination, we run 10,000 simulations of independent ICU visits. As tpr (sensitivity) goes up, physicians inspect patients at higher frequency, and mortality rate decrease. We can estimate the minimum mortality rate (MR) given inspection rate (IR) constraints $IR \leq 1\%, 2\%, 3\%, 4\%, 5\%$ and 6% respectively. We also list the minimum mortality rates corresponding to the same set of IR constraints for the Random Policy and BBP baseline (Table 1). We also plot the average number of times patient enter state C and State P per encounter(Figure 4.15 and 4.16).

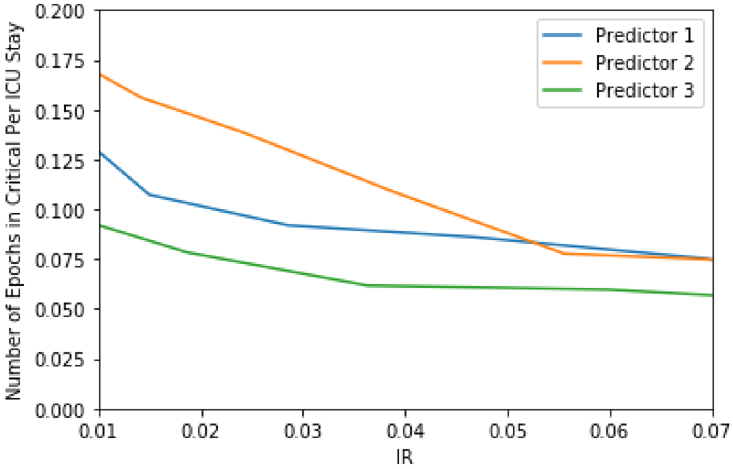


Figure 4.15: Average Number of Epochs in State C per ICU visit

There are a few interesting findings from the simulations. First, increased inspection rate

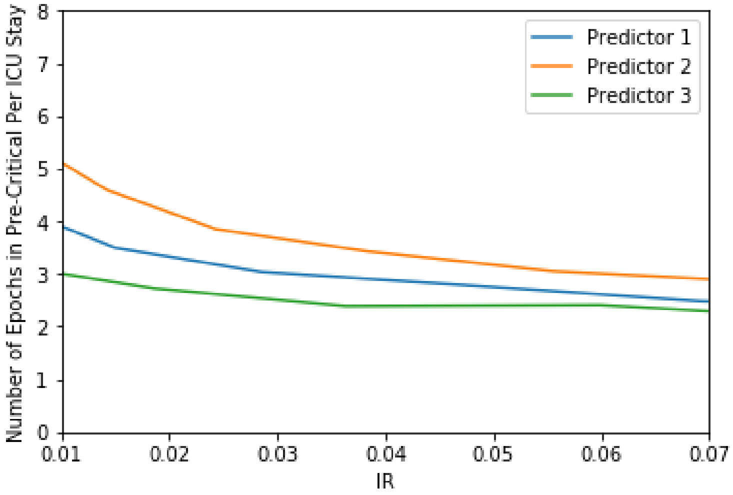


Figure 4.16: Average Number of Epochs in State P per ICU visit

lead to lower mortality in all policies. Second, given a fixed inspection rate, the mortality rate of different policies has the following order: BBP-DSS > ATP > BBP-Baseline > Random Policy.

Given a 4 percent inspection rate, triggered policies can avoid more than 50% of mortality compared to the BBP Baseline. Belief Based Policies can avoid more than 75% of mortality compared to the BBP Baseline. Third, at 4% inspection rate, an increase of AUROC by 0.1 can decrease mortality rate by 0.8%. As the inspection rate goes up, the benefit of increasing the AUROC goes down. At a lower inspection rate, higher accuracy can make a bigger difference. This indicates that accurate mortality predictors are more needed in relatively busy units where physicians are more time constrained.

In addition, the optimal true positive rates (sensitivity) for both ATP and BBP-DSS are in 0.3-0.5 range. Optimal true positive rates for both ATP and BBP-DSS increases as inspection rate goes up. This indicates that as physicians have more time to inspect patients, they tend to use lower risk score thresholds that correspond to higher sensitivity level. The alert based policy is much less efficient than the belief based policy.

	BBP1	BBP2	BBP3	ATP1	ATP2	ATP3	Random	BBP Baseline
MR^*,1% IR	5.2%	3.1%	2.9%	9.2%	7.9%	5.4%	45.3%	25.5%
tpr^*	0.2	0.3	0.3	0.3	0.3	0.35	-	-
fpr^*	0.032	0.032	0.032	0.062	0.062	0.046	-	-
MR^*,2% IR	2.9%	2.1%	1.7%	7.1%	5.5%	4.6%	40.2%	14.9%
tpr^*	0.25	0.35	0.4	0.35	0.35	0.4	-	-
fpr^*	0.062	0.062	0.062	0.083	0.062	0.062	-	-
MR^*,3% IR	2.4%	2.0%	1.5%	6.3%	4.9%	4.2%	36.5%	13.5%
tpr^*	0.35	0.40	0.45	0.4	0.4	0.4	-	-
fpr^*	0.083	0.083	0.083	0.107	0.083	0.062	-	-
MR^***, 4%, IR**	2.3%	1.8%	1.5%	5.8%	4.7%	4.0%	32.2%	10.9%
tpr^*	0.35	0.4	0.45	0.4	0.4	0.45	-	-
fpr^*	0.083	0.083	0.083	0.107	0.083	0.083	-	-
MR^*, 5% IR	2.1%	1.5%	1.4%	5.2%	4.5%	3.8%	28.1%	8.5%
tpr^*	0.35	0.4	0.5	0.4	0.45	0.45	-	-
fpr^*	0.083	0.083	0.107	0.107	0.107	0.083	-	-
MR^*, 6% IR	1.9%	1.4%	1.3%	4.6%	4.3%	3.7%	26.3%	6.8%
tpr^*	0.35	0.45	0.5	0.45	0.45	0.5	-	-
fpr^*	0.083	0.107	0.107	0.134	0.107	0.107	-	-

Table 4.1: Minimum Mortality Rate for Different Policies under Different Inspection Rate Constraint

4.6 Discussion

Despite interests in developing advanced ICU mortality prediction tools, there's been little literature talking about how these well these predictor can be used as early warning systems in practice. In this chapter, we proposed a POMDP framework to evaluate the potential benefits brought by these mortality predictions.

We come to the conclusion that more accurate mortality predictions can reduce ICU mortality rate by helping each patient's belief state to converge faster to the real state. With the help of mortality prediction algorithms, physicians are able to identify Pre-Critical patients sooner and thus keep more patients from entering Critical state.

We also found that the optimal tpr usually falls between 0.3 and 0.5. This indicates that a mortality prediction algorithm with higher AUROC is not equivalent to a better decision support tool. Instead, local performance in certain parts of ROC curves might be more important. The significant out-performance of the ATP policy by the BPP policy shows that finding ways to integrate a DSS signal into the decision making process of the physician may be more fruitful than seeking to replace the physician. This finding is inline with the results described in Chapter 3.

We recognize that the gap between mortality rate of BBP Baseline and BBP DSS has been over-estimated. There are two limitations in our model assumption that lead to this overly optimistic result.

First, the Bayesian update that incorporates both experience and alerts is difficult for physicians to do precisely in practice. But while in theory the alert-triggered policy appears easier to implement it will be very difficult to convince physicians to cede control to a black box algorithm.

Second, we are overestimating the extra information provided by DSS by assuming that alerts generated on the same patients are independent. However, consecutive predictions within a short time frame are usually highly correlated, due to the strong correlation between training data.

Bibliography

[1] J.-L. Vincent, A. De Mendonça, F. Cantraine, R. Moreno, J. Takala, P. M. Suter, C. L. Sprung, F. Colardyn, and S. Blecher. “Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study”. In: *Critical care medicine* 26.11 (1998), pp. 1793–1800.

[2] E. K. Adams, R. Houchens, G. E. Wright, and J. Robbins. “Predicting hospital choice for rural Medicare beneficiaries: the role of severity of illness.” In: *Health Services Research* 26.5 (1991), p. 583.

[3] H. Zucker, K. Adler, D. Berens, R. Bleich, R. Brynner, K. Butler, et al. “Ventilator allocation guidelines”. In: *Albany: New York State Department of Health Task Force on Life and the Law* (2015).

[4] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence. “APACHE-acute physiology and chronic health evaluation: a physiologically based classification system.” In: *Critical care medicine* 9.8 (1981), pp. 591–597.

[5] S. Lemeshow, D. Teres, J. Klar, J. S. Avrunin, S. H. Gehlbach, and J. Rapoport. “Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients”. In: *Jama* 270.20 (1993), pp. 2478–2486.

[6] J.-R. Le Gall, S. Lemeshow, and F. Saulnier. “A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study”. In: *Jama* 270.24 (1993), pp. 2957–2963.

[7] M. J. Rothman, S. I. Rothman, and J. Beals IV. “Development and validation of a continuous measure of patient condition using the electronic medical record”. In: *Journal of biomedical informatics* 46.5 (2013), pp. 837–848.

[8] G. D. Finlay, M. J. Rothman, and R. A. Smith. “Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system”. In: *Journal of hospital medicine* 9.2 (2014), pp. 116–119.

[9] D. M. Sow, J. Sun, A. Biem, J. Hu, M. Blount, and S. Ebadollahi. “Real-time analysis for short-term prognosis in intensive care”. In: *IBM Journal of Research and Development* 56.5 (2012), pp. 3–1.

[10] O. Badawi, X. Liu, E. Hassan, P. J. Amelung, and S. Swami. “Evaluation of ICU risk models adapted for use as continuous markers of severity of illness throughout the ICU stay”. In: *Critical care medicine* 46.3 (2018), pp. 361–367.

[11] F. L. Ferreira, D. P. Bota, A. Bross, C. Mélot, and J.-L. Vincent. “Serial evaluation of the SOFA score to predict outcome in critically ill patients”. In: *Jama* 286.14 (2001), pp. 1754–1758.

[12] A. L. Holder, E. Overton, P. Lyu, J. A. Kempker, S. Nemati, F. Razmi, G. S. Martin, T. G. Buchman, and D. J. Murphy. “Serial daily organ failure assessment beyond ICU day 5 does not independently add precision to ICU risk-of-death prediction”. In: *Critical care medicine* 45.12 (2017), p. 2014.

[13] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. “Scalable and accurate deep learning with electronic health records”. In: *NPJ Digital Medicine* 1.1 (2018), pp. 1–10.

[14] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. “Unfolding physiological state: Mortality modelling in intensive care units”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2014, pp. 75–84.

[15] J. Ramsay and B. Silverman. “Principal components analysis for functional data”. In: *Functional data analysis* (2005), pp. 147–172.

[16] E. Bradley, O. Yakusheva, L. I. Horwitz, H. Sipsma, and J. Fletcher. “Identifying patients at increased risk for unplanned readmission”. In: *Medical care* 51.9 (2013), p. 761.

[17] J. O. Ramsay. “Functional data analysis”. In: *Encyclopedia of Statistical Sciences* 4 (2004).

[18] M. Unser, A. Aldroubi, and M. Eden. “B-spline signal processing. I. Theory”. In: *IEEE transactions on signal processing* 41.2 (1993), pp. 821–833.

[19] F. O’Sullivan. “Nonparametric estimation of relative risk using splines and cross-validation”. In: *SIAM Journal on Scientific and Statistical Computing* 9.3 (1988), pp. 531–542.

[20] G. D. Costanzo. *Functional Principal Component Analysis of Financial Time Series.* 2005.

[21] M. A. Hael. “Modeling of rainfall variability using functional principal component method: a case study of Taiz region, Yemen”. In: *Modeling Earth Systems and Environment* 7.1 (2021), pp. 17–27.

[22] J. Peng and D. Paul. “A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data”. In: *Journal of Computational and Graphical Statistics* 18.4 (2009), pp. 995–1015.

[23] D. R. Cox. “Regression models and life-tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[24] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. “Random forest: a classification and regression tool for compound classification and QSAR modeling”. In: *Journal of chemical information and computer sciences* 43.6 (2003), pp. 1947–1958.

[25] W. Pan and R. Chappell. “Estimation in the Cox proportional hazards model with left-truncated and interval-censored data”. In: *Biometrics* 58.1 (2002), pp. 64–70.

[26] T. Therneau and T. Lumley. *R survival package.* 2013.

[27] L. Breiman. “Random forests”. In: *UC Berkeley TR567* (1999).

[28] A. Liaw, M. Wiener, et al. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[29] C. C. Earle, M. B. Landrum, J. M. Souza, B. A. Neville, J. C. Weeks, and J. Z. Ayanian. "Aggressiveness of cancer care near the end of life: is it a quality-of-care issue?" In: *Journal of clinical oncology* 26.23 (2008), p. 3860.

[30] J. Lynn, J. M. Teno, R. S. Phillips, A. W. Wu, N. Desbiens, J. Harrold, M. T. Claessens, N. Wenger, B. Kreling, and A. F. Connors Jr. "Perceptions by family members of the dying experience of older and seriously ill patients". In: *Annals of internal medicine* 126.2 (1997), pp. 97–106.

[31] J. M. Teno, P. L. Gozalo, J. P. Bynum, N. E. Leland, S. C. Miller, N. E. Morden, T. Scupp, D. C. Goodman, and V. Mor. "Change in end-of-life care for Medicare beneficiaries: site of death, place of care, and health care transitions in 2000, 2005, and 2009". In: *Jama* 309.5 (2013), pp. 470–477.

[32] A. B. Mariotto, K. Robin Yabroff, Y. Shao, E. J. Feuer, and M. L. Brown. "Projections of the cost of cancer care in the United States: 2010–2020". In: *Journal of the National Cancer Institute* 103.2 (2011), pp. 117–128.

[33] K. R. Yabroff, J. Lund, D. Kepka, and A. Mariotto. "Economic burden of cancer in the United States: estimates, projections, and future research". In: *Cancer Epidemiology and Prevention Biomarkers* 20.10 (2011), pp. 2006–2014.

[34] S. C. Kao, P. Butow, V. Bray, S. J. Clarke, and J. Vardy. "Patient and oncologist estimates of survival in advanced cancer patients". In: *Psycho-Oncology* 20.2 (2011), pp. 213–218.

[35] B. E. Kiely, A. J. Martin, M. H. Tattersall, A. K. Nowak, D. Goldstein, N. R. Wilcken, D. K. Wyld, E. A. Abdi, A. Glasgow, P. J. Beale, et al. "The median informs the message: accuracy of individualized scenarios for survival time based on oncologists' estimates". In: *Journal of clinical oncology* 31.28 (2013), pp. 3565–3571.

[36] N. A. Christakis, J. L. Smith, C. M. Parkes, and E. B. Lamont. "Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort

studyCommentary: Why do doctors overestimate? Commentary: Prognoses should be based on proved indices not intuition". In: *Bmj* 320.7233 (2000), pp. 469–473.

[37] W. J. Mackillop and C. F. Quirt. "Measuring the accuracy of prognostic judgments in oncology". In: *Journal of clinical epidemiology* 50.1 (1997), pp. 21–29.

[38] S. Subramaniam, A. Thorns, M. Ridout, T. Thirukkumaran, and T. R. Osborne. "Accuracy of prognosis prediction by PPI in hospice inpatients with cancer: a multi-centre prospective study". In: *BMJ supportive & palliative care* 3.3 (2013), pp. 324–329.

[39] N. de Paula Pantano, B. S. R. Paiva, D. Hui, and C. E. Paiva. "Validation of the Modified Glasgow Prognostic Score in advanced cancer patients receiving palliative care". In: *Journal of pain and symptom management* 51.2 (2016), pp. 270–277.

[40] K. J. Ramchandran, J. W. Shega, J. Von Roenn, M. Schumacher, E. Szmuilowicz, A. Rademaker, B. B. Weitner, P. D. Loftus, I. M. Chu, and S. Weitzman. "A predictive model to identify hospitalized cancer patients at risk for 30-day mortality based on admission criteria via the electronic medical record". In: *Cancer* 119.11 (2013), pp. 2074–2080.

[41] H. Bozdogan. "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions". In: *Psychometrika* 52.3 (1987), pp. 345–370.

[42] M. Stone. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 111–133.

[43] T. Saito and M. Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3 (2015), e0118432.

[44] B. Ozenne, F. Subtil, and D. Maucort-Boulch. "The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases". In: *Journal of clinical epidemiology* 68.8 (2015), pp. 855–859.

[45] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila. "Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients". In: *Critical care medicine* 34.5 (2006), pp. 1297–1310.

[46] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data". In: *BMC bioinformatics* 10.1 (2009), pp. 1–16.

[47] R. S. Pritchard, E. S. Fisher, J. M. Teno, S. M. Sharp, D. J. Reding, W. A. Knaus, J. E. Wennberg, J. Lynn, and S. Investigators. "Influence of patient preferences and local health system characteristics on the place of death". In: *Journal of the American geriatrics society* 46.10 (1998), pp. 1242–1250.

[48] N. E. Morden, C.-H. Chang, J. O. Jacobson, E. M. Berke, J. P. Bynum, K. M. Murray, and D. C. Goodman. "End-of-life care for Medicare beneficiaries with cancer is highly intensive overall and varies widely". In: *Health affairs* 31.4 (2012), pp. 786–796.

[49] D. A. Gruenberg, W. Shelton, S. L. Rose, A. E. Rutter, S. Socaris, and G. McGee. "Factors influencing length of stay in the intensive care unit". In: *American Journal of critical care* 15.5 (2006), pp. 502–509.

[50] S. L. Kane-Gill, E. E. Castelli, L. Kirisci, T. L. Rice, and M. P. Fink. "Effectiveness Study of rHuEPO in the ICU". In: *Crit Care Shock* 10 (2007), pp. 53–62.

[51] R. K. Amaravadi, J. B. Dimick, P. J. Pronovost, and P. A. Lipsett. "ICU nurse-to-patient ratio is associated with complications and resource use after esophagectomy". In: *Intensive care medicine* 26.12 (2000), pp. 1857–1862.

[52] J. Ma, D. K. Lee, M. E. Perkins, M. A. Pisani, and E. Pinker. "Using the shapes of clinical data Trajectories to predict mortality in ICUs". In: *Critical care explorations* 1.4 (2019).

[53] C. W. Hug and P. Szolovits. "ICU acuity: real-time models versus daily models". In: *AMIA annual symposium proceedings*. Vol. 2009. American Medical Informatics Association. 2009, p. 260.

[54] K. Yu, M. Zhang, T. Cui, and M. Hauskrecht. "Monitoring ICU mortality risk with a long short-term memory recurrent neural network". In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.* Vol. 25. World Scientific. 2020, pp. 103–114.

[55] G. S. Krishnan and S. S. Kamath. "A supervised learning approach for ICU mortality prediction based on unstructured electrocardiogram text reports". In: *International Conference on Applications of Natural Language to Information Systems.* Springer. 2018, pp. 126–134.

[56] S. N. Shukla and B. M. Marlin. "Integrating Physiological Time Series and Clinical Notes with Deep Learning for Improved ICU Mortality Prediction". In: *arXiv preprint arXiv:2003.11059* (2020).

[57] J. N. Mandrekar. "Receiver operating characteristic curve in diagnostic test assessment". In: *Journal of Thoracic Oncology* 5.9 (2010), pp. 1315–1316.

[58] H. B. Gershengorn, D. A. Harrison, A. Garland, M. E. Wilcox, K. M. Rowan, and H. Wunsch. "Association of intensive care unit patient-to-intensivist ratios with hospital mortality". In: *JAMA internal medicine* 177.3 (2017), pp. 388–396.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA