Predicting the Risk and Trajectory of Intensive Care Patients Using Survival Models

by

Caleb W. Hug

B.S., Whitworth College (2004)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2006

Author . . . .
Department of Electrical Engineering and Computer Science
August 30, 2006

Certified by . . . .
Peter Szolovits
Professor
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Predicting the Risk and Trajectory of Intensive Care Patients Using Survival Models

by

Caleb W. Hug

Submitted to the Department of Electrical Engineering and Computer Science
on August 30, 2006, in partial fulfillment of the
requirements for the degree of
Masters of Science in Computer Science and Engineering

Abstract

Using artificial intelligence to assist physicians in patient care has received sustained interest over the past several decades. Recently, with automated systems at most bedsides, the amount of patient information collected continues to increase, providing specific impetus for intelligent systems that can interpret this information. In fact, the large set of sensors and test results, often measured repeatedly over long periods of time, make it challenging for caregivers to quickly utilize all of the data for optimal patient treatment.

This research focuses on predicting the survival of ICU patients throughout their stay. Unlike traditional static mortality models, this survival prediction is explored as an indicator of patient state and trajectory. Using survival analysis techniques and machine learning, models are constructed that predict individual patient survival probabilities at fixed intervals in the future. These models seek to help physicians interpret the large amount of data available in order to provide optimal patient care.

We find that the survival predictions from our models are comparable to survival predictions using the SAPS score, but are available throughout the patient's ICU course instead of only at 24 hours after admission. Additionally, we demonstrate effective prediction of patient mortality over fixed windows in the future.

Thesis Supervisor: Peter Szolovits
Title: Professor

Acknowledgments

Many people are responsible for the successful completion of this thesis. First, I would like to thank my adviser Peter Szolovits. He has an amazing talent for understanding the exact struggles that I faced during this research; his constructive advise was invaluable. I am also grateful to all of the other MEDG members at the Computer Science and Artificial Intelligence Laboratory at MIT. Bill Long was a particularly helpful resource while I was working through issues with the data used in this project. Tom Lasko's medical expertise, which he graciously shared with me, was also invaluable as I conducted this research.

Several members of the Laboratory for Computational Physiology helped supply necessary data for this research. First, I would like to thank professor Roger Mark for his input. I am also indebted to him for the great amount of work that he has done in making the BRP project successful. I am grateful to Gari Clifford for answering many of my specific questions regarding the MIMIC II project. I would also like to thank Mauricio Villarroel for providing responsive support for database problems. Mohammed Saeed also provided specific advise at various stages of this research.

I have been blessed with a wonderfully supporting family. My mother and father were particularly helpful for conversations during times of frustration and homesickness. My siblings, Joshua (and his wife Celeste), Titus (and his wife Elizabeth), Levi, Elizabeth, Justin, and Miranda were also great. I have enjoyed my frequent conversations with each of them and their communication has made the transition to living on the East coast much easier. I would especially like to thank my oldest brother Joshua; he has been an amazing source of inspiration and encouragement throughout my education, pushing me towards excellence and freely sharing from his experience.

The friendships that I have cultivated here in Boston are invaluable. I would like to thank all of the members of the Eastgate Bible study. The prayer and support they have given me has been wonderful, and I am richly blessed to have each of them in my life. James Zimbardi and Kim and Ben Beaudoin have also made my time here in Boston quite rewarding.

I would be helplessly lost without my lovely wife Kendra. She is such an encouragement to me—cooking me wonderful food, providing needed distractions, and making sure that I don't take life too seriously. My life is immeasurably richer since she has joined me—she is a living testament to Proverbs 31:10-31.

Lastly, I want to thank my Heavenly Father. His grace has carried me from a young boy engulfed in "special programs" at my elementary school (for the *slower* children) to a nerdy college student excited about computer science; from a college sophomore nearly paralyzed by a tumor in my lower back, to where I am today. Because of His many blessings, my life is full of abundance. For this I am eternally grateful.

This thesis is based upon research supported by a National Library of Medicine Medical Informatics Traineeship, grant LM 07092.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Using artificial intelligence to assist physicians in patient care has received sustained interest over the past several decades. This chapter provides motivation for intelligent systems that predict patient outcome. First, it gives some general arguments for intelligent monitoring and then it discusses the case for mortality prediction. The chapter concludes with an outline for this thesis.

1.1.1 Intelligent Patient Monitoring

Recently, with automated systems at most bedsides, the amount of patient information collected continues to increase, providing specific impetus for intelligent systems that can interpret this information. In fact, the large set of sensors and test results, often measured repeatedly over long periods of time, make it challenging for caregivers to quickly utilize all of the data for optimal patient treatment. This problem is exacerbated by ineffective alarms, unreliable measurements, and lack of standards between alarm manufacturers. In a study by Tsien [62], as many as 86 percent of alarms were false positives and an additional 6 percent were clinically irrelevant. Other researchers have come to similar conclusions (i.e. [9], [8], and [40]).

More recent work in this area has shown improvement in these alarms. Zhang [65] indicates that newer alarms provide technical alerts, which are triggered by situations involving signal quality problems and equipment malfunctions. She found that by excluding these alerts, the false positive rates dropped to around 5 percent. One of the specific features that manufactures have used to verify that a signal is not being generated by a disconnected wire is a simple continuity test. Despite this progress, many problems remain.

These problems have a high cost. As Rothschild's study shows, serious medical errors are quite common in the intensive care unit (ICU) [48]. The ICU is especially susceptible to these errors due to the time sensitivity and the general complexity of the environment.

This project aims to reduce the complexity of this environment. The approach we take is to fuse the data from multiple sensors into an overall assessment of immediate risks for the patient and a prediction of the patient's trajectory. The high-level indication of a patient's risks from such an indicator could assist caregivers in interpreting the many individual alarms and possibly even remove some of them. Furthermore, the trajectory prediction could assist physicians in making preemptive care decisions. In summary, the integration of these sensors could allow caregivers to focus their needed attention to critical areas and help reduce costly errors in the ICU.

1.1.2 Mortality Prediction

Early warning for patient mortality would obviously be useful for physicians. This could potentially help physicians focus their attention on patients that have the most acute need of intervention. From an administration's point of view, reliable models would also help direct resources that are often dedicated to patients with effectively no chance of survival.

Misunderstandings of a patient's trajectory occur regularly in practice. For example, many patients are discharged from the ICU too soon. As Goldhill and Anne state,

> Many patients die after discharge from ICU and this mortality may be

> decreased by minimizing inappropriate early discharge to the ward, by the provision of high-dependency and step-down units, and by continuing advice and follow-up by the ICU team after the patient has been discharged [23].

In their study using a large group of patients from British ICUs, they found that 27 percent of the observed mortalities occurred after discharge from the ICU.

Existing models for predicting patient mortality have had limited success. Physicians continue to be much better at predicting final patient outcome than the various scoring metrics available [47]. One difficulty these scores encounter, as Goodhill and Anne also point out, is the contribution of limited resources to patient mortality [23]. In their study, postoperative patients, with low predicted mortality, were especially susceptible to premature discharge. They attribute much of this to a "widely perceived shortage" of ICU beds in the United Kingdom, and reference work that suggests that the average risk of death in the ICU is substantially higher in the United Kingdom than it is in the United States. Current mortality models are largely based on the typical pattern of care for similar patients. Premature discharge along with other care decisions (both good and bad) are unaccounted for by these mortality prediction models. This can make calibration difficult—simply looking at a patient's first 24 hours and his or her final outcome fails to include many important confunders.

Problems like these stem from models having a weak understanding of the underlying patient risk and trajectory. Ultimately, a simplistic score calculated by adding together the influences from different predictive variables falls short of a physician's more complex representation of the patient. This indicates the need for more powerful models that offer insight into specific patients rather than providing a statistical summary of "similar" patients. Even if the predictions from these metrics were comparable to estimates from physicians, their impersonal nature still might be a problem; in order for them to be useful in practice, their accuracy needs to be decisively better.

This research focuses on predicting the survival of ICU patients throughout their stay. Unlike traditional mortality models, this survival prediction is explored as an indicator of patient state and trajectory. Using survival analysis techniques and machine learning, models are constructed that predict individual patient survival probabilities at fixed intervals in the future. These models seek to help physicians interpret the large amounts of data available in order to provide optimal patient care.

1.2 Thesis Organization

This thesis is organized into the following chapters. First, Chapter 2 provides necessary background for the techniques used to construct survival models and to evaluate them. The Chapter starts with an introduction to survival analysis methods, then it discusses the SAPS mortality metric, and finally it describes various classification methods.

In Chapter 3, the details of the dataset preparation are explored. This includes the assumptions made while constructing the dataset and an overview of the final datasets used for modeling in this project.

Chapter 4 presents the feature selection methods used and the specific models created. These models are evaluated using a held-out test set and ROC curves. This chapter concludes with a discussion of these models and compares them with the SAPS I mortality metric.

Chapter 5 explores some work related to this project. Specifically, it provides an overview of the additional mortality scores available, as well as alternative survival analysis techniques. The chapter ends with a discussion of related intelligent monitoring work.

Finally, Chapter 6 summarizes the contributions of this work and offers direction for future research.

Chapter 2

Background

In this chapter we provide the background necessary to understand the rest of this thesis. With this intent we start by discussing the statistical techniques known as survival analysis. This includes a brief overview of the Kaplan-Meier estimator, the proportional hazards model and the accelerated lifetime model. Next, we discuss a common mortality metric known as the Simplified Acute Physiology Score (SAPS). We conclude the chapter by providing a high-level description of two classification algorithms used in this thesis.

2.1 Survival Analysis

Survival analysis refers to the analysis of time-to-event data from a known origin. These distributions are typically positively skewed, having a greater mean than median—in other words, they have a longer tail to the right of the median than to the left. Survival analysis was originally developed to predict patient survival, but the same techniques can be applied to any event that occurs within a certain time. In the following description of survival analysis, we use the term "length of survival" in this more general sense. A special characteristic of survival data is that they often contain censored instances.

2.1.1 Censored Data

Censored data provide the central motivation for survival analysis techniques. While there are multiple types of censoring, *right-censored* data is the most common. Other types of censoring include *left-censoring* and *interval censoring.* Each of these types of censoring is described below.

For a set of data that exhibits right-censoring, the final outcome for some subset of the instances is unknown. This commonly occurs, for example, in situations where a cancer treatment is observed for a fixed number of years, but at the end of the study some of the patients are still alive. Figure 2-1 illustrates right censoring. In the figure, each line represents a different case being observed. The solid circle indicates the start time, the open rectangle represents the death time and the open circle represents the time that the instance was right-censored.

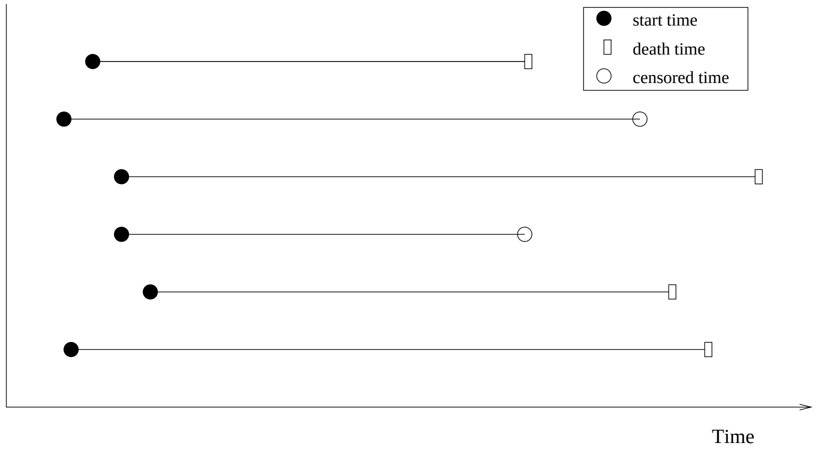


Figure 2-1: Illustration of Right-Censored Data

Data can also be *left-censored.* This less-frequent form of censoring occurs when the observed survival time is less than the actual survival time. This may occur, for example, if the exact date of birth is unavailable for a patient. By reversing the time axis this form of censoring can be treated similarly to the right-censored case.

Additionally, survival data can be *interval-censored.* This type of censoring applies to circumstances where the failure occurs within a fixed interval. For example, suppose that a patient was monitored for a particular event for two days, but the patient was discharged before the event occurs. If the patient is readmitted five days later and the event occurred sometime during the interim, then the length of survival is between two and seven days.

2.1.2 Definitions

It is helpful to start with the following standard survival analysis definitions. First, following the notation used by [56], let

$$Z_i = \min(Y_i, t_i)$$

where $Y_1, Y_2, ..., Y_n$ are the lengths of survival for each of the n observations and $t_1, t_2, ..., t_n$ are the right-censored times. Additionally, we define the censoring indicator, δ_i, as

$$\delta_i = \begin{cases} 1 & \text{if } Y_i \leq t_i \text{ (event observed),} \\ 0 & \text{if } Y_i > t_i \text{ (censored).} \end{cases}$$

Using this formalism, the ordered-pair (Z_i, δ_i) is observed for each instance in the data.

Survival Function

Obtaining an unbiased survival function is of primary concern in survival analysis. If a random variable T has the underlying probability density function $f(t)$, then the cumulative distribution function $F(t)$ can be obtained by simply integrating $f(t)$ from 0 to t. If this function represents the probability that the failure occurs between 0 and t, the desired survival function, $S(t)$, provides the probability that the failure occurs at or after t. Consequently,

$$\begin{aligned} S(t) &= P(Y \geq t) \\ &= 1 - F(t) \\ &= 1 - \int_0^t f(u)\, du. \end{aligned} \tag{2.1}$$

Hazard Function

Next, we define the hazard function of a patient. The hazard at time t for a patient is the patient's risk over a small time segment given that the patient survived until time t. Dividing this conditional probability by the small time segment gives a rate. Taking the limit of this rate results in the hazard function $h(t)$. Formally,

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \leq Y < t + \Delta t | Y \geq t)}{\Delta t} \right\}. \tag{2.2}$$

The hazard function and the survival function are closely related. In fact, performing some simple manipulation, Equation 2.2 can be written as

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.3}$$

Consequently

$$h(t) = -\frac{d}{dt}\{\log(S(t))\}, \tag{2.4}$$

and

$$S(t) = e^{-H(t)}, \tag{2.5}$$

where

$$H(t) = \int_0^t h(u)\, du. \tag{2.6}$$

The cumulative hazard function, $H(t)$, can also be defined in terms of the survival function as follows

$$H(t) = -\log S(t). \tag{2.7}$$

2.1.3 Kaplan-Meier Estimator

There are various parametric and non-parametric methods available for estimating the survival and hazard functions from the observed and censored survival times. The Kaplan-Meier estimate for the survival function is the simplest and most used of all of these methods.

The Kaplan-Meier estimator [28], also referred to as the product-limit estimator, provides a simple method for estimating the survival function from a set of right-censored data. While this estimator was used prior to 1958, Kaplan and Meier derived it using the maximum-likelihood method such that the function represents the distribution which maximizes the likelihood of the observed data.

In order to construct this estimator, the individual survival times Z_i are first sorted into order of increasing magnitude. Notationally, this is indicated by parenthesized subscripts. The following derivation, adapted from [56], starts by splitting the observed ordered-pairs into the following disjoint intervals

$$I_j = (Z_{(j-1)}, Z_{(j)}], \quad j = 1, 2, 3, ..., n, \quad \text{such that } Z_0 = 0.$$

Next, we define the risk set. If the uncensored survival times contain no ties and are ordered as

$$y_{(1)} < y_{(2)} < y_{(3)} < ... < y_{(k)},$$

we can define the **risk set** R at time u as

$$R(u) = \{\text{set of patients at risk prior to time } u\}. \tag{2.8}$$

In other words, the set of patients "at risk" is the set of patients still alive (and uncensored) prior to u.

Now, let $\hat{p_j}$ be the estimate that the patient survives through I_j given that they were alive at the beginning of I_j. Formally,

$$\hat{p_j} = 1 - \frac{\text{number dying in } I_j}{\text{number with the potential to die in } I_j},$$

for $j = 1, 2, 3, ..., n$. To simplify, let N_j represent the denominator in this equation.

We can now estimate the survival function as follows:

$$\begin{aligned}
\hat{S}(u) &= \prod_{j:Z_{(j)}\leq u} \hat{p_j} \\
&= \prod_{j:Z_{(j)}\leq u, \delta_{(j)}=1} (1 - \frac{1}{N_j}) \\
&= \prod_{j:Z_{(j)}\leq u} (1 - \frac{1}{N_j})^{\delta_{(j)}} \\
&= \prod_{j:Z_{(j)}\leq u} (1 - \frac{1}{n-j+1})^{\delta_{(j)}}.
\end{aligned}$$

The resulting estimator is given by Equation 2.9:

$$\hat{S}(u) = \prod_{j:Z_{(j)}\leq u} (\frac{n-j}{n-j+1})^{\delta_{(j)}}. \tag{2.9}$$

The above derivation relies on the assumption that there are no ties in the observed length of survival times. In other words, no two patients can have the same time recorded for the occurrence of the event of interest. In reality, with discrete times measured, this assumption might be broken. Fortunately, to account for these cases, Equation 2.9 can easily be generalized (see [56]).

The Kaplan-Meier estimate, $\hat{S}(u)$, has several noticeable properties. First, it is a piecewise constant function, meaning that the only changes occur at "jump points". These jump points are located at uncensored observations. Additionally, when the observations are not subject to censoring, the estimate becomes the expected empirical survival function, $S_n(u) = \frac{\text{number of observations} > u}{n}$.

2.1.4 Survival Models with Covariates

In addition to the Kaplan-Meier estimator, there are several alternative models that have been examined extensively for modeling survival data. Many of these models offer greater sophistication by using explanatory variables to customize the model for an individual instance. Parametric models are a common alternative to the non-parametric Kaplan-Meier estimator. These models typically use the method of maximum likelihood to fit a parametric distribution to the data. Some of the common distributions used are the Weibull, the exponential, the log-normal, and the log-logistic. Other models use a non-parametric model for the data that is similar to—and sometimes based on—the Kaplan-Meier estimator.

The two primary methods for dealing with covariates—that is, explanatory variables or features which are expected to influence the survival curve of interest—are the accelerated lifetimes model and proportional hazards model. Each of these methods allow the important features, such as age, to be incorporated into the model.

Accelerated Lifetimes

The accelerated-lifetime model is quite simple: the idea is that the covariates directly effect the length of survival, extending it or shrinking it accordingly. For example, the length of survival after the influence of the covariates can be written as

$$Y_{\mathbf{x}} = Y_{\mathbf{0}} a(\mathbf{x}), \tag{2.10}$$

where $Y_{\mathbf{0}}$ is the baseline length of survival, $\mathbf{x}$ is the vector of covariate observations, and the function $a()$ provides the accelerating factor and meets the requirement that $a(\mathbf{0}) = 1$.

Now, the accelerated-lifetime model appears when we calculate the probability that this new $Y_{\mathbf{x}}$ exceeds y

$$\begin{aligned} S_{\mathbf{x}}(y) &= P(Y_{\mathbf{0}} a(\mathbf{x}) > y) \\ &= P(Y_{\mathbf{0}} > \frac{y}{a(\mathbf{x})}) \\ &= S_{\mathbf{0}}(\frac{y}{a(\mathbf{x})}), \end{aligned} \tag{2.11}$$

where S_0 is the baseline survival function that represents the survival function for the case where $a(\mathbf{x}) = 1$. The function $a(\mathbf{x})$ is usually set to be $e^{\boldsymbol{\beta}^T \mathbf{x}}$, where $\boldsymbol{\beta}$ is the coefficient vector with the weights for the individual covariates. This results in the log-linear function,

$$Y_{\mathbf{x}} = Y_{\mathbf{0}} e^{\boldsymbol{\beta}^T \mathbf{x}}.$$

If the contribution from the covariates, $\boldsymbol{\beta}^T \mathbf{x}$, is positive, then the length of survival is increased relative to the baseline. Likewise, if the contribution is negative, then the length of survival is decreased relative to the baseline.

One of the nice consequences of this choice, on top of the fact that it satisfies the requirement that $a(\mathbf{0}) = 1$, is that it allows for easy calculation of the hazard function. By simply taking the derivative of 2.11 we obtain $h_{\mathbf{x}}(y) = h_0[\frac{y}{a(\mathbf{x})}]\frac{1}{a(\mathbf{x})}$.

Now by choosing a parametric form for baseline survival function, S_0, the model can be trained. In practice, the **Weibull distribution** is typically used for this. The Weibull probability distribution is defined as

$$f(y; \alpha, \beta) = \frac{\alpha}{\beta}\left(\frac{y}{\beta}\right)^{\alpha-1} e^{-(y/\beta)^{\alpha}},$$

which leads to the following survival function for this distribution:

$$S(y; \alpha, \beta) = e^{-(\frac{y}{\beta})^{\alpha}}, \quad \text{for } y > 0. \tag{2.12}$$

In order to account for covariates, the β parameter is typically replaced by some function $w(\mathbf{x})$. It is easy to show that accounting for covariates in this manner, using something such as

$$\beta = a(\mathbf{x}) = e^{\boldsymbol{\beta}^T \mathbf{x}},$$

does not violate the accelerated lifetimes assumption [56].

Other common distributions used with the accelerated lifetimes model include the exponential, the log-normal, or the log-logistic distributions.

Cox Proportional Hazards

Another more common way to handle covariates is to use the Cox proportional hazards model. In order to describe this model, we start with the more general proportional hazards model. The proportional hazards model is similar to the accelerated-lifetime model, except that the baseline hazard is scaled instead of the length of survival,

$$h_{\mathbf{x}}(y) = h_0(y)p(\mathbf{x}), \tag{2.13}$$

where p is a positive function of $\mathbf{x}$. Similar to the survival case, h_0 is the baseline hazard occurring when $p(\mathbf{x}) = 1$. Solving for p clearly illustrates the proportion created between the two hazards. As was the case for the $a()$ function in the accelerated-lifetimes model, the convenient $p(\mathbf{x}) = e^{\boldsymbol{\beta}^T\mathbf{x}}$ is typically chosen for p. By integrating equation 2.13, the we can arrive at the adjusted $S_{\mathbf{x}}(y)$:

$$S_{\mathbf{x}}(y) = [S_{\mathbf{0}}(y)]^{p(\mathbf{x})}. \tag{2.14}$$

This shows that the survival function resulting from the proportional hazards model is the baseline hazard raised to the power $p(\mathbf{x})$.

If we assume $Y \sim Weibull(\alpha, \beta)$, then the proportional hazards model has the same parametric survival function as Equation 2.12. As in the accelerated lifetimes case, the covariates are typically accounted for by specifying some function of $\mathbf{x}$ for the scale parameter β. Doing this does not violate the proportional hazards assumption. In fact, the Weibull distribution has the unique property that for length of survival data sampled from it the proportional hazards model and the accelerated lifetime model coincide. If a different distribution is used, then a number of diagnostic tests are available to help select the more appropriate model to use.

In 1972 Cox showed that the weights for the explanatory variables, $\boldsymbol{\beta}$, could be estimated independently from the baseline hazard function, $h_{\mathbf{0}}(u)$ [17]. This finding allows the estimation of the baseline hazard to be deferred until after the coefficient vector is estimated. The method is generally referred to as **Cox Proportional Hazards**. It is considered a semi-parametric technique because the baseline hazard is estimated using a non-parametric method (typically the Kaplan-Meier estimator) but the $\boldsymbol{\beta}$-parameters are still used to customize the model to specific covariate values.

Cox arrived at the following expression to represent the likelihood of the $\boldsymbol{\beta}$-

parameters, which is typically referred to as the **Cox partial likelihood**:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{e^{\boldsymbol{\beta}^T \mathbf{x}_{(j)}}}{\sum_{l \in R_j} e^{\boldsymbol{\beta}^T \mathbf{x}_l}}, \tag{2.15}$$

This partial likelihood is different from a traditional likelihood because the length of survival times are not used directly. Instead, it relies on the censored instances passively influencing the likelihood by falling out of the risk set in the denominator. Each of the terms in the product represents the probability that the individual with observations $\mathbf{x}_{(j)}$ at time $y_(j)$ dies given the risk set at the same time. Finding the maximum of this partial likelihood using typically maximum likelihood estimation methods yields the best $\boldsymbol{\beta}$ values for the model.

As with the Kaplan-Meier model, additional adjustments to Equation 2.15 are necessary to account for cases where multiple patients have the same length of survival. If the number of these ties is small, Breslow [4] suggests, as a good approximation, treating them as if they occur sequentially. This is the method for handling ties that is typically used in practice.

2.2 ICU Mortality Scoring Systems

A number of ICU Mortality scores have been introduced during the past 25 years. The models created in this project utilize the SAPS metric for comparison. The discussion of additional scoring metrics is deferred until Chapter 5.

2.2.1 SAPS

The Simplified Acute Physiology Score (SAPS) was introduced as a simpler and less time-consuming alternative to the Acute Physiology and Chronic Health Evaluation (APACHE) scoring system [33]. SAPS uses 14 easily measured variables and is able to yield very similar results to the more complicated APACHE scores. On the original 679 patients that the score was tested on, researchers demonstrated that the mortality rate consistently increased from a low SAPS of 4, with zero deaths, to a high SAPS greater than 20 with about 81 percent of the patients dying.

The SAPS metric is calibrated for use during the first 24 hours after ICU admission. For variables with multiple measurements during this period, the worst value is selected. One of the limitations of SAPS, as with the original APACHE score, is that they are used for separating patients into groups with similar probabilities of death, limiting its utility for predicting individual patient survival. The authors caution against use on specific patients. The predicted mortality obtained from these measures are useful for epidemiological purposes. The custom mortality risk prediction is for a specific patient population.

SAPS II [21] and Extended SAPS II [34] were developed using a much larger, international sample of patients. These scores directly provide a probability of hospital mortality. Using the SAPS II mortality prediction model, the creators found the area under the Receiver Operating Characteristics (ROC) curve to be about 0.86 on a held-out validation set. In most studies these scores have been found to perform slightly better than SAPS I. For example, one study on a large set of patients found the ROC area to be about 0.78 for SAPS I and 0.85 for SAPS II [7]. SAPS II, however, is less convenient for the current purposes because it requires specifying the chronic diseases that the patient has and the reason for the patient's admission. Other than this difference the two scores are quite similar.

2.3 Classification Models

There are numerous options available for classification algorithms. Logistic regression has been used on classification problems for many years and has some convenient characteristics. More recently, support vector machines (SVMs) have emerged as a very general classification framework capable of creating robust non-linear classifiers.

2.3.1 Logistic Regression

Logistic regression is a powerful regression technique for use with a binary response variable. It is similar in many respects to ordinary least squares regression, but uses a linear combination of the explanatory variables, $\mathbf{x}$, to model the log-odds (or "logit") transformation of the response variable y:

$$\log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x},$$

where x_0 is defined as 1 and $\boldsymbol{\beta}$ includes the intercept as β_0. Now, we can solve this equation for $P(y = 1|\mathbf{x})$. The resulting logistic regression model is shown in Equation 2.16:

$$P(y=1|\mathbf{x}) = \frac{1}{1+e^{-\boldsymbol{\beta}^T \mathbf{x}}}. \tag{2.16}$$

The parameters for a Logistic regression model, $\boldsymbol{\beta}$, are typically fit using maximum likelihood estimation. The primary benefit of this model is that the logit function has a sigmoid shape. This causes large values for continuous variables to be scaled into a bounded region between 0 and 1 and makes the model robust to large values of continuous variables.

2.3.2 Support Vector Machines

Support vector machines (SVMs) are routinely used for non-linear classification problems. Fundamentally, they rely on finding the maximum margin hyperplane—that is, the linear boundary between the two classes which maximizes the margins between the boundary and the two classes. By utilizing the "kernel trick", the feature space can be transformed so that these linear methods can be applied to non-linear classification problems.

A number of kernels are available for classification with an SVM. The radial basis function (RBF) kernel is quite common. This kernel requires two parameters: the cost, providing the penalty for miss-classified training instances, and the gamma parameter. These parameters need to be set, requiring some form of model selection. A simple way to find these parameters is to perform a grid search and use cross-validation. This is the technique employed for the SVMs used in this project.

After strong parameters have been found, the trained SVM model is generally quite robust due to the relative stability of the maximum margin separator [10]. Using a simple generalization, SVMs can also be extended for use with non-linear regression [38].

Chapter 3

Dataset Preparation

This chapter discusses the data used by this project. It starts by describing the data sources, specifically the MIMIC II project. Next, it describes the preprocessing used to clean the data and the various assumptions made for fusing the data into one synchronized dataset. Finally, it concludes with a summary of the final datasets.

3.1 Data Sources

Modern ICUs collect a wealth of patient information. The primary source of data for this project was the MIMIC II data. These data were collected as part of the Biomedical Research Partnership (BRP) between academia, industry, and clinical medicine [49]. As part of the BRP project, waveform data at 2 Hz and 1-minute trend data were collected from the monitors at several ICU beds. The MIMIC data also includes the deidentified nurse-recorded events and the administered medications for these patients. The final discharge states (alive or dead), the nursing notes, and the discharge summaries are available for a number of the patients. Each of these pieces of the MIMIC II data are described in this section.

3.1.1 High Resolution Data

The high-resolution waveform data, output by the bedside monitors, is valuable for exploring specific patient disorders. However, this project did not directly utilize this source for features. Instead, we relied on the one-minute averages of the high-resolution waveform data. The bedside monitors include this 1/60 Hz trend data as part of their output from the bedside monitors. The trend values for each signal are calculated by the monitor using a proprietary algorithm that averages the waveform values.

The set of trend signals available varies between patients. As a result, many patients do not have specific signals available. There are, however, a few signals that are widely monitored. These include the heart rate and the oxygen saturation (SpO_2). Invasive blood pressure readings are available for a large subset of the patients. Non-invasive blood pressure readings are also widely available, but only at infrequent intervals.

Another constraint with using the high resolution data is the number of patients available. While the general set of patients includes more than 17000 patients, at the time of this research only 2412 patients included high resolution data.

3.1.2 ISM Data

The primary source of data used for this project came from the Philips Medical Systems' CareVue Information Support Mart (ISM) database. This database is used to archive events during a patient's stay in the ICU. Most of the ISM data is event based. Most events are recorded on a rather low temporal resolution because the times attributed to them depend on when the nurse updates the patient's chart. Other events are recorded automatically, but are still generally subject to the frequency of the nursing rounds. For example, when an intravenous medication is administered it is automatically recorded but the machine recording the administration of these events has its own time that is not necessarily synchronized with the ISM time. The data entered during the nursing rounds usually consists of the standard vitals (such as blood pressure) which are generally recorded at least once per hour. For patients requiring more attention, the frequency of these updates increases.

The three tables of primary interest in the ISM database are the *ChartEvents* table, the *MedEvents* table, and the *IOEvents* table. The ChartEvents table contains a wide array of measurements and lab results related to the patient's condition. Many of the measurements, such as blood pressure and heart rate are quite frequent (i.e. once per hour), whereas many others, such as lab results, occur less frequently (i.e. once per day). The MedEvents table contains dosages, changes in dosages, start times, and end times for intravenous drugs. Finally, the IOEvents table provides the recorded fluid input and output events for the patient's stay. Using these numbers, the fluid I/O balance totals for the patient can be calculated for a given period of time.

3.1.3 Additional Data

In addition to the data outlined above, other data were needed for this research. Most importantly, the patient's status on discharge was needed for training and validating the survival models. This data indicates if the patient was alive or dead when he or she left the ICU. Because these data were not included in the ISM, it was necessary to retrieve them from the hospital. In addition to the outcome data, the ICD9 codes, nursing notes and discharge summaries were available for specific patients. These were useful for reconstructing a more complete view of a patient's progression through the ICU.

3.2 Data Preprocessing

Before merging the data from different sources into one large dataset, the data were preprocessed. This required several assumptions. This section provides an overview of these assumptions and briefly outlines the preprocessing steps used.

3.2.1 Trend Data Preprocessing

Several steps were taken in order to preprocess the trend data. For the specific details of this preprocessing, the reader is referred to Appendix A. A more general overview is provided below.

As in [61], derived features were calculated for many of the features. The slope was found by fitting a line to the points using standard least squares regression. Table 3.1 lists the features and the derived features from the trend data. In this table, the *Window* column provides the length of window used to calculate the slope (or mean) of the feature. The relation between the window and a given instance is graphically illustrated in Figure 3-1.

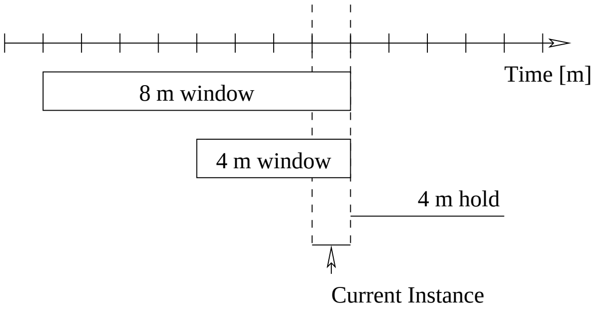


Figure 3-1: Derived Feature Window and Hold Window

Since the set of signals collected varies between patients (i.e. some patients don't have invasive blood pressure readings), it was necessary to require features to be present in order to include a given patient. The one feature listed in Table 3.1 that was not available for most patients was the central venous pressure (CVP). This is expected because the CVP reading requires a central venous catheter that is not necessary for most patients.

To start with, we applied a 3-minute median filter to each of the trend signals to remove noise. However, the trend data still had problems with noise and artifacts after applying this filter. Invalid values for blood pressure measurements were particularly troubling: many of the mean arterial pressure values did not lie between the systolic and the diastolic pressures for the same time instance. This is clearly impossible. Consequently, a number of rules were necessary to ensure the validity of blood pressure

Feature	Window [m]	History [m]
Heart Rate (HR) Mean	3	0, 15, 30, 45
HR Slope	15	0, 15, 30, 45
HR Slope	45	0, 45
HR Slope	120	0, 45
HR Standard Deviation	20	0, 15
Systolic Blood Pressure (ABPSys) Mean	3	0, 15, 30, 45
ABPSys Slope	15	0, 15
ABPSys Slope	45	0, 45
ABPSys Slope	120	0, 45
Diastolic Blood Pressure (ABPDias) Mean	3	0, 15, 30, 45
ABPDias Slope	15	0, 15
ABPDias Slope	45	0, 45
ABPDias Slope	120	0, 45
Mean Arterial Pressure (ABPMean) Mean	3	0, 15, 30, 45
ABPMean Slope	15	0, 15
ABPMean Slope	45	0, 45
ABPMean Slope	120	0, 45
Respiratory Rate (RESP) Mean	3	0, 15, 30, 45
RESP Slope	15	0, 15
RESP Slope	45	0, 45
RESP Slope	120	0, 45
Oxygen Saturation (SpO2) Mean	3	0, 15, 30, 45
SpO2 Slope	15	0, 15
SpO2 Slope	45	0, 45
SpO2 Slope	120	0, 45
Central Venous Pressure (CVP) Mean[a]	3	0, 15, 30, 45
CVP Slope	15	0, 15
CVP Slope	45	0, 45
CVP Slope	120	0, 45

Table 3.1: Vitals Measurements (Trend Data)

[a] Requiring the CVP signal shrinks the set of patients that include outcome information from 2037 down to 747. Because of this large reduction, the trend data were considered with and without this feature.

Systolic Blood Pressure > Mean Arterial Pressure
Systolic Blood Pressure > Diastolic Blood Pressure
Systolic Blood Pressure < 300 mm Hg
Mean Arterial Pressure > Diastolic Blood Pressure
Mean Arterial Pressure < 260 mm Hg
Diastolic Blood Pressure < 220 mm Hg

Table 3.2: Validation Rules

values. These rules are outlined in Table 3.2. If a rule does not pass for a given instance (row), then the features present in the rule are set to *NA* (missing).

3.2.2 ISM Data Preprocessing

The ISM data, in general, have less noise than the trend data. This is due to the direct human involvement in recording the measurements and filtering out clearly erroneous values. As with the trend data, however, there were cases where the diastolic blood pressure was recorded as greater than the systolic blood pressure. Consequently, it was necessary to apply the same validation rules that were used on the trend data.

The same procedure was used for calculating the slope for the ISM data as was used for the trend data. An additional challenge present in the ISM data, however, was sparsity. In order to get the windows between distinct entries to overlap, certain features were held for a fixed period of time. In other words, we assume that the last known measurement continues to be valid throughout a bounded window of time. Figure 3-1 illustrates a 4-minute hold on a particular instance. For “raw” features, this period is given under the *Window* column in Tables 3.3 and 3.4. For derived features, the *Window* column provides the length of the window used for calculating the feature.

In order to capture some of the dynamics of the data preceding a given instance, the previous values of various features at given times in the past were also included in each instance. The number of minutes for these shifts are included in the History column of each table. Multiple entries indicate unique features, with a shift of zero indicating the current value.

Tables 3.3, 3.4 and 3.5 show all of the features obtained from the ISM database.

Feature	Window [m]	History [m]
Systolic Blood Pressure (SBP)	5 hold	0
SBP Slope	60	0
SBP Slope	240	0
Diastolic Blood Pressure (DBP)	5 hold	0
DBP Slope	60	0
DBP Slope	240	0
Mean Arterial Pressure (MAP)	5 hold	0
MAP Slope	60	0
MAP Slope	240	0
Heart Rate (CV_HR)	5	0
CV_HR Slope	60	0
CV_HR Slope	240	0
Oxygen Saturation (SpO2)	5	0
SpO2 Slope	60	0
SpO2 Slope	240	0
Respiratory Rate (RESP)	5	0
RESP Slope	60	0
RESP Slope	240	0
Central Venous Pressure (CVP)	5	0
CVP Slope	60	0
CVP Slope	240	0

Table 3.3: ISM Vital Measurements

Feature	Window [m]	History [m]
Age	const	0
Weight	const	0
Sex	const	0
Censor Indicator	1	0
Service Type	720	0, 120
Survival Time	1	0
Input Total	720	0, 60, 120
Output Total	720	0, 60, 120
Arterial Oxygen Saturation (SaO2)	720	0, 120
Arterial pH (Art_pH)	720	0, 120
Arterial Base Excess (Art_BE)	720	0, 120
Carbon Dioxide (Art_CO2)	720	0, 120
Partial Pressure Carbon Dioxide (Art_PaCO2)	720	0, 120
Partial Pressure Oxygen (Art_PaO2)	720	0, 120

Table 3.4: Additional ISM Features

Medication	History [m]
Aggrastat	0, 60
Amicar	0, 60
Aminophylline	0, 60
Amiodarone	0, 60
Amrinone	0, 60
Argatroban	0, 60
Ativan	0, 60
Atracurium	0, 60
Cisatracurium	0, 60
Dilaudid	0, 60
Diltiazem	0, 60
Dobutamine	0, 60
Dopamine	0, 60
Doxacurium	0, 60
Epinephrine	0, 60
Epinephrine-k	0, 60
Esmolol	0, 60
Fentanyl	0, 60
Fentanyl (Conc)	0, 60
Heparin	0, 60
Insulin	0, 60
Integrelin	0, 60
Labetolol	0, 60
Lasix	0, 60
Lepirudin	0, 60
Levophed	0, 60
Levophed-k	0, 60
Lidocaine	0, 60
Midazolam	0, 60
Milrinone	0, 60
Morphine Sulfate	0, 60
Natrecor	0, 60
Neosynephrine	0, 60
Neosynephrine-k	0, 60
Nicardipine	0, 60
Nitroglycerine	0, 60
Nitroglycerine-k	0, 60
Nitroprusside	0, 60
Pancuronium	0, 60
Pentobarbitol	0, 60
Precedex	0, 60
Procainamide	0, 60
Propofol	0, 60
Reopro	0, 60
Sandostatin	0, 60
TPA	0, 60
Vasopressin	0, 60
Vecuronium	0, 60

Table 3.5: ISM Medications

Three of the variables in Table 3.4 are discrete. The set of possible values for these variables are listed in Table 3.6. Because the *Service Type* feature is unordered and is not binary, it was converted into seven binary indicator variables for the dataset. The intravenous medications are listed in Table 3.5. While this list is quite long, many of the medications are eventually ignored because of their sparsity (or complete absence).

3.2.3 Outcome Data

For survival analysis, it is necessary to have the final discharge status for the set of patients used. Patient outcome information for the MIMIC II data, however, was limited. When the dataset for this project was constructed, only 2058 patient records

Feature	Values	Description
Censor Indicator	0	Uncensored
	1	Censored
Sex	0	Female
	1	Male
Service	svOther	Other Care Unit
	svCSICU	Cardiac Surgery ICU
	svNSICU	Neurological Surgery ICU
	svMICU	Medical ICU
	svMSICU	Medical Surgery Surgery ICU
	svCCU	Cardiac Care Unit
	svCSRU	Cardiac Surgery Recover Unit

Table 3.6: Discrete Variables

included outcome data. This set of patients is a subset of the set of patients with trend data available.

A method was found to extract a small additional set of patient outcomes. The CareVue system includes a table entitled "CensusEvents". This table provides outcome information for a small subset of the MIMIC II patients. With this additional set of patients included, most of the patients still do not have final discharge information. The CensusEvents table provides the discharge status for a total of 1310 patients. For 252 of these patients, the discharge status was already known. Combining these two sources for outcomes resulted in a set of 3116 patients with outcome information. The number of patients with trend data, however, remained at 2058.

The ability to gain over 1000 patient outcomes prompted the consideration of two different datasets. The first dataset, **Dataset 1**, is composed of the vitals from the trend data along with the ISM medications. This dataset includes a total of 747 patients. If CVP is not required, then this number increases to 2037. The second dataset, **Dataset 2**, uses only ISM vitals, allowing it to utilize the larger set of 3116 patients. Table 3.7 lists the components of these two datasets.

Dataset 1	Dataset 2
199 features	132 features
747 (2037) patients	3116 patients
Trend Vitals	ISM Vitals
ISM Meds	ISM Meds
Additional ISM Features	Additional ISM Features

Table 3.7: Datasets

3.3 Final Dataset

The final datasets were constructed by merging the features in 3.7 into one time-series dataset for both **Dataset 1** and **Dataset 2**. Feature selection was done on the uncensored patients using backward selection with linear regression (described in the next chapter). Feature selection on **Dataset 1** (with the trend data features) returned nearly the same set of features that were selected from the ISM vitals (**Dataset 2**). This was also the case if the trend data did not require CVP measurements. In the end, the difference in the amount of correlation between the features selected from each set and the patient's survival time was negligible. This was rather surprising as it indicated that the higher resolution information included in the trend signals did not significantly contribute to predicting a patient's length of survival. Since **Dataset 2**, utilizing only the ISM data, has a larger set of patients available than **Dataset 1**, it was selected for use as the final dataset. The descriptions that follow refer exclusively to **Dataset 2** (and subsets of this data).

3.3.1 Missing Values

Before performing feature selection, it was necessary to deal with the numerous missing data points (NA's). If all of the instances that contain a NA value were simply omitted, the size of the dataset would be severely reduced. It was therefore helpful to target problematic features. The procedure used was guided by the observation that the number of censored instances greatly outnumbered the number of uncensored instances. The following steps were taken, in the following order, to remove all of the NA's from the dataset:

1. The dataset was split into censored patients and uncensored patients
2. Using the uncensored patients, the following features were removed
   (a) Features that are never present
   (b) Features missing more than 40 percent of their values
   (c) Features with zero variance (constant)
3. The remaining features, not fitting the above criteria for removal, were selected from the censored cases and the two partitions were recombined into a new set
4. All instances (rows) in this new set that contained any missing values were removed

This procedure eliminated 53 features from the original data. The remaining 85 “clean” features are listed in Table 3.8. The original data contained 372,282 instances, 273,153 of which were censored. Removing instances that were missing values for any of the clean features reduced these numbers to 115,128 and 77,340, respectively. This indicates that the cleaned data, with only 85 features, was still quite sparse. In fact, the cleaning procedure eliminated about one half of the patients.

There are several reasons for this sparsity. First, although the measurements considered in the original dataset should be widely available, reality does not always reflect this. Often some of these features are unavailable for a moderate segment of a patient’s stay—or in some cases they are missing altogether. Another cause of this sparsity is merging data of different temporal resolutions together. Although point values were held for fixed windows of time in order to alleviate this issue, many measurements still did not align. This was responsible for a large portion of the instances with missing values. By extending these windows, some of the less frequently measured features could be held for a longer period of time. This would help increase the total number of instances, but over time the validity of these measurements decreases. The two-hour window used for most of the features in Table 3.4 is rather conservative, erring on the side of ensuring that measurements are still valid. Many of these problems occur early in the patient's stay because valid lab results have not yet been obtained. Because of this, many patients that have a short overall stay in the ICU do not include any valid instances and are completely removed from the dataset. The future work section of Chapter 6 discusses some more elaborate data imputation schemes that might help further reduce these issues.

Two suffixes are used for features in Table 3.8. First, if a feature is followed by an "_h", then the number n that directly follows the "_h" indicates that the measurement represents the value n-minutes prior to the current time. For example, the feature "Labetolol_h60", gives the administered value of Labetolol 60 minutes prior to the time of the instance. Similarly, features followed by "_Slope_60" indicate derived features. For example, "SBP_Slope_240" is the slope of the systolic blood pressure derived from a 240-minute window.

3.3.2 Dataset Partitions

In order to train and validate patient models, it is necessary to have data for training and data for testing. Initially, it was assumed that individual instances within a patient could be treated independently from other instances within the same patient. We later found that this assumption was too strong. To prevent classification algorithms from detecting intra-patient similarly, it is necessary to partition the data into disjoint sets of patients rather than disjoint sets of instances.

Consequently, as a final step, the cleaned dataset was randomly partitioned into two sets of patients. The first set, referred to as the **training set** was composed of 70 percent of the patients. The second set, referred to as the **test set**, contained the remaining 30 percent of the patients and was set aside for testing purposes.

3.3.3 Data Summary

Table 3.9 provides a brief summary of the final dataset (referred to as *cleaned*). It is clear from this table that by removing problematic (highly sparse) features, the number of valid instances was kept relatively high. The two partitions of the *cleaned*

survTime	index	Morphine_Sulfate
Morphine_Sulfate_h60	Lasix	Diltiazem
Diltiazem_h60	Dobutamine	Dobutamine_h60
Nitroglycerine	Nitroglycerine_h60	Sandostatin
Sandostatin_h60	Nitroprusside	Nitroprusside_h60
Lidocaine	Lidocaine_h60	Labetolol
Labetolol_h60	Milrinone	Milrinone_h60
Epinephrine	Epinephrine_h60	Neosynephrine
Neosynephrine_h60	Heparin	Heparin_h60
Fentanyl	Fentanyl_h60	Amiodarone
Amiodarone_h60	Ativan	Ativan_h60
Levophed	Levophed_h60	SBP
SBP_Slope_240	SBP_Slope_60	DBP
DBP_Slope_240	DBP_Slope_60	MAP
MAP_Slope_240	MAP_Slope_60	CV_HR
CV_HR_Slope_240	CV_HR_Slope_60	SpO2
SpO2_Slope_240	SpO2_Slope_60	RESP
RESP_Slope_240	RESP_Slope_60	CVP_Slope_240
Sex	Age	Input_Sum_720
Input_Sum_720_h60	Input_Sum_720_h120	Output_Sum_720
Output_Sum_720_h60	Output_Sum_720_h120	Weight
Weight_h120	Art_pH	Art_pH_h120
Art_BE	Art_BE_h120	Art_CO2
Art_CO2_h120	Art_PaCO2	Art_PaCO2_h120
Art_PaO2	Art_PaO2_h120	svCSICU
svCSICU_h120	svNSICU	svNSICU_h120
svMICU	svMICU_h120	svMSICU
svMSICU_h120	svCCU	svCCU_h120
svCSRU	svCSRU_h120	Censored

Table 3.8: Clean Features

Dataset	Features	Instances	Censored	Patients
raw	138	372,282	273,153	1,842
raw (*NA*'s omitted)	138	51,732	35,031	708
cleaned	85	115,128	77,340	880
train	85	81,374	56,712	615
test	85	33,754	20,628	265

Table 3.9: All Patients: Final Datasets

Dataset	Features	Instances	Censored	Patients
raw	138	90,005	61,856	487
raw (*NA*'s omitted)	138	6,575	4,165	143
cleaned	72	45,056	29,665	338
train	72	31,830	21,237	235
test	72	13,225	8,428	103

Table 3.10: MICU Patients: Final Datasets

dataset are listed in this table as **train** and **test**. The **train** dataset is used for the feature selection and modeling described in the next chapter. The **test** dataset is used for model evaluation.

Subsets of this final dataset were also examined. Specifically, the following sets of patients were selected: patients in the MICU, patients who were in the MICU with an ICD9 code indicating hypovolemia, and all patients with an ICD9 code indicating hypovolemia. For each of these groups, the same procedure was followed as outlined for the aggregate set of patients. Table 3.10 gives a summary of the resulting datasets for the MICU patients.

Chapter 4

Patient Models

Using the dataset described in the previous chapter, we now explore various outcome models. To start with, we consider models built using all patients in the training set. Later we focus on more specific models trained on subsets of those patients.

4.1 Aggregate Dataset Models

The aggregate dataset includes all of the patients that remained after preprocessing. Table 3.9 shows that there are a total of 880 patients in this dataset before being split into a training set used for training models and a testing set used only for evaluation purposes. We start our analysis by attempting to predict which of the patients in this dataset are censored. Next we explore the feature space through feature selection. Using a reduced set of features, we fit the two most common survival regression models. After fitting these models, we provide some brief diagnostics to test their validity.

4.1.1 Predicting Censoring

One assumption that virtually all survival analysis techniques make is that the process that generates the censored times is independent of the covariates being explored. Initially, this requirement might seem too strong for the setting being explored here, where leaving the ICU alive would presumably depend on the patient's vitals and other physiological measurements. In one case, if a patient's monitored signals continue to improve, then the patient is likely to be discharged (censored). Another case might be that the patient's state has been steadily declining and as a result the patient is discharged to hospice care.

However, attempts to predict patient censoring were largely unsuccessful. Using an SVM classifier, the best predictor found did little better than a random guess. This classifier was trained using the training set. The ROC curve for this predictor on the test set is shown in Figure 4-1. An ROC curve plots the true positive rate (*sensitivity*) versus the false positive rate (*1-specificity*) for different prediction thresholds. An ideal classifier has a total area of 1, while a diagonal line is equivalent to a random guess. The area under the curve (AUC) is a typical metric used to evaluate the strength of a classifier from the ROC curve. The AUC for this predictor is 0.628, indicating significant room for improvement.

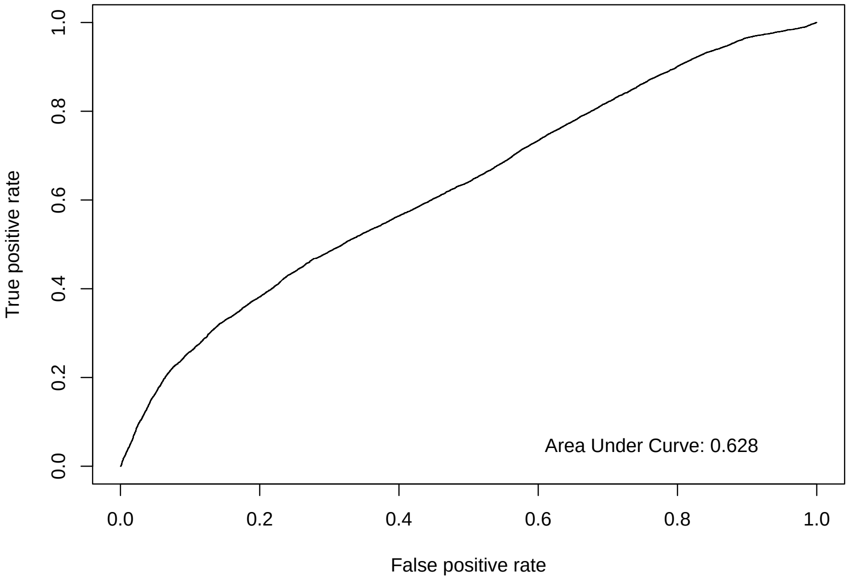


Figure 4-1: SVM Censored Prediction

The *LibSVM* library [11] along with an interface to *R* written by Dimitriadou

[18] were used to train this model. The ROC curves were generated using the *ROCR* package [53].

Looking at the failure times for censored versus uncensored instances supports the conclusion that the censoring is uninformative. As Figure 4-2 and Figure 4-3 show, the distribution of failure times for the uncensored cases is quite similar to the distribution of censored times. These failure time distributions also appear to be exponential in nature. If the failure times are transformed using a log scale, it appears that a lognormal distribution would be appropriate for modeling them.

Figures 4-2 and 4-3 are both based on the total survival time for each patient. We also examined the failure distributions if time dependence is not considered, using the time until failure for each instance in a particular patient's stay. This resulted in the distributions shown in Figures 4-4 and 4-5.

These distributions are more difficult than the time-dependent ones. The primary complication is missing data. If all instances were available for each patient then the failure time distribution would be monotonically decreasing. But in reality, many of the patients that survived for a long time have significant gaps, making the distribution less informative. While the transformed distributions are still relatively lognormal in nature, they clearly have more noise. This suggests that a time-dependent model would likely perform better than one that ignores the relationship between the failure times of instances within a particular patient.

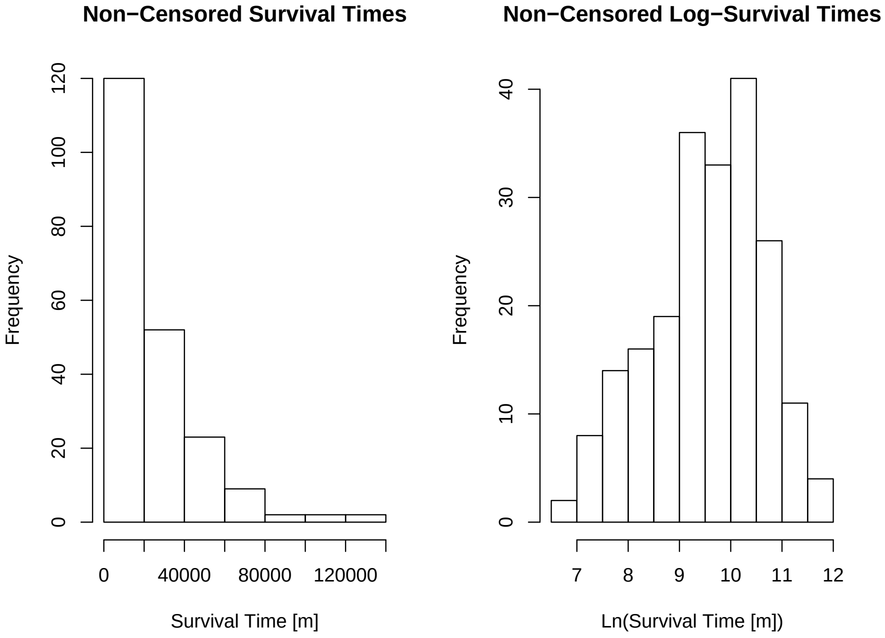


Figure 4-2: Survival Histograms for Uncensored Patients

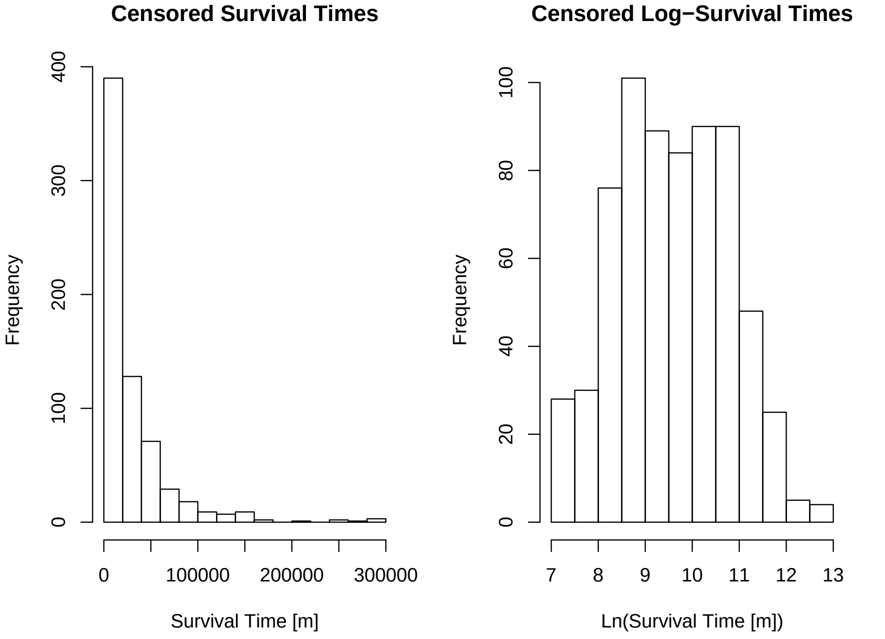


Figure 4-3: Survival Histograms for Censored Patients

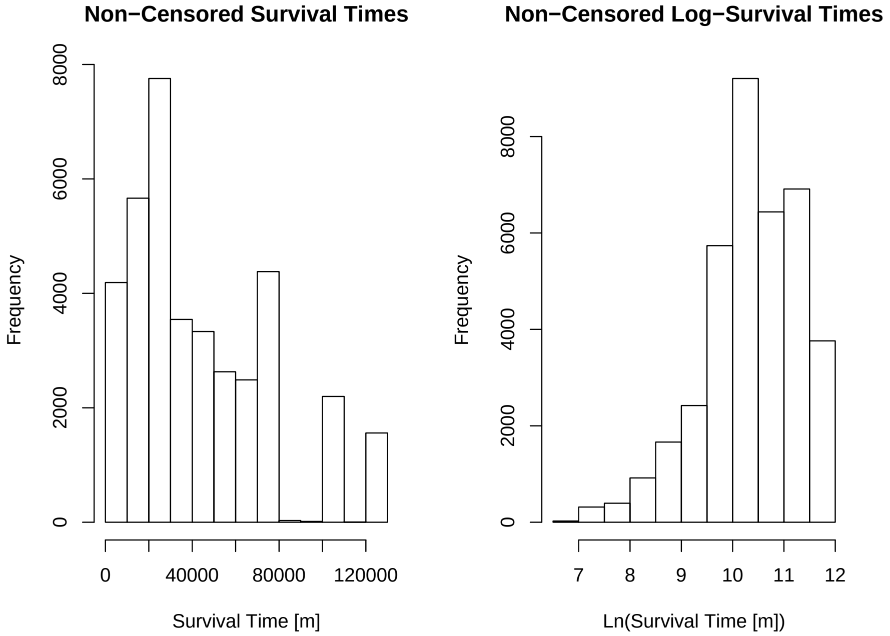


Figure 4-4: Survival Histograms for Uncensored Patients (all instances)

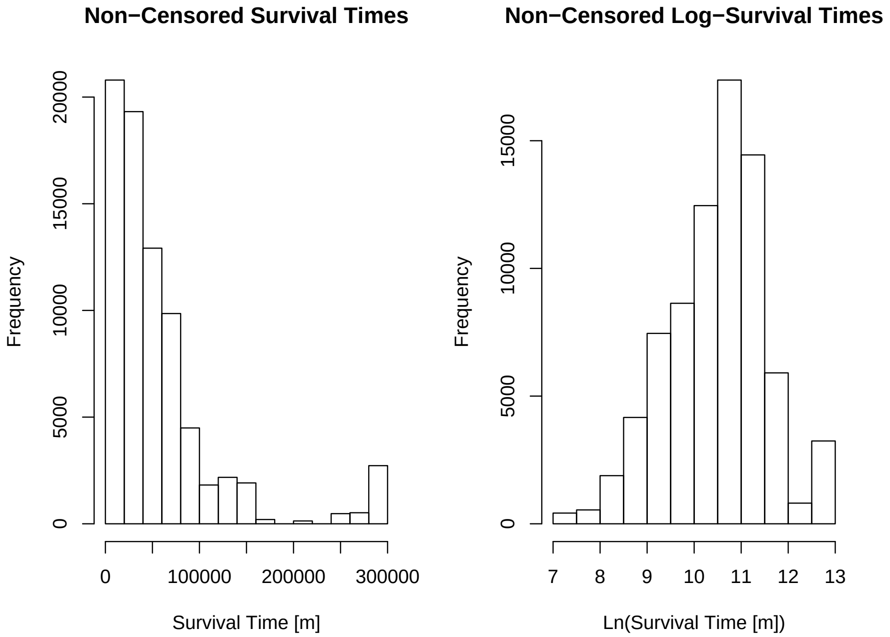


Figure 4-5: Survival Histograms for Censored Patients (all instances)

4.1.2 Feature Selection

Using the cleaned dataset described in the previous chapter, we start by finding the subset of the features that are most related to the patient's length of survival. To do this the uncensored training instances were fitted using forward and backward variable selection with linear regression.

Forward selection is a simple approach that seeks to add one variable at a time to the set of explanatory variables. At each step, the most significant excluded variable (using its p-value) is added to the model. Table 4.1 shows the top 15 features found using this selection process on the uncensored portion of the dataset. This table also lists the cumulative R^2 value, indicating the amount of correlation obtained as the feature set increases. The *LEAPS* package was used with *R* to perform the forward selection [63].

Backward selection is similar to forward selection, except that it starts with all of the variables included in the model. The least significant variables are sequentially dropped, and the model is refitted. For this particular dataset, the forward and backward selection methods resulted in the same set of features. Table 4.2 shows the top 15 features found using backward selection. It also lists the corresponding cumulative R^2 values.

Finally, forward and backward selection were also performed when the model response was the logarithm of the survival time instead of the actual survival time. As expected, this resulted in a slightly better correlation because the penalty for very long survival times was decreased. The features resulting from the two selection methods are shown in Tables 4.3 and 4.4.

Feature	R^2 (Cumulative)
svCSRU	0.0651
svCSICU_h120	0.1066
Input_Sum_720_h120	0.1328
DBP	0.1461
Fentanyl	0.1574
RESP	0.1685
Dobutamine	0.1780
Output_Sum_720_h60	0.1876
Art_PaO2	0.1932
Neosynephrine_h60	0.1980
SBP	0.2024
svNSICU	0.2057
Heparin	0.2096
index	0.2134
Amiodarone_h60	0.2169

Table 4.1: Top Fifteen Features using Forward Selection

Feature	R^2 (Cumulative)
svCSRU	0.0651
svCSICU	0.1065
Input_Sum_720_h120	0.1327
Fentanyl_h60	0.1457
Output_Sum_720	0.1588
Dobutamine_h60	0.1690
DBP	0.1777
Art_PaO2	0.1856
Neosynephrine_h60	0.1913
svNSICU	0.1962
Heparin	0.2012
SBP	0.2062
RESP	0.2092
index	0.2129
Amiodarone_h60	0.2164

Table 4.2: Top Fifteen Features using Backward Selection

Feature	R^2 (Cumulative)
Input_Sum_720	0.04716
svCSICU	0.08506
svCSRU	0.12395
Output_Sum_720	0.14864
DBP	0.16470
SpO2	0.17608
RESP	0.18323
Amiodarone_h60	0.19104
svCCU	0.19621
Art_BE	0.20176
Milrinone_h60	0.20735
Weight	0.21207
Neosynephrine_h60	0.21677
Morphine_Sulfate_h60	0.22094
Fentanyl	0.22450

Table 4.3: Top Fifteen Features using Forward Selection and Log Transform

Feature	R^2 (Cumulative)
Input_Sum_720_h120	0.04618
svCSICU	0.08482
svCSRU	0.12374
Output_Sum_720	0.14709
DBP	0.16339
SpO2	0.17637
RESP	0.18418
Fentanyl	0.19018
Art_BE	0.19664
Sex	0.20091
Weight	0.20714
Milrinone_h60	0.21241
svCCU	0.21753
Morphine_Sulfate_h60	0.22195
Neosynephrine_h60	0.22619

Table 4.4: Top Fifteen Features using Backward Selection and Log Transform

From these tables it is evident that the particular care unit that the patient is in is one of the best predictors of the patient's survival time. Figure 4-6 illustrates the different Kaplan-Meier survival curves for patients who are in the Cardiac Surgery Recovery Unit (CSRU) versus those who are not. This curve was fitted using the time-dependent version of the survival times (effectively making them interval censored) in order to prevent single patients from obscuring the curves. Each of these estimated curves include the associated confidence bars.

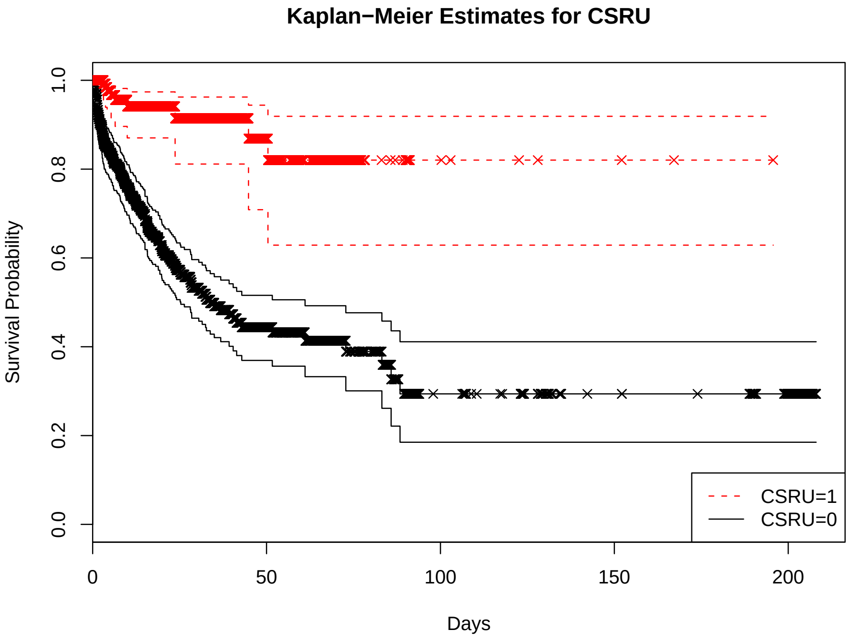


Figure 4-6: Kaplan-Meier Estimates for CSRU and non-CSRU Patients

4.1.3 Survival Regression

Table 4.5 lists the features found by modeling the survival time combined with those found by modeling the log-survival time. The *Age* and *Sex* (with values 1 corresponding to "male" and 0 corresponding to "female") features were also added to this set. Given these features found using forward and backward selection, we can proceed to fit the various survival models that were described in Chapter 2 to the data.

Input_Sum_720	svCSICU
svCSRU	Output_Sum_720
DBP	SpO2
RESP	Amiodarone_h60
svCCU	Art_BE
Milrinone_h60	Weight
Neosynephrine_h60	Morphine_Sulfate_h60
Fentanyl	svCSICU_h120
Input_Sum_720_h120	Dobutamine
Output_Sum_720_h60	Art_PaO2
SBP	svNSICU
Heparin	index

Table 4.5: Merged Feature Set

Cox Proportional Hazards Model

Since the Cox Proportional Hazards (CPH) model is likely the most widely used survival regression model, it was explored first. For this model, the *Design* package [27] written for *R* by Harrell was used.

Typically it is desirable to perform feature selection using the survival model under consideration. Using linear regression and goodness of fit to select the important features could possibly miss features that would be useful for a given survival model. The large number of features under consideration, however, made feature selection directly from the survival model infeasible by preventing the process that fits the CPH model from converging. By first taking a liberal set of features that have the most correlation with length of survival and training the survival model on this set, feature selection can be repeated on the survival model with little risk of missing important features. This is the approach we took. Even with this reduced set of features, manual adjustment was necessary to find survival models that successfully converged. The process used is described below.

1. The features in Table 4.5 were given to the CPH model as covariates

2. The *Milrinone_h60* feature caused the covariate matrix to be singular so it was removed

3. The *index* (minutes since admission) feature prevented the model from converging so it was removed

The successfully fitted model had the following characteristics:

```
Cox Proportional Hazards Model

   Obs   Events Model L.R.   d.f.    P    Score   Score P
 81374      133     256.24     20    0   638.98         0
    R2
 0.182

                          coef se(coef)        z        p
Input_Sum_720         0.000132 3.89e-05  3.39276 6.92e-04
Output_Sum_720       -0.000174 1.10e-04 -1.58379 1.13e-01
svCSICU              -0.334986 3.48e-01 -0.96271 3.36e-01
svCSRU               -1.537861 3.58e-01 -4.29391 1.76e-05
DBP                   0.002568 8.69e-03  0.29554 7.68e-01
SpO2                 -0.039776 4.70e-03 -8.46278 0.00e+00
RESP                  0.006327 1.34e-02  0.47125 6.37e-01
Amiodarone_h60        0.012505 6.67e-01  0.01876 9.85e-01
svCCU                -0.073452 2.47e-01 -0.29716 7.66e-01
Art_BE               -0.093450 1.66e-02 -5.63053 1.80e-08
Weight               -0.000562 3.38e-03 -0.16652 8.68e-01
Neosynephrine_h60    -0.000239 2.53e-02 -0.00946 9.92e-01
Morphine_Sulfate_h60  0.069753 2.93e-02  2.38237 1.72e-02
Fentanyl              0.003492 9.83e-04  3.55129 3.83e-04
Age                   0.017686 7.00e-03  2.52698 1.15e-02
Sex                   0.571557 1.95e-01  2.93800 3.30e-03
Art_PaO2              0.004281 1.86e-03  2.29750 2.16e-02
Heparin              -0.000614 3.02e-04 -2.03411 4.19e-02
SBP                  -0.016532 4.35e-03 -3.79632 1.47e-04
svNSICU              -0.308775 2.57e-01 -1.20087 2.30e-01
```

The final column labeled "p" shows the significance of the various features used in the model. Several of the features do not significantly improve the model. Namely, *720min Output Sum*, *CSICU*, *Diastolic blood pressure*, *Respiratory rate*, *Amiodarone*, *CCU*, *Weight*, *Neosynephrine*, and *NSICU* appear to be of little use to the model. Performing backward selection on the model removed all of these features along with *Heparin* and *Arterial PaO2*.

Finally, before accepting the CPH model, interactions that seemed meaningful were added to see if they would improve the performance. Specific interactions that we tried included *Heart rate and Systolic blood pressure*, *Heart rate and SpO2*, *Systolic blood pressure and Diastolic blood pressure*, *Respiratory Rate and Heart Rate* and *720min input sum and 720min Output Sum*. The Input_Sum-Output_Sum interaction prevented the model from converging. The others allowed the model fitting to converge, but did not significantly improve the model performance.

The final resulting model is shown in Equation 4.1. Values for the baseline survival function, S_0, are listed in Table 4.6.

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{\boldsymbol{\beta}^T \mathbf{x}}}, \tag{4.1}$$

where

$$\begin{aligned}
&\boldsymbol{\beta}^T \mathbf{x} = \\
&4.741622 + 9.601624 \times 10^{-5} \text{Input\_Sum\_720\_h120} + 0.01762138\ \text{Age} \\
&- 1.495267\ \text{svCSRU} - 0.04092318\ \text{SpO}_2 - 0.1053184\ \text{Art\_BE} \\
&+ 0.002741086\ \text{Fentanyl} + 0.4232253\ \text{Sex} - 0.01523348\ \text{SBP} \\
&+ 0.05494363\ \text{Morphine\_Sulfate\_h60}.
\end{aligned}$$

t	$S_0(t)$
0	1.000
30	0.781
60	0.705
90	0.608
120	0.608
150	0.608
180	0.608

Table 4.6: Baseline Survival Values

The limited amount of correlation present in this model gives it a rather large level of uncertainty. Figure 4-7 compares the CPH estimates for patients in the CSRU to those who are not. In order to contrast the CSRU patients with the non-CSRU patients, this model uses the median values for all of the other covariates. We can see that the CPH model's predictions are quite similar to the Kaplan-Meier estimates for this same comparison (Figure 4-6).

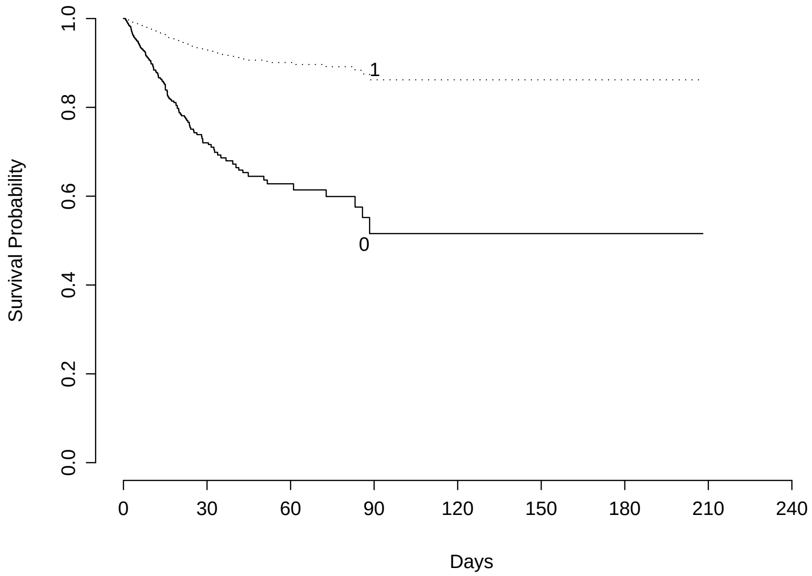


Figure 4-7: Cox Proportional Hazards Estimates for CSRU and non-CSRU Patients

Accelerated Lifetime Model

For comparison purposes, a parametric Accelerated Lifetimes Model (ALM) was also fitted to the data. Both the Weibull distribution and the lognormal distributions were evaluated. Using *R*, the *Design* package [27] provides a routine for fitting parametric ALMs. This routine is based on the *Survival* package [44] originally developed by Therneau, and adds automatic backward selection and analysis of variance. This routine had problems similar to those encountered using the CPH routine for attempting to handle a large number of features.

For the Weibull distribution, the successfully fitted model had the following characteristics:

```
   Obs     Events Model L.R.    d.f.     P      R2
 81374      24662   13881.52       9     0    0.16

                         Value Std. Error      z         p
(Intercept)           9.10e+00   1.04e-01  87.13  0.00e+00
Input_Sum_720_h120   -9.08e-05   3.05e-06 -29.74 2.37e-194
Age                  -1.84e-02   5.74e-04 -32.09 5.28e-226
svCSRU                1.13e+00   1.92e-02  58.77  0.00e+00
SpO2                  2.62e-02   9.44e-04  27.70 6.38e-169
Art_BE                5.90e-02   1.83e-03  32.22 9.15e-228
Fentanyl             -3.32e-03   8.24e-05 -40.30  0.00e+00
Sex                  -7.75e-01   1.81e-02 -42.90  0.00e+00
SBP                   1.55e-02   3.52e-04  43.98  0.00e+00
Morphine_Sulfate_h60  2.02e-01   2.19e-02   9.21  3.25e-20
Log(scale)            2.13e-01   5.00e-03  42.60  0.00e+00

Scale= 1.24
```

A model was fit using the lognormal distribution as well. Based on the R^2 value, this performs marginally worse than the model fitted using the Weibull distribution. The model using the lognormal distribution has an R^2 of about 0.15 versus an R^2 of about 0.16 for the Weibull model.

One of the benefits of the parametric model is the ability to plot the corresponding hazard function. For example, as in the CPH model, we can examine the effect of one of the covariates on the curves using the median values for the other covariates. Figure 4-9 shows a set of curves from the Weibull ALM for different ages. For comparison, Figure 4-8 shows the same set of curves from the lognormal ALM. As expected, the hazard curve is higher and the survival curve is lower as the age increases. Unlike the Weibull model, the lognormal model is not restricted to being monotonically decreasing. This can be seen in the lognormal model's non-zero maximum for the hazard function.

It is important to note that the accelerated lifetime models do not account for time dependencies, effectively considering each instance as an independent patient. Using this construction, individual patients who stay in the ICU for a particularly long time can bias the results.

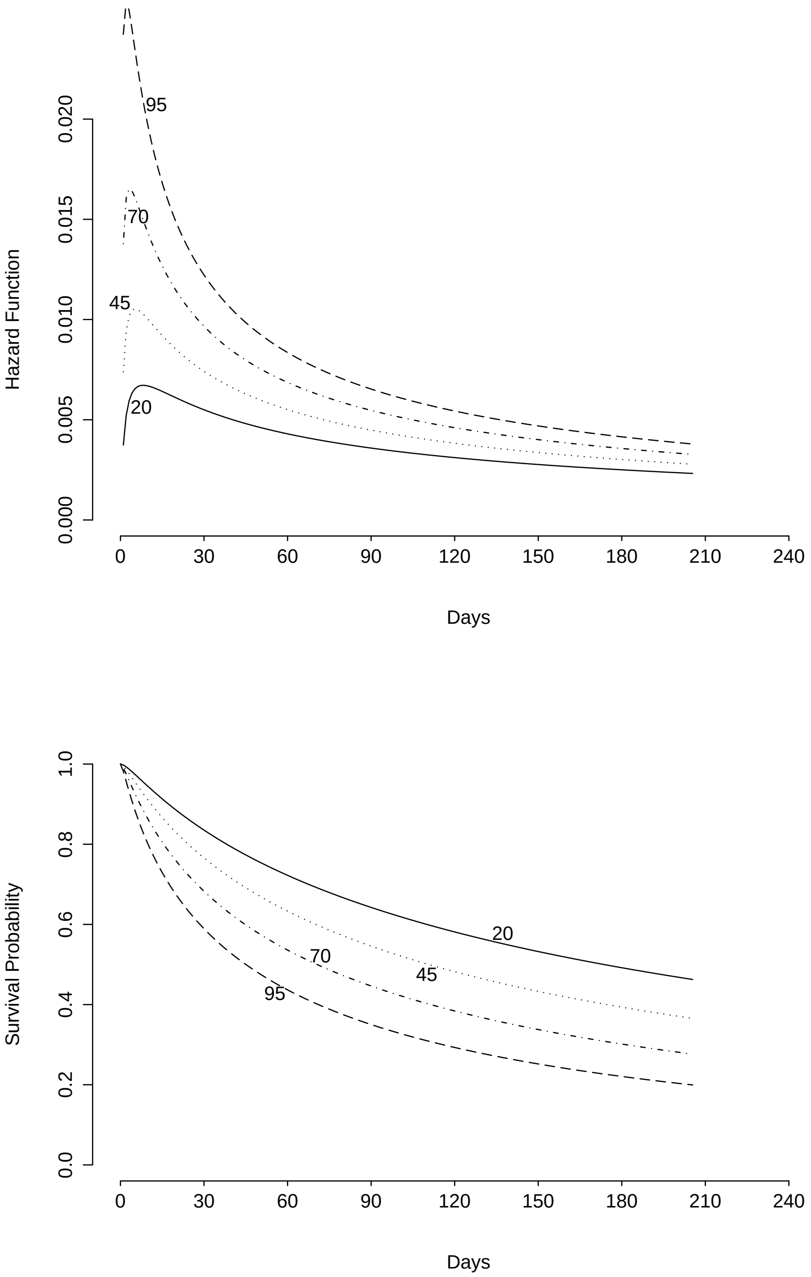


Figure 4-8: Lognormal ALM: Survival and Hazard curves for Age

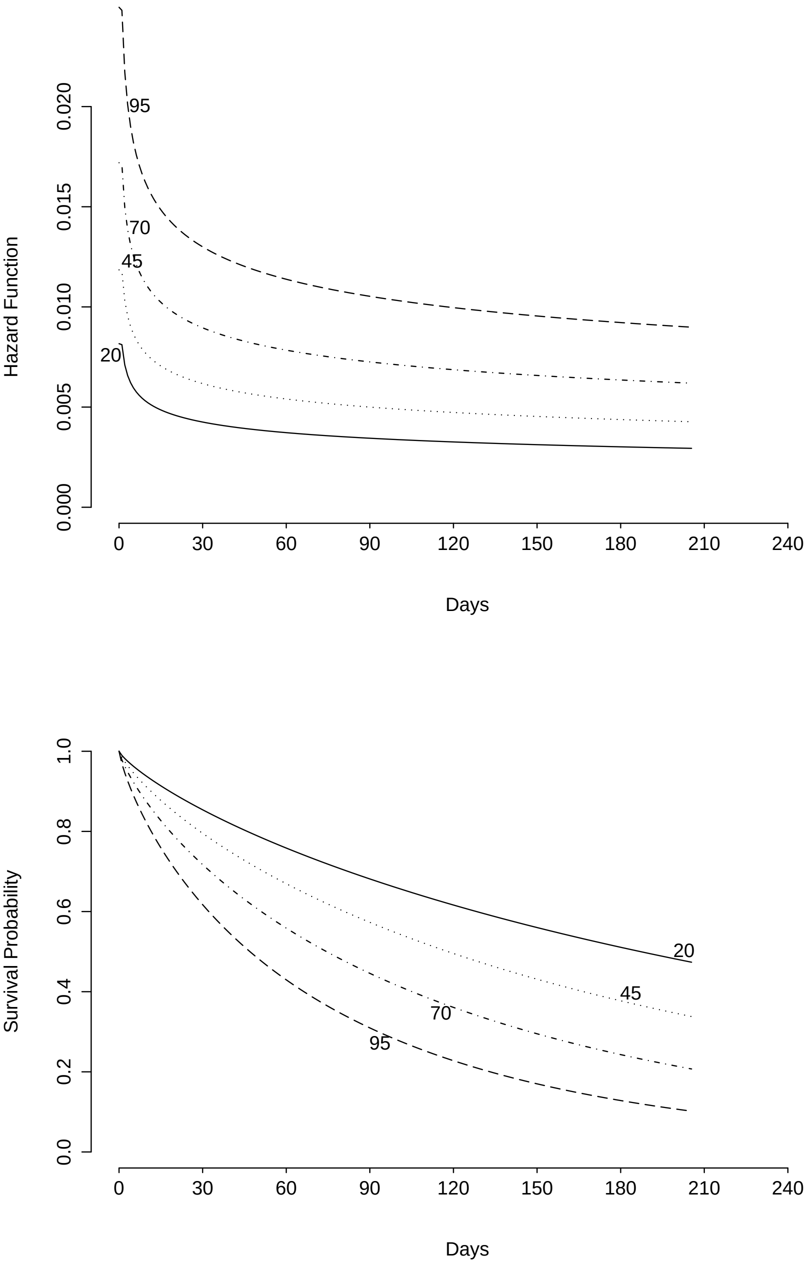


Figure 4-9: Weibull ALM: Survival and Hazard curves for Age

4.1.4 CPH Model Diagnostics

A number of diagnostics exist for testing if a fitted Cox proportional hazards model is appropriate for the data. We consider three tests. First, in order to verify the assumption of proportional hazards, we plot the *scaled Schoenfeld residuals*. Next, we use *standardized delta-betas*, also known as *dfbetas*, for determining overly influential observations. Finally, we examine the possibility of nonlinearity between the log hazard and the covariates by plotting the *martingale residuals* against individual covariates.

Scaled Schoenfeld Residuals

Defining residuals in ordinary linear regression is quite straightforward. For survival models, however, a more complicated definition of a residual is necessary. Several different residuals have been suggested for CPH models and each provides insight into different characteristics of the model.

To determine the overall fit of the proportional hazards model, we calculated the scaled Schoenfeld residuals for each feature. Equation 4.2 shows the Schoenfeld residual [51], for feature j at observation i:

$$r_{ji} = \delta_i \left(x_i^{(j)} - \frac{\sum_{m \in R(y_i)} x_m^{(j)} e^{\boldsymbol{\beta}^T \mathbf{x}_m}}{\sum_{m \in R(y_i)} e^{\boldsymbol{\beta}^T \mathbf{x}_m}} \right). \tag{4.2}$$

As before, $\boldsymbol{\beta}$ is the vector of coefficients for each feature, $R(y)$ is the risk set at time y, and δ_i is the censoring indicator for instance i. The parenthesized superscript, $x^{(j)}$, selects only the jth feature from the vector of covariates.

As a result of the censoring indicator in this equation, censored observations always have a residual of zero. The other residuals are simply the difference between a covariate's value at a particular observation and the weighted average of that covariate from the set of individuals at risk (see [59]). We can test the significance of the residuals for each covariate and also perform a global test for the residuals as a whole using the *cox.zph* routine [27]:

```
                         rho    chisq      p
Input_Sum_720_h120   0.01047 1.69e-02 0.8967
Age                  0.06202 5.69e-01 0.4507
svCSRU              -0.03055 1.24e-01 0.7249
SpO2                -0.13103 2.23e+00 0.1357
Art_BE               0.12305 2.96e+00 0.0855
Fentanyl             0.11904 1.93e+00 0.1644
Sex                  0.10751 1.52e+00 0.2182
SBP                  0.16080 3.11e+00 0.0778
Morphine_Sulfate_h60 -0.00239 6.71e-04 0.9793
GLOBAL                    NA 1.17e+01 0.2323
```

From this output, it appears that the residuals are insignificant for all of the covariates, with the possible exception of *SBP* and *Art_BE*. Graphically examining these residuals is recommended for verifying the overall fit of a proportional hazards model to the data. Figure 4-10, on page 69, shows the residuals for each covariate. In this figure the solid line is a local regression line, and the dotted lines represent the ±2-standard-error thresholds. None of the fitted lines for the covariates appear to have a slope that consistently deviates from zero. This indicates that the proportional hazards assumption is likely to be reasonable.

Standardized Delta-Betas

To examine the influence of individual observations we calculate *dfbetas* for each observation. The *dfbeta* for an observation is quite simple in concept. First, we let $\hat{\boldsymbol{\beta}}$ be the estimate of the $\boldsymbol{\beta}$ vector in the model. Now, denote the estimated coefficients obtained without the ith observation as $\hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}}$. The *dfbeta* for the ith observation is $dfbeta_i = \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}_{(i)}}$. In practice this can be expensive to compute for large datasets, and an efficient approximation is used. Additionally, it is useful to normalize these values using their standard error values. Using these normalized approximations, Figure 4-11 on page 70 shows the *dfbetas* for each observation. The majority of the observations appear to have a reasonably small influence on the estimates for the coefficients.

There are, however, a few overly influential observations. The most blatant of these occurs at 69689. Examining this observation closer reveals why it is so large. This observation is the last observation for a patient that stayed in the ICU for more than 76 days before being discharged alive. The anomaly comes from the fact that the last of the six observations available for the patient occurred only two days into the patient's stay. This resulted in the particularly influential observation representing a huge time interval of about 74 days. With the exception of this observation, all but one of the *dfbetas* are below 0.4 standard errors of the coefficients, making them quite reasonable.

Martingale Residuals

From probability theory, a continuous-time martingale is a stochastic process with the following property:

$$E(X_t|X_r, r \leq s) = X_s,$$

for all $s \leq t$. The martingale residual is constructed as follows. First, define the counting process $N_j(t)$ as 1 when patient j has died at or before time t and 0 if the patient is still alive. Also, let $R_j(t)$ indicate whether or not patient j is at risk at time t. The following process is a martingale if the proportional hazards model is correctly specified [29]:

$$M_j(t) = N_j(t) - \int_0^t R_j(u) e^{\boldsymbol{\beta}^T \mathbf{x}_j(u)} dH_0(u),$$

where $x_j(t)$ is the vector of time-dependent covariates and $H_0(u)$ is the baseline cumulative hazard at time u.

Martingale residuals result when the estimates for $\boldsymbol{\beta}$ and $H_0()$ are used instead of the actual values [60]. The residual, $\hat{M}_j(t)$, can then be defined as

$$\hat{M}_j(t) = N_j(t) - \int_0^t R_j(u) e^{\hat{\boldsymbol{\beta}}^T \mathbf{x}_j(u)} d\hat{H}_0(u).$$

In effect, the martingale residual at time t is the difference between the observed number of deaths over the interval $[0, t]$ minus the expected number of deaths given by the specific model. Because the actual number of deaths for patient j can not be greater than 1, the residuals are always less than or equal to 1. On the other hand, when the expected number of deaths over the segment is large, the residual can be less than -1.

The functional form of each of the covariates can be explored by partitioning the set of covariates. In partitioning, we let the covariate under investigation be x_1 and we let $\mathbf{x}^*$ be the set of remaining covariates. The CPH model can then be written as

$$h(y|\mathbf{x}^*, x_1) = H_0(y)e^{f(x_1)}e^{(\boldsymbol{\beta}^{*T}\mathbf{x}^*)}.$$

Therneau *et al* show that a smoothed plot of $\hat{M_j}$ versus the different values of the covariate x_1 will generally provide the correct form for $f()$. If the smoothed plot is linear, then x_1 does not need to be transformed.

Figure 4-12 shows the martingale residuals for each of the six covariates that are not dichotomous. For this figure, there is a dashed horizontal line that was fit to the data using linear least squares and a smoothed solid line that was fit using local regression. The *Loess* method (also known as locally weighted polynomial regression), natively available in the *R* language [45], was used for the local regression[1]. The default values for each of the parameters to *Loess* were used. The linear appearance of these plots indicates that the covariates do not need to be transformed. In fact, looking at the plots, the two lines can not be distinguished from each other.

We can also add the components contributed by each covariate ($\beta_i x_i$) to the martingale residuals. If we plot these versus the different covariate values (x_i), we can again see that the plots are quite linear by observing that the line fit using linear least squares (dashed) is virtually identical to the smoothed line fit using local regression (solid line). These plots are shown in Figure 4-13.

There appears to be two “groups” of residuals in most of the martingale residual

[1] Loess [14] is a common local regression method that works by fitting low-order polynomials to localized subsets of the data using weighted least squares. Weighted least squares gives more weight to points near the point being estimated than points further away.

plots. The group of residuals at or near 1 are cases where the patient died but the model predicted no deaths (or a small, fractional number of deaths). While this group appears to be rather common, the majority of the residuals are clearly zero.

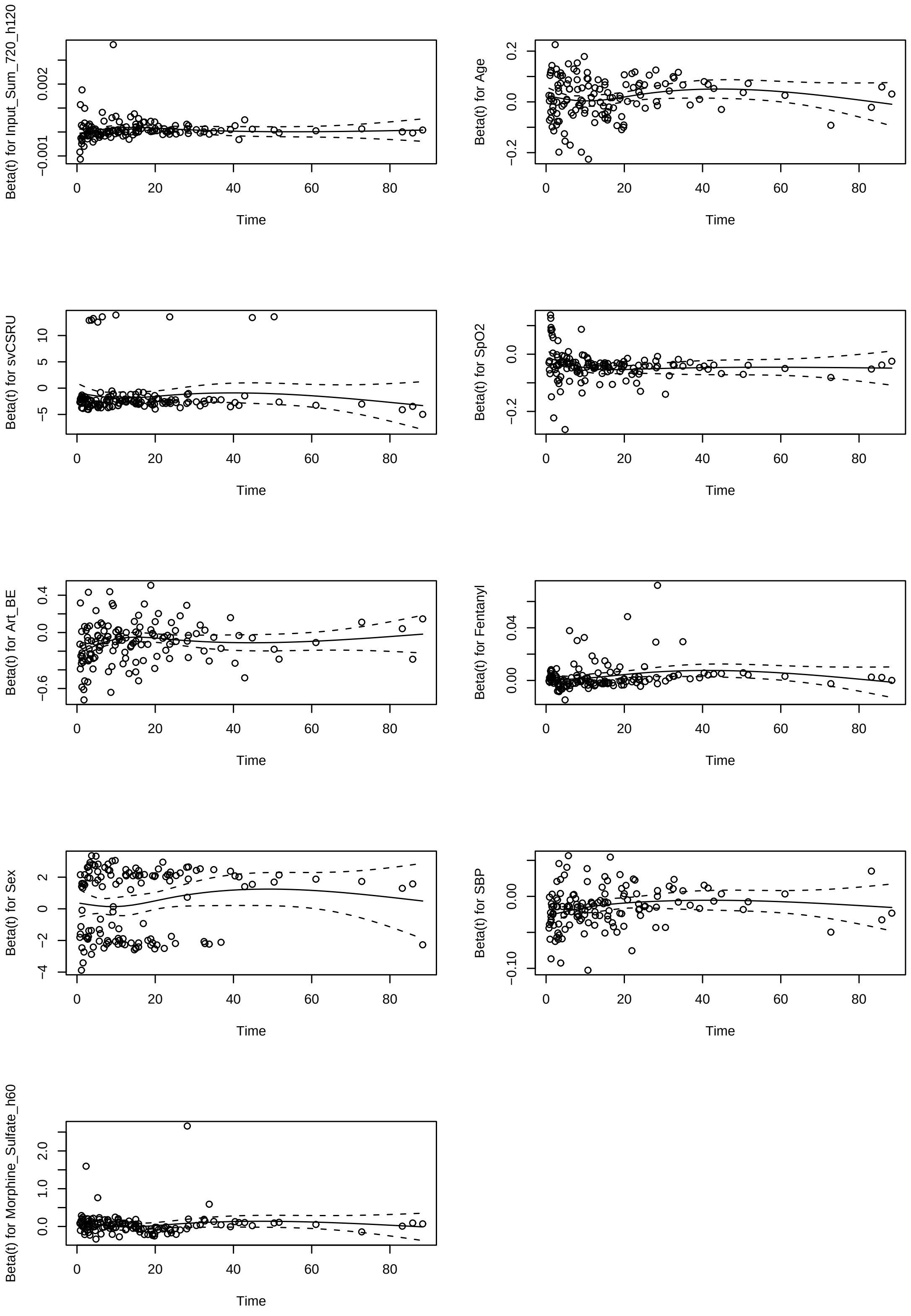


Figure 4-10: Scaled Schoenfeld Residuals vs time in days

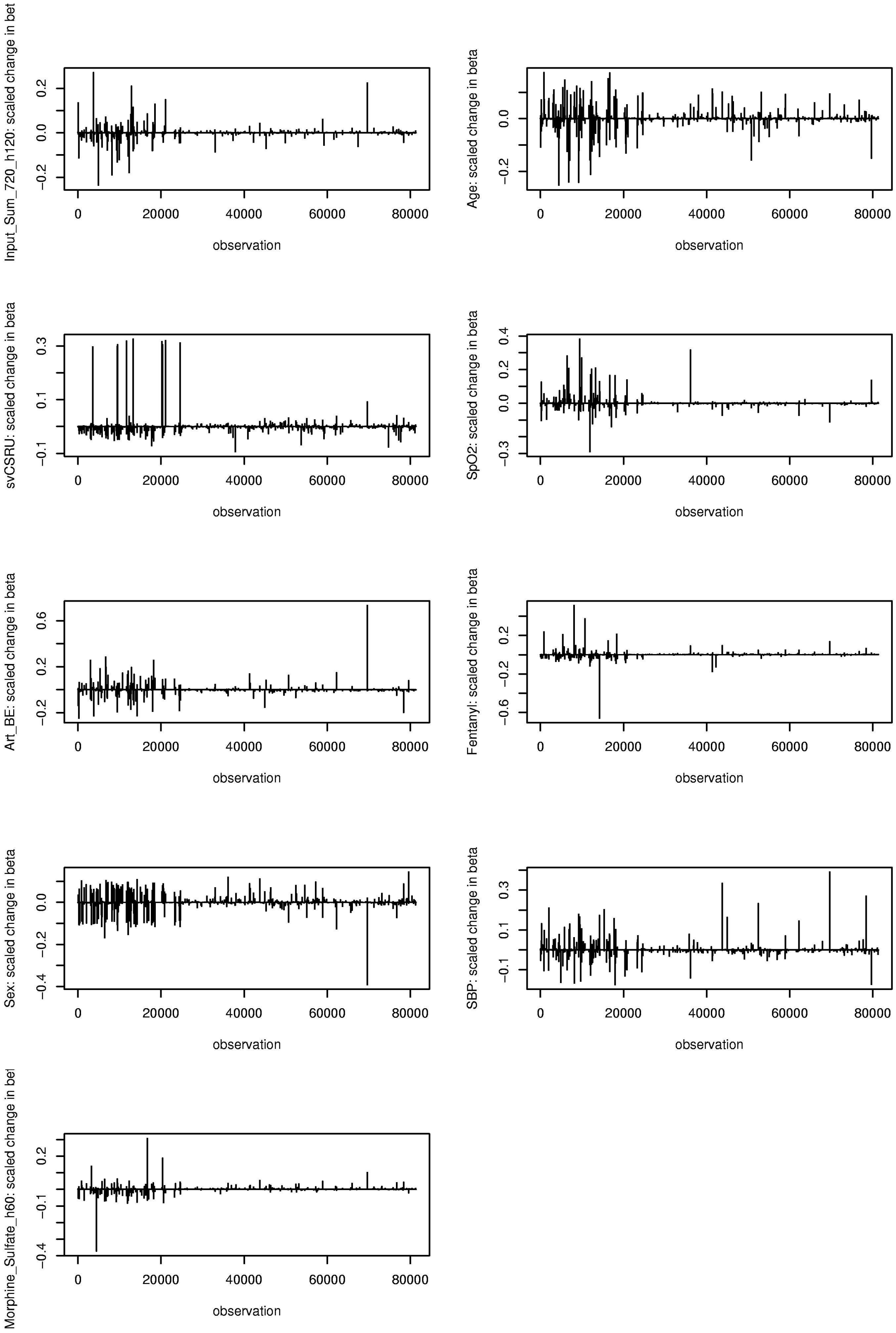


Figure 4-11: Index plots of dfbeta

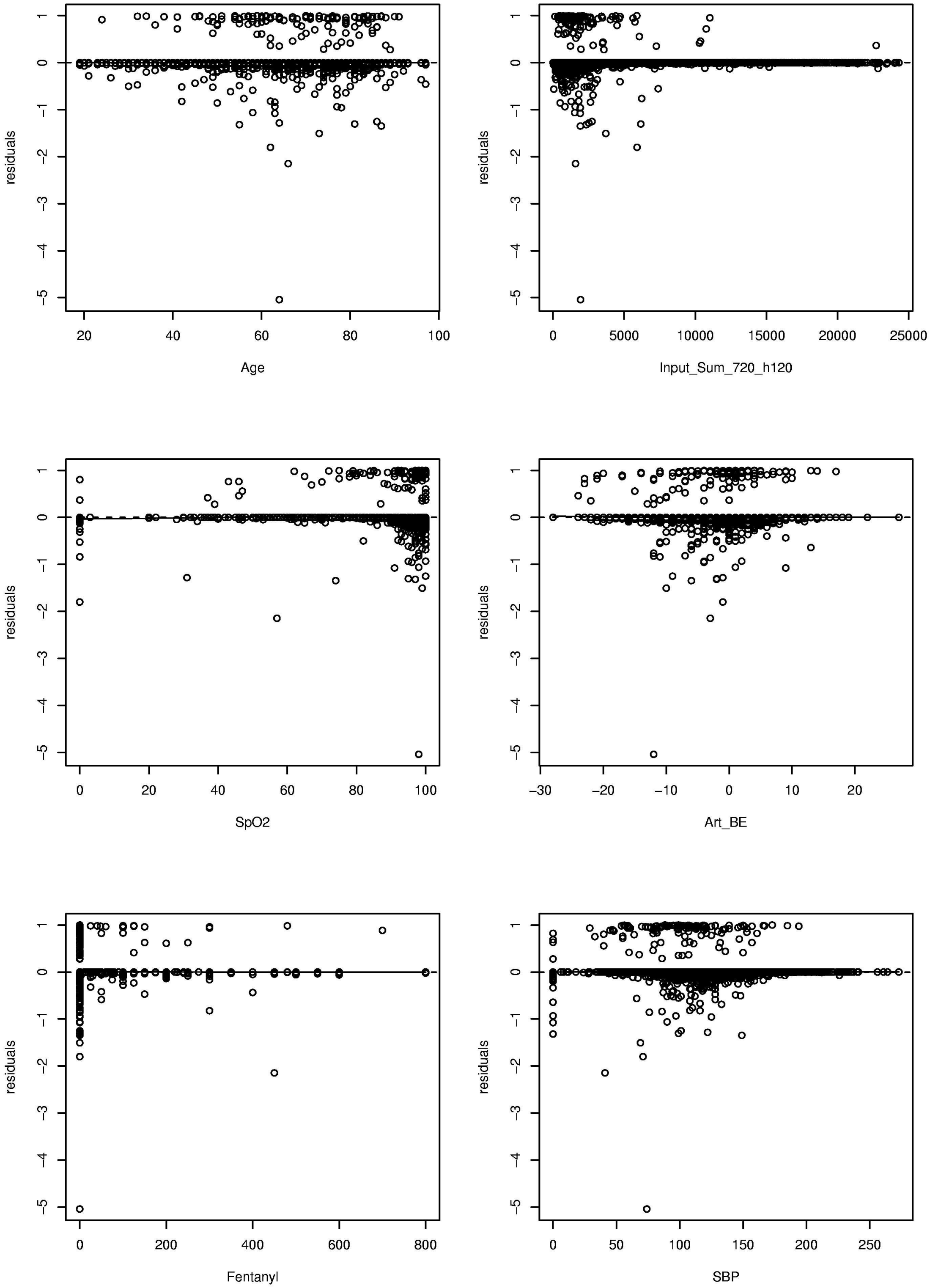


Figure 4-12: Martingale Residuals

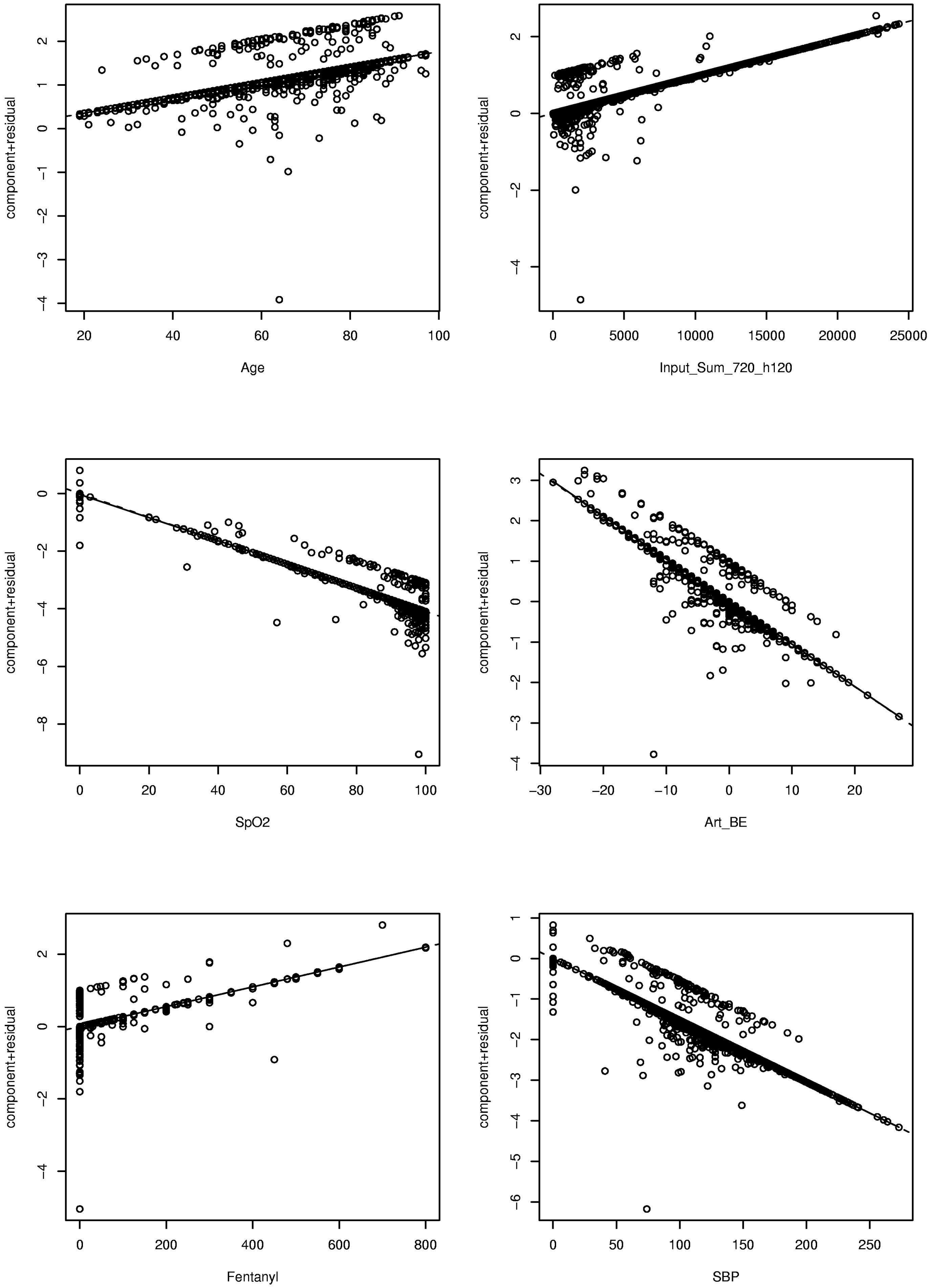


Figure 4-13: Component plus Martingale Residuals

4.1.5 Accelerated Lifetime Model Diagnostics

The accelerated lifetime model did not perform as well as the CPH model. The primary reason for this is likely the CPH's ability to utilize time-dependent covariates. This better performance, combined with the diagnostics indicating that the assumption of proportional hazards appears valid for these data, led us to use the CPH model. All further discussion in this thesis focuses on the CPH model, unless explicitly stated otherwise.

4.1.6 Model Evaluation

In order to evaluate the effectiveness of this model at predicting mortality, individual patients were examined. First, for two randomly selected patient instances—one from a censored patient and one from an uncensored patient—survival curves were estimated. Figure 4-14 shows the comparison of the two patients. It is clear that the censored patient in this case has a much better outlook, which is expected. Considering the uncensored patient only survived 1.6 days after this instance, the survival curve does not drop as quickly as expected; according to the curve, the patient has a $93 \pm 3\%$ chance of survival at 1.6 days in the future.

Next, the amount of self consistency within a patient was examined. To do this, the 10-day survival estimates for consecutive instances in the same patients were analyzed. The same patients shown in Figure 4-14 were used, and their estimates were plotted against the instance indexes. The estimates for the censored patient are shown in Figure 4-15 and the estimates for the uncensored patient are shown in Figure 4-16. To help visualize the trend in the points, the figures include a least-squares fitted line and a smoothed curve that was fitted using local regression.

It is interesting to note that the censored patient has a positive trend in the survival prediction, indicating that the patient is becoming more stable with time. The survival prediction for the patient that died, however, has a decreasing trend. It also appears that the variance in the estimates might be slightly smaller for the censored case than the uncensored case, although it is difficult to tell from these two

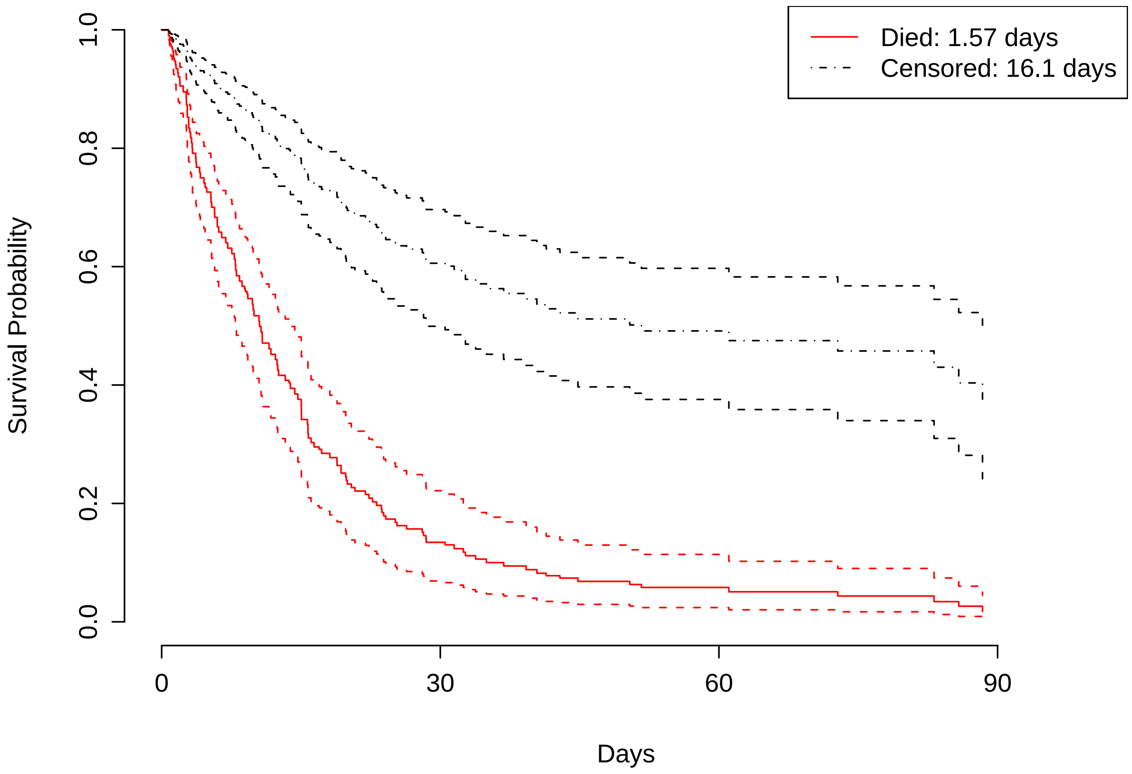


Figure 4-14: CPH Survival Curves for Censored Patient and Dead Patient with error bands

figures.

After looking at some specific differences between these two patients, general differences between censored patients and uncensored patients were explored. To do this, the cumulative mean values for the 10-day survival estimate were calculated for each instance within a particular patient. The cumulative variance was also calculated for each instance in a given patient. The observations noted for the two specific cases generalize to the complete set of patients. Table 4.7 summarizes these important differences. Additional plots of randomly selected patients are available in Appendix B for visualizing some of these trends.

We can also look at the histograms for differences in the distribution of these estimates between the censored patients and the uncensored patients. Figure 4-17

Feature	Mean Censored	Mean Uncensored
CPH 10-day Estimate	0.90	0.81
Cumulative CPH 10-day Est	0.88	0.80
Cumulative CPH 10-day Est Var	0.0057	0.0096

Table 4.7: 10-day Prediction Trends

shows histograms for the 10-day survival estimates, the cumulative 10-day survival estimates, and the cumulative variance of the 10-day survival estimates. In this figure, the first row shows these histograms for all patients, the second row shows these histograms for only censored patients and the final row shows these histograms for only uncensored patients. It is clear from the histograms that are there are slight differences between the survival estimates for censored patients and survival estimates for uncensored patients.

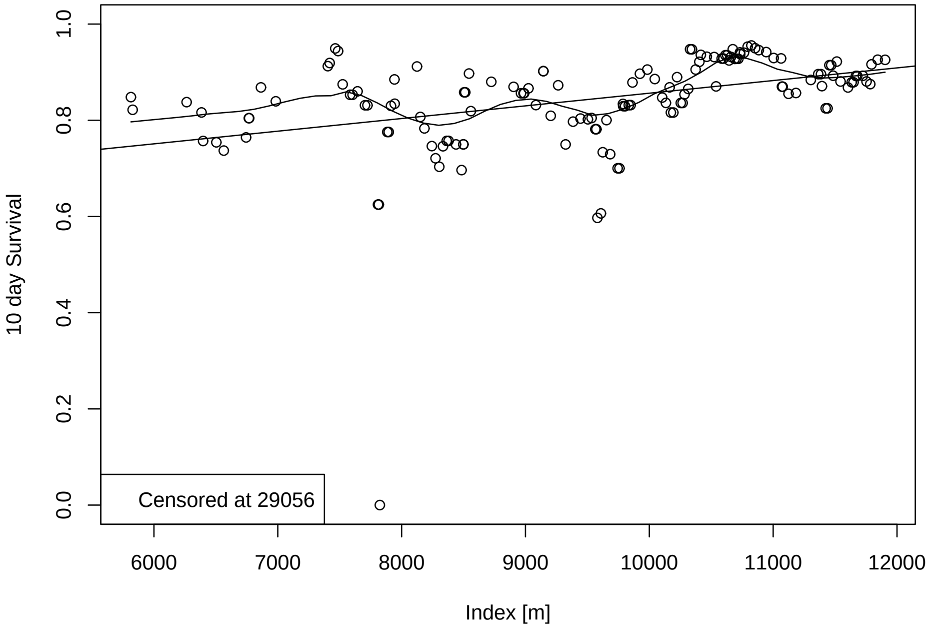


Figure 4-15: CPH 10-day Survival Estimates for Censored Patient

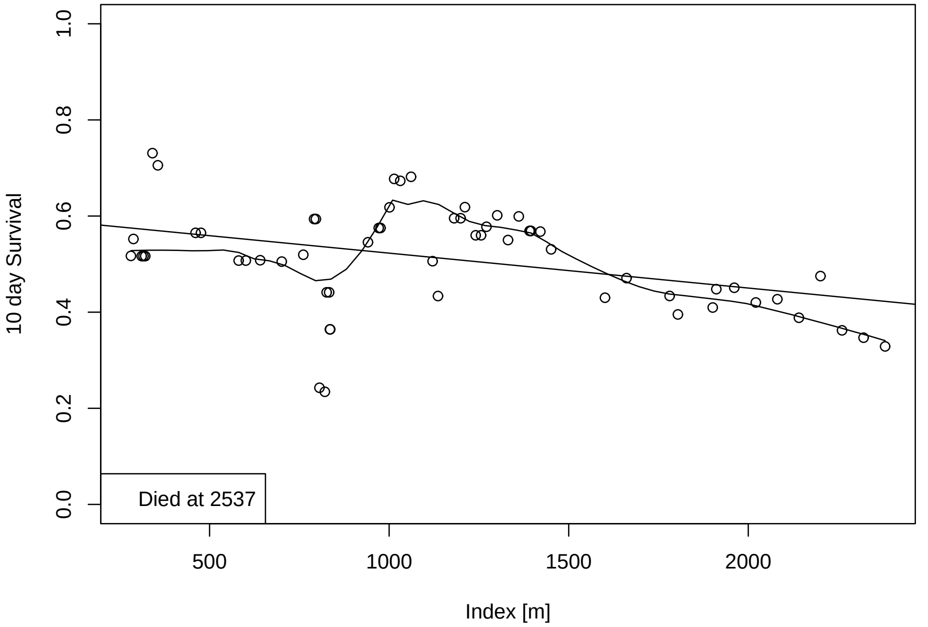


Figure 4-16: CPH 10-day Survival Estimates for Uncensored Patient

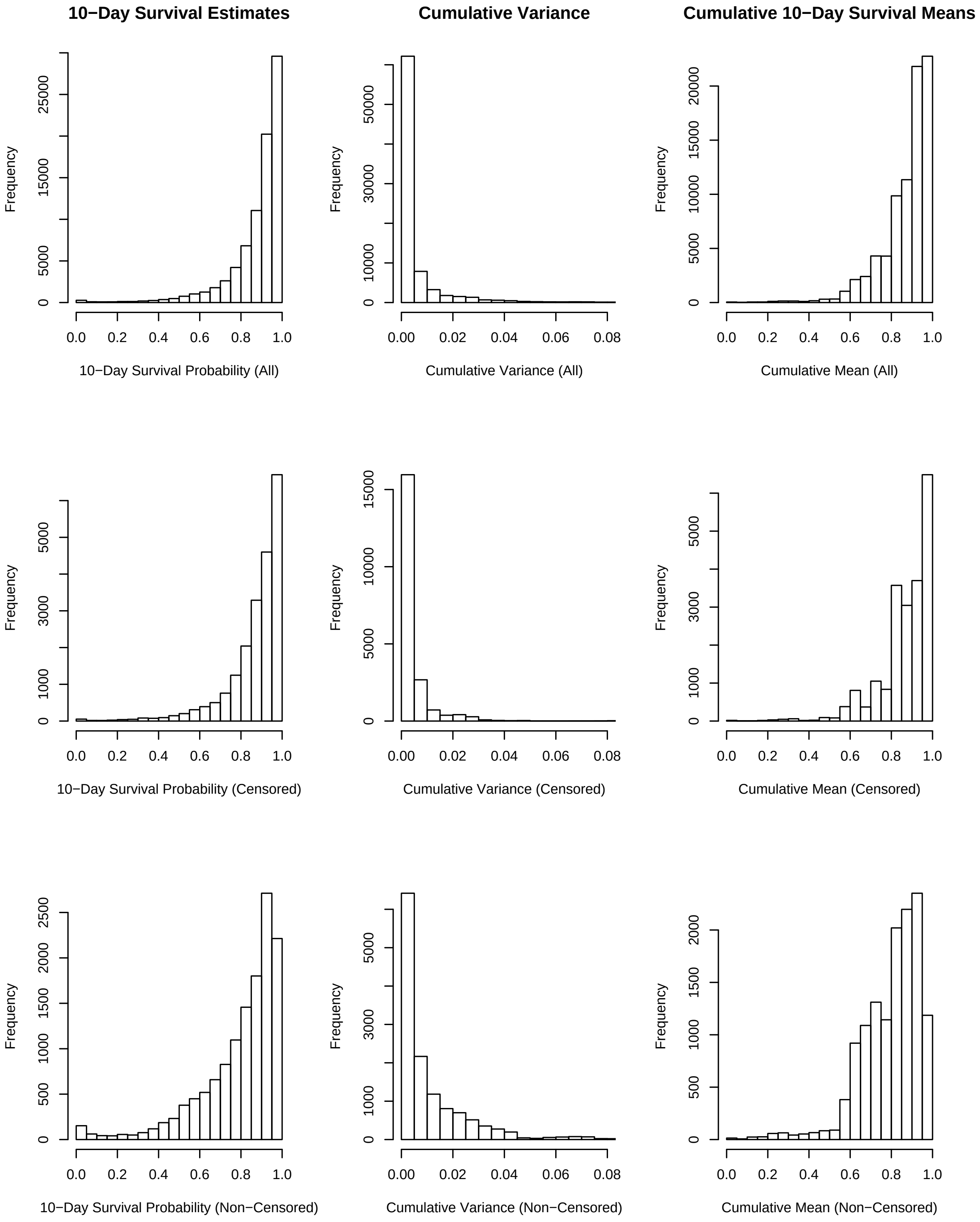


Figure 4-17: Histograms Comparing Censored and Uncensored patients

4.2 Patient Subsets

The small correlations shown in Tables 4.1 through 4.4, combined with the large differences between specialized ICUs, indicate that it might be useful to examine smaller, more specific, subsets of the data. In this section, we examine models trained and evaluated using specific subsets of the data.

4.2.1 MICU Patients

The set of MICU patients was obtained by selecting all instances where the MICU indicator variable (*svMICU*) was set to 1. After preprocessing, a total of 338 patients matched this criterion. Of these patients, 117 died while in the ICU. Table 3.10 shows how these patients were separated between the training set and the test set.

Feature Selection

Again, both forward selection and backward selection resulted in effectively the same features (with the difference being Nitroprusside_h60 vs Nitroprusside). Table 4.8 shows the top 15 features for linear regression on the uncensored data.

Following the same methodology used for the more general set of patients, a CPH model was trained for this set of patients. Using this model, survival predictions were obtained for the training data, and these predictions were then incorporated as features into a new training dataset. Table 4.9 lists these additional features and compares their mean values between the censored group of patients and the uncensored group of patients. In general, the differences between these values are very similar to the differences found for the general set of patients in Table 4.7.

4.2.2 MICU Hypovolemic Patients

For the most specific dataset, the MICU patient set used previously was further reduced to include only hypovolemic patients. These patients were selected based on their ICD9 codes indicating that they were hypovolemic at some point in their ICU stay. Given the number of total patients, the set of patients meeting these two

Feature	R^2 (Cumulative)
Weight	0.07901
Age	0.11670
CV_HR	0.16552
svNSICU	0.19240
Amiodarone_h60	0.21656
Sex	0.23264
Input_Sum_720	0.24718
svCCU	0.25984
Heparin_h60	0.27223
Nitroprusside_h60	0.28398
SBP	0.29334
Levophed	0.30128
Fentanyl	0.30653
Art_PaO2	0.31190
Lidocaine	0.31386

Table 4.8: MICU: Top Fifteen Features using Forward Selection

Feature	Mean Censored	Mean Uncensored
CPH 10-day Estimate	0.81	0.72
Cumulative CPH 10-day Est	0.77	0.71
Cumulative CPH 10-day Est Var	0.0097	0.0120

Table 4.9: MICU: 10-day Prediction Trends

criteria was quite small—25 patients in all. Of these 25 patients, 16 were used for the training set, and 9 were used for testing. Each of these sets included three patients that died, while the remaining patients lived. With only six cases of patient mortality, this dataset was deemed to be too small for useful analysis.

4.2.3 Hypovolemic Patients

The final dataset examined was the set of patients that were marked as being hypovolemic sometime during their stay without the MICU requirement. After preprocessing the data, 60 unique hypovolemic patients were available. Only 10 of these, however, died before they were discharged from the ICU. While over twice as large as the set of MICU hypovolemic patients, it was deemed to be too small for useful analysis.

4.3 Outcome Prediction

One possible way to evaluate patient survival models is to look at the final state of the patients in the ICU. This view of patient survival is clearly violated by numerous patients, such as those that are still severely ill when they leave the ICU for hospice care. These types of cases are expected to make prediction of the patient's final state upon departure from the ICU difficult. Despite these inherent difficulties, predicting which patients die *while in the ICU* is still useful.

4.3.1 General Patients

First, in order to evaluate the model's ability to predict which patients leave the ICU alive, we predict which patients were censored. Figure 4-18 shows the ROC curve for the model's prediction. Comparing this figure to Figure 4-1, it clearly does better than the plain SVM at predicting censoring. This is due to the estimates from the CPH model being appended to the original training set.

Next, several models were trained to predict death within a fixed time window.

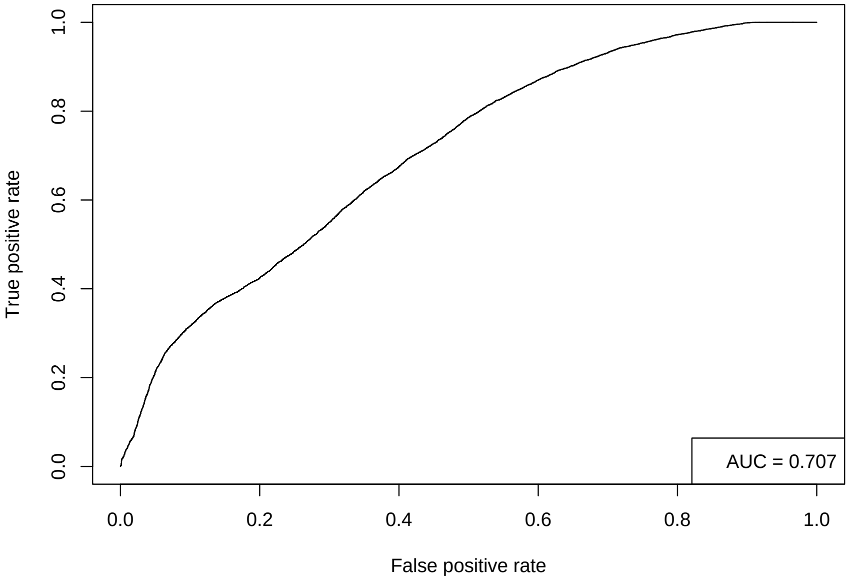


Figure 4-18: ROC Curve for Logistic Regression Censor Prediction

As expected, the shorter this window is, the more effective the model is at predicting patient mortality. Figures 4-19 through 4-22 show the predictive power of these models using their respective time windows. Each of these ROC curves includes the total area under the curve.

As the curves indicate, the survival predictions from these models appear to be reasonable. At 48 hours, for example, a false positive rate of only 0.2 (or a specificity of 0.8) has a corresponding true positive rate (sensitivity) of over 0.5. If the length of prediction is decreased to 24 hours, the false positive rate falls to about 0.1—less than half of the previous value—for the same true positive rate of 0.5.

None of these curves (with the possible exception of the 12 hour prediction model) represent particularly strong predictions. For the 12 hour model, despite its reasonably strong performance with an AUC of 0.877, the usefulness of such a short prediction is likely to be minimal in practice; physicians typically know if a patient is going to die within 12 hours.

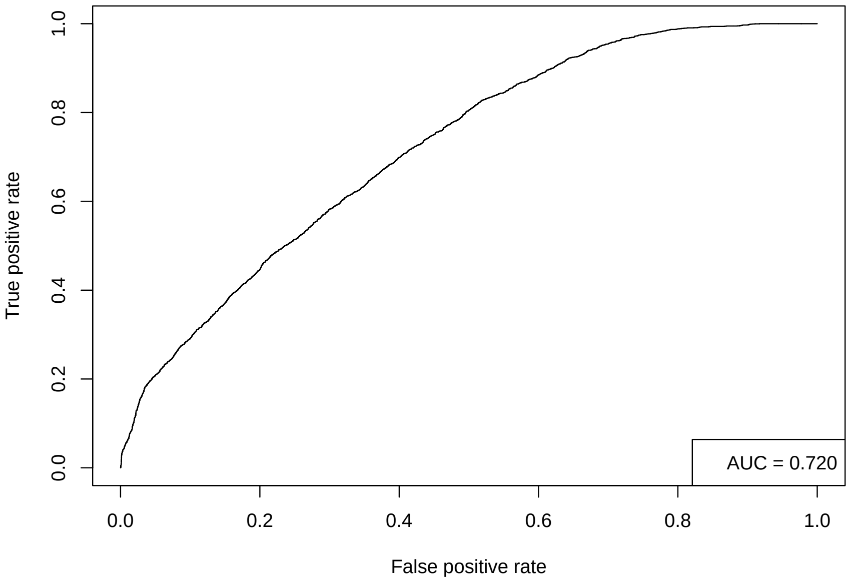


Figure 4-19: ROC Curve for Logistic Regression: 96h Survival Prediction

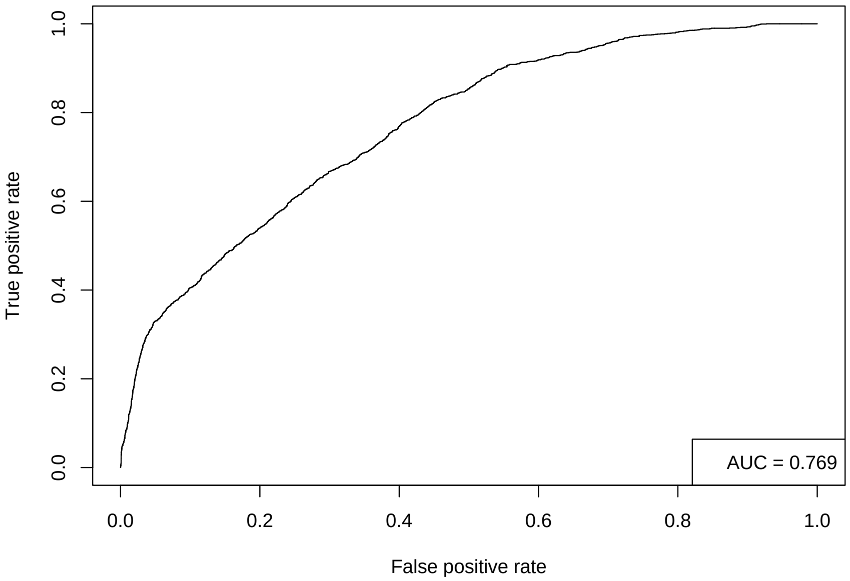


Figure 4-20: ROC Curve for Logistic Regression: 48h Survival Prediction

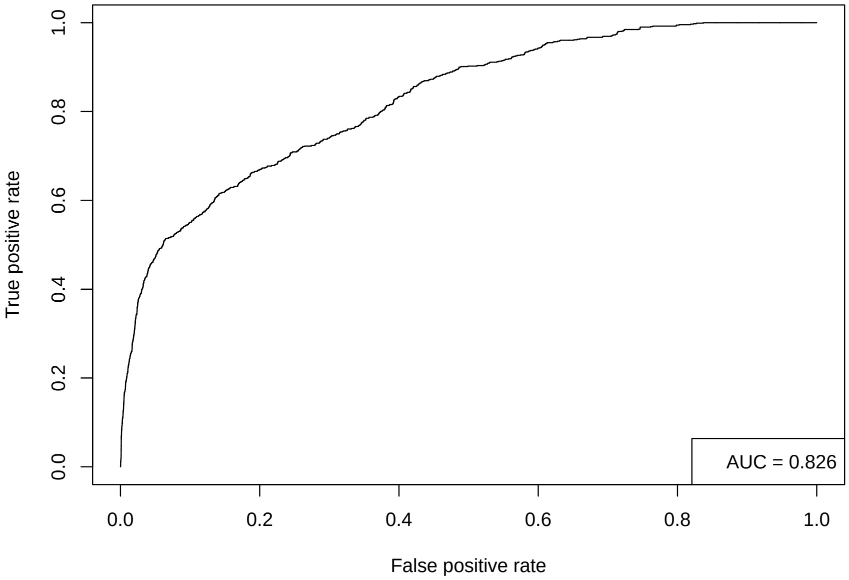


Figure 4-21: ROC Curve for Logistic Regression: 24h Survival Prediction

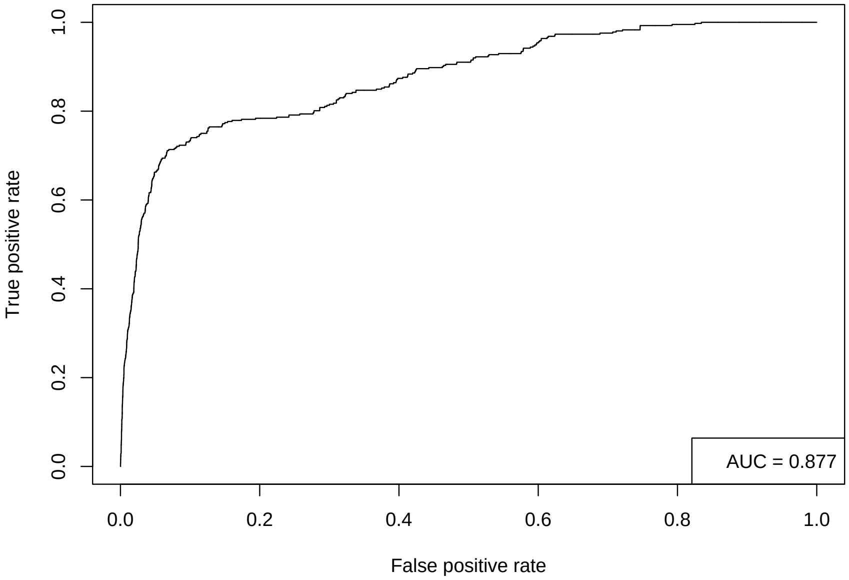


Figure 4-22: ROC Curve for Logistic Regression: 12h Survival Prediction

4.3.2 MICU Patients

Censoring was also predicted using a model created with only a subset of MICU patients. As done with the aggregate set of patients, a test set was held out to evaluate the performance of this model. The ROC curve in Figure 4-23 shows that the logistic regression model for predicting censored patients in the MICU performed slightly worse than the model that was found for the general set of patients (Figure 4-18).

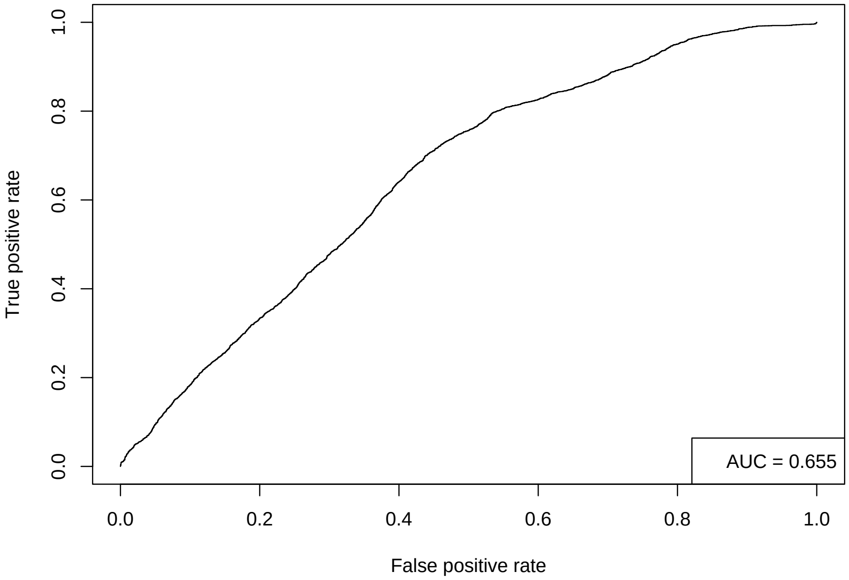


Figure 4-23: MICU: ROC Curve for Logistic Regression Censor Prediction

Finally, using logistic regression models as previously done for the general set of patients, a number of models were created for predicting patient death within fixed windows. The respective ROC curves, generated using the test data, are shown in Figures 4-24 through 4-27. Looking at the area under these curves, it appears that the models for this smaller dataset consistently perform worse than the models that use the more general dataset.

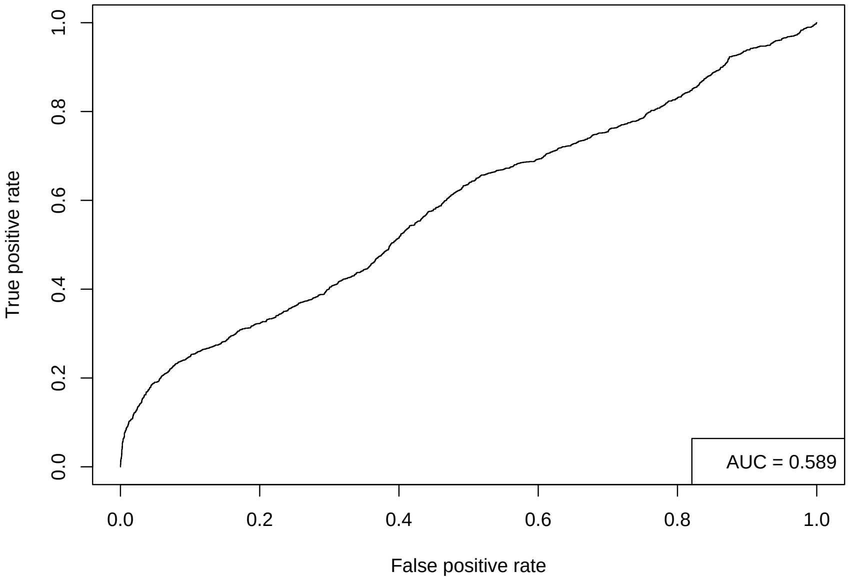


Figure 4-24: MICU: ROC Curve for Logistic Regression: 96h Survival Prediction

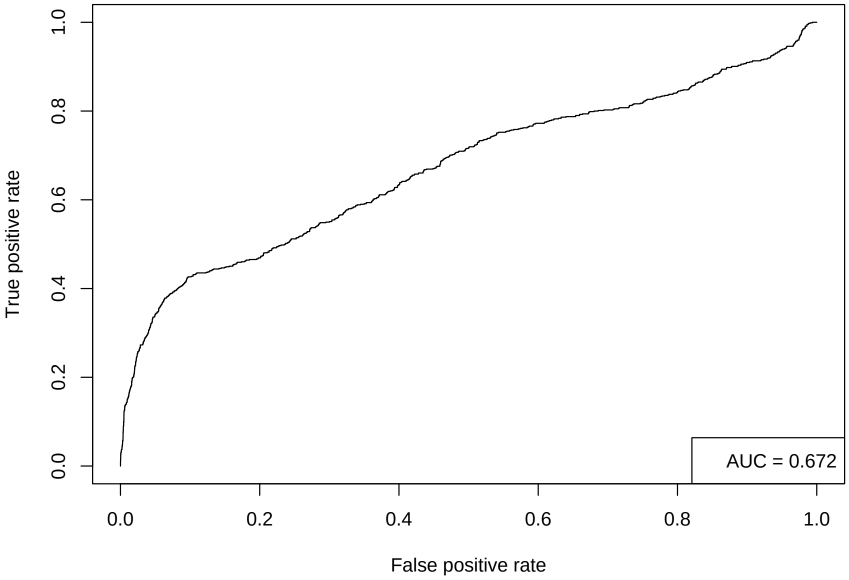


Figure 4-25: MICU: ROC Curve for Logistic Regression: 48h Survival Prediction

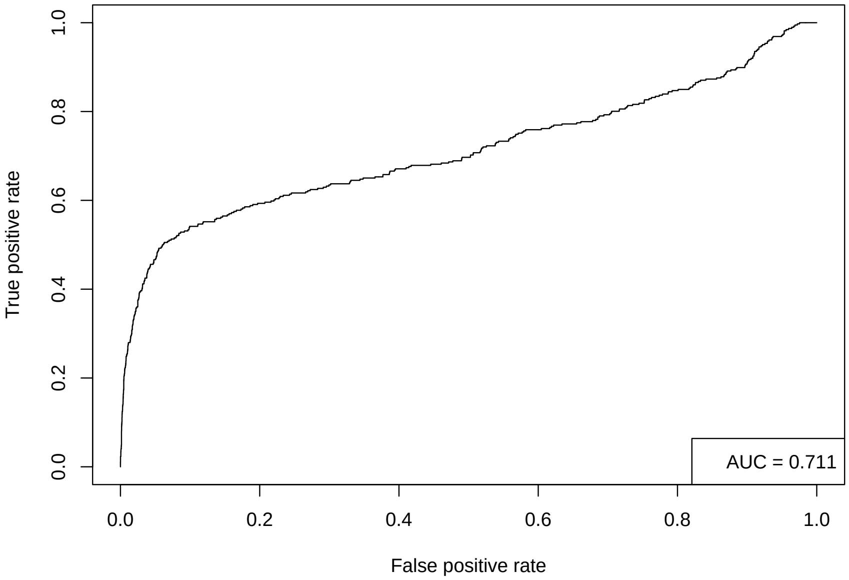


Figure 4-26: MICU: ROC Curve for Logistic Regression: 24h Survival Prediction

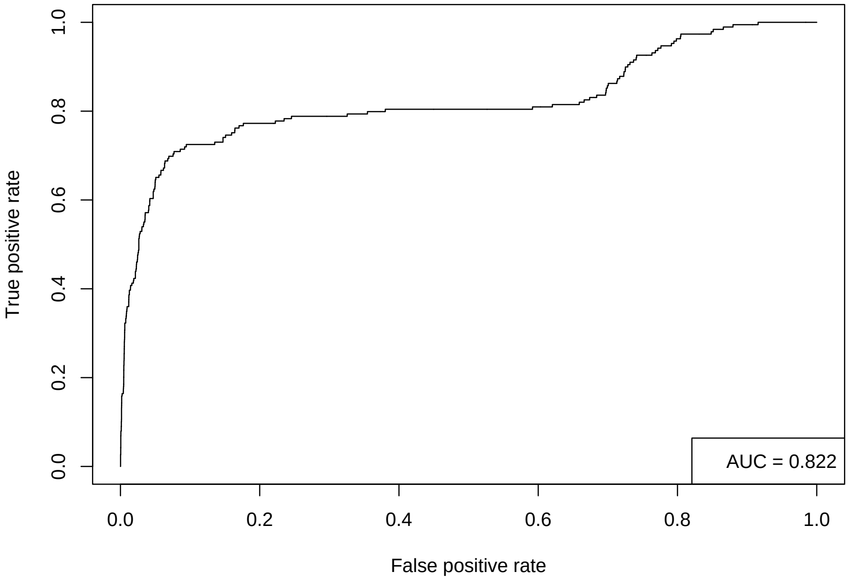


Figure 4-27: MICU: ROC Curve for Logistic Regression: 12h Survival Prediction

4.4 Comparison with SAPS

As a comparison to the above models, we calculated SAPS scores for all of the patients with outcome information available. In calculating these scores, the methodology presented in [33] was followed. Figure 4-28 shows the mortality rate for each individual score, along with the number of patients receiving each score. Although it is important to consider the small number of patients used to calculate the mortality rate for higher SAPS values, it is clear from this figure that there is a strong correlation between an increasing SAPS value and a patient's risk of death.

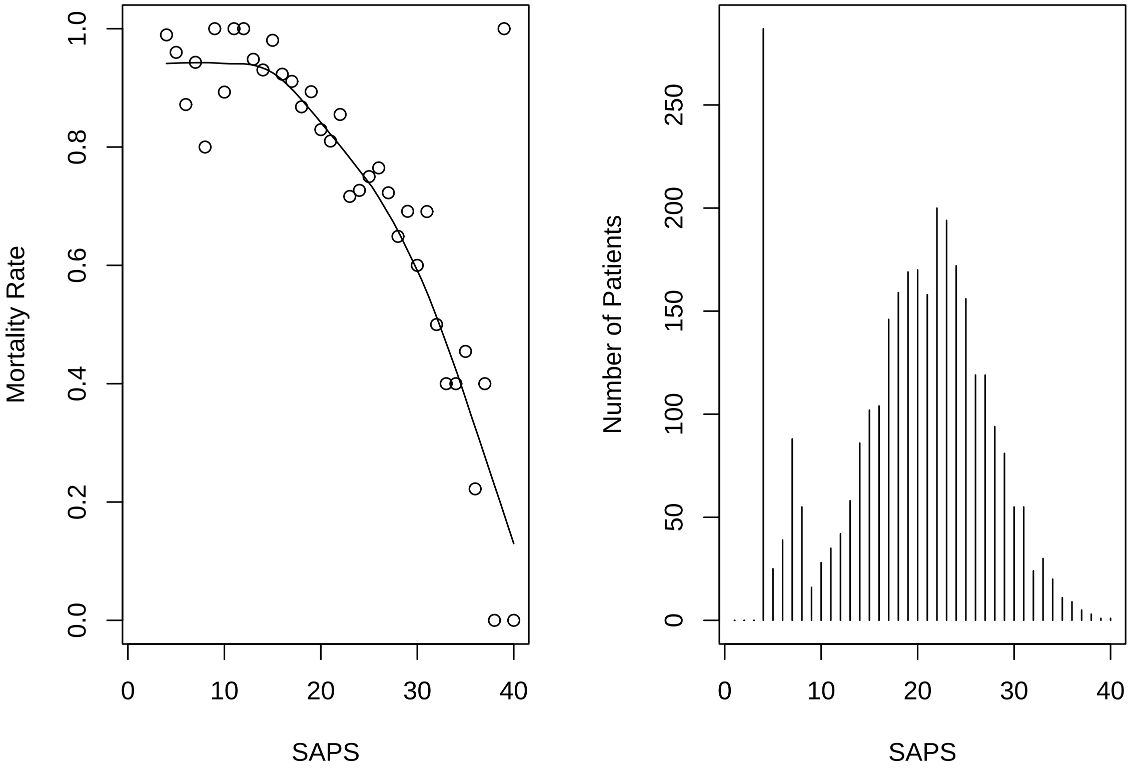


Figure 4-28: Mortality Rate vs SAPS value

As in the previous models, an ROC curve can be created using different cut-off thresholds for SAPS values. This allows us to compare the relationship between *sensitivity* and *1 - Specificity*. Figure 4-29 shows this ROC curve along with the area under it. An area under the curve of 0.731 indicates that the SAPS value does slightly better at predicting mortality than the 96-hour model that had an AUC of 0.720 found for the aggregate set of patients (Figure 4-19). Additionally, it should be slightly easier to predict 96-hour mortality than the final mortality as the SAPS score does. If the 96-hour model is adjusted to look at the final ICU discharge state (resulting in predicting censoring, as in Figure 4-18), then the AUC falls to 0.707. One of the complaints with SAPS is the fact that if the patient dies within the first 24 hours, which is not uncommon, by using the worst score over the period the metric sometimes diagnosis death instead of predicting death.

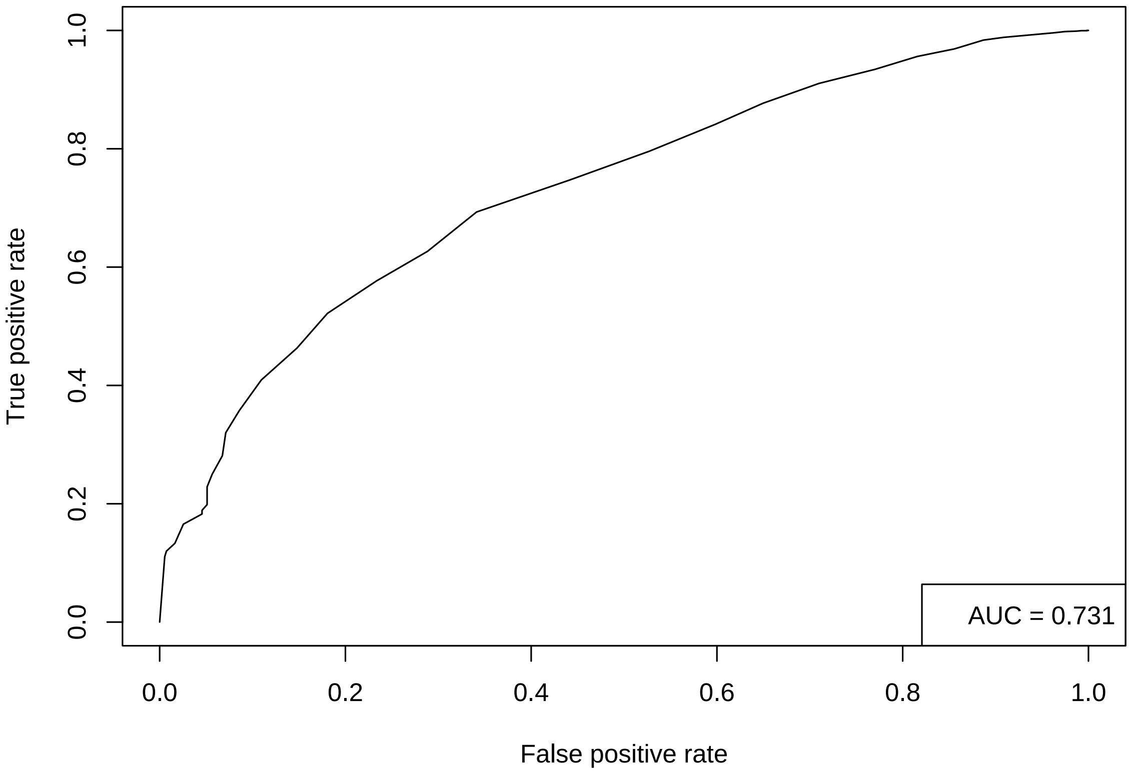


Figure 4-29: ROC Curve for SAPS

The SAPS value for a patient is calculated only over the patient's first 24 hours in the hospital. This constraint can also be placed on the outcome prediction models.

In fact, by adding this constraint, the model's performance appears to improve. As a result of this constraint, the number of unique patients in the test set decreases from 265 patients to 111 patients. Of the 33754 test instances only 2249 of them occur in the first 24 hours. With this reduced test set, the ROC curve (Figure 4-30) has an AUC of 0.768. This value could likely be improved further by allowing the model to utilize the first 24 hours (i.e. using the worst values like the SAPS metric) of each patient's stay instead of only looking at isolated instances.

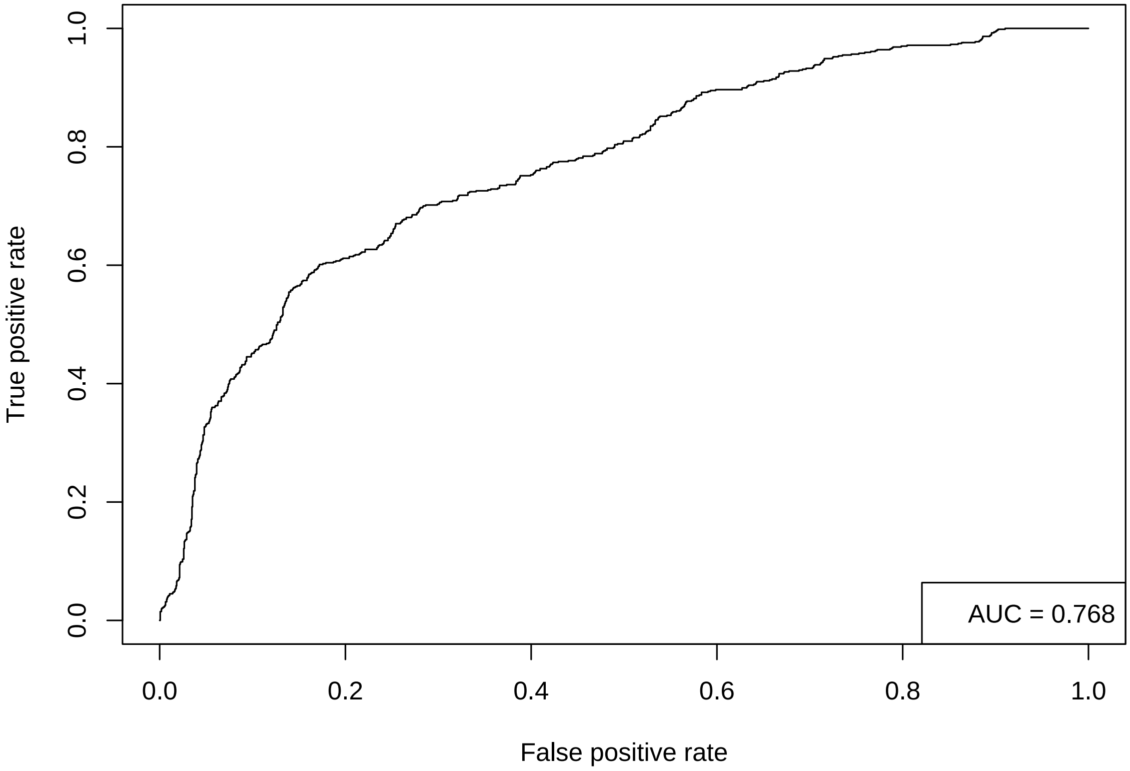


Figure 4-30: ROC Curve for 96h Outcome Prediction (first 24hrs)

It is noteworthy that the model appears to significantly improve when only the first 24 hours of the patient's stay are examined. Under this light, the outcome prediction models suggested in this paper appear to perform marginally better than the SAPS value for predicting patient mortality.

Chapter 5

Related Work

This chapter describes other research that has been done that relates to the goal of this thesis. First, an overview of the many mortality prediction scores is provided. Next, additional survival analysis techniques are discussed. Lastly, a brief discussion of related intelligent monitoring approaches and decision support systems is provided.

5.1 Mortality Prediction Scores

A variety of mortality scores have been developed over the past 25 years. In addition to SAPS, common mortality metrics include the Acute Physiology and Chronic Health Evaluation (APACHE) score [32] and the Mortality Probability Model (MPM) [35]. Like SAPS, each of these scoring systems has undergone multiple revisions [30] [31] [36]. The SAPS II, the APACHE III and the MPM II metrics are routinely used. Each of these can be calculated from published information, but the translation of the APACHE III score into the corresponding probability of hospital mortality is proprietary.

The Apache III and SAPS II models rely on the worst value recorded over the first 24 hours of the patient's stay. Bosman showed that because of this methodology, significantly different outcome predictions are obtained by using an intensive care information system versus manual recordings [2]. Additionally, the APACHE score shares the problem that SAPS encounters when the patient dies during the first 24 hours after their admission—that is, the metric reflects patient mortality rather than predicting patient mortality.

The new revisions of each of the three models (APACHE III, SAPS II, and MPM II) improve upon their earlier versions [7]. Most of this improvement is due to the inclusion of chronic illnesses. Between SAPS II and APACHE III, Moreno found that the SAPS II model slightly outperformed the APACHE III model [42]. But in another case, specific to outcome from acute renal failure, the APACHE II score outperformed the SAPS II and MPM-24 II models [19]. Other studies support the notion that the performance of each model is quite sensitive to the population of patients being evaluated and the population used for calibration.

The MPM accounts for the limitation of being calibrated for the first 24 hours of the patient's stay by introducing multiple models. These models are specialized for different periods in the patient's stay. This includes a model for use at admission, 24 hours after admission, and 48 hours after admission and a model that combines all three of these. For the purposes of the current project, the MPM models were inappropriate, because they utilize several factors in the patient's record that are not easily accessed. Examples of the more difficult features, which would probably require careful parsing of free-text notes to find in the datasets available in our study, include "CPR prior to ICU admission" or "Cancer part of present problem". The complete list of features utilized by the MPM-Admission model are listed in Table 5.1.

Feature	type
Coma (Glasgow 3-5)	Boolean
Emergency admission	Boolean
CPR prior to ICU admission	Boolean
Cancer part of present problem	Boolean
Chronic renal failure	Boolean
Probable infection	Boolean
Previous ICU admission within 6 months	Boolean
Surgical service at ICU admission	Boolean
Age	Continuous
Heart rate at ICU admission	Continuous
Systolic Blood Pressure	Continuous

Table 5.1: Mortality Prediction Models — Admission

Other researchers have looked at the SAPS and APACHE metrics in regards to specific illness. The SAPS metric has been shown to be effective across different intensive care units. In a prospective study, Schuster found that the metric was equally applicable to coronary care unit patients [52]. As another example, Chen and others recently explored the use of the APACHE score to predict prognosis for acute renal failure (ARF) patients. They found that the score, calculated during the 24 hours immediately prior to the onset of ARF (instead of the normal 24 hours after patient admission) was a statistically significant predictor of patient survival [12].

The survival prediction models mentioned above all rely on relatively simple logistic regression. The limitations of this type of model have led some researchers to explore the use of more computationally intensive models that have shown considerable classification power, such as neural networks. Motivated by findings that indicate that binary logistic regression for high-risk patients is relatively inaccurate and inconsistent, Goss shows improvement in patient outcome prediction using a neural network, but concludes that all current prediction systems suffer from high error rates[24]. More recently, researchers have used differentiable approximation to the concordance index (a common survival model quality metric) directly as the objective function for training various classification algorithms [64]. Using this metric with a neural network model, they demonstrate improvement in separating low-risk and high-risk groups of patients.

Other researchers have focused their research on simple models derived from more specific indicators of patient mortality. One recent study by Smith *et al* has shown that arterial base excess and lactate concentrations by themselves each have strong prognostic power at the time of patient admission [55]. Using base excess and lactate individually, they obtained an area under the ROC curve of 0.73 and 0.78 for each, respectively. Furthermore, they showed that at 24 hours into the patient's stay, the predictive ability of base excess is slightly decreased while the predictive power of lactate marginally increases. They also point out that while the lactate value is predictive of mortality, this increase in risk is dependent on the cause of the rise in lactate.

Physicians still generally outperform these static mortality models. A recent review of by Sinuff *et al* compared the predictions from these models to the predictions of physicians using a large set of published literature [54]. This research found that the physicians outperformed the scoring metrics during the first 24 hours after ICU admission. They reported an area under the ROC curve of 0.85 ± 0.03 for physicians versus an area of 0.63 ± 0.06 for the scoring metrics. A study by Rocker *et al* found that mortality estimates lower than 10 percent by physicians "strongly influence patterns of life support provision and limitation and vary in impact according to the severity of organ dysfunction and the presence or absence of preferences to limit life support" [47]. This observation could explain some of the difficulty of predicting fixed survival windows encountered in this project.

5.2 Survival Analysis

In addition to the most common survival analysis techniques described in Chapter 2, several additional survival models have been developed. One of the more interesting of these is the Buckley-James regression method. There has also been considerable work recently focused on informative censoring.

5.2.1 Survival Regression

Buckley-James regression is one of several techniques for survival regression [5]. It is based on the linear relationship between the expectation of the survival time and the covariates,

$$E[\delta_i Y_i + (1-\delta_i)E(Y_i|Y_i > t_i)] = \beta_0 + \boldsymbol{\beta_1}^T\mathbf{x}. \tag{5.1}$$

It is clear from this model that if all of the data are uncensored, then the relationship is simply an ordinary least squares model for survival time.

Equation 5.1 shows that the censored data points can be replaced by their expected values. This can be done without biasing the regression equation. The Buckley-James estimator uses this idea by replacing the variable y_i with the Kaplan and

Meier estimator. The solution is then obtained by solving the ordinary least squares equations iteratively. In the end, this technique effectively replaces the censored data points with their expected values and then proceeds to minimize the sum of squares. It repeats this process until the procedure converges, or gets trapped in a loop.

Miller and Halpern concluded that the Buckley-James estimator performed better than the two other most common linear regression schemes for survival data (Miller's estimator and Koul, Susarla & Van's estimator)[41]. By using the standard Standford heart transplant dataset, they found that the Cox and Buckley-James regression methods both performed comparably. This indicates that in cases where the assumption of proportional hazards is clearly inappropriate, Buckley-James might be a strong alternative. However, additional research has indicated that the Buckley-James regression technique is weak under heavy censoring (i.e. greater than 60 percent) [26]. Cox's proportional hazards model is preferable when heavy censoring is present and the R^2 correlation is below about 0.55 [57]. Considering these criteria, the Cox proportional hazards model is a reasonable choice for this project.

5.2.2 Informative Censoring

Finally, additional survival analysis research has been directed to the case where there are multiple risks that could cause the event of interest. For example, a patient might die from a myocardial infraction that is unrelated to a specific illness being treated in the ICU. Conventional survival analysis looking, at survival from the illness under consideration, would have to consider this case censored. This would be incorrect as the patient is no longer at risk of the event of interest.

A *competing risks* framework can include this as a separate "competing risk" [16][20]. In fact, a patient being discharged alive from the ICU can also be considered as a competing risk [46]. This method addresses the assumption that the censoring is non-informative. Using *competing risks* has an intuitive appeal for use in the ICU, as patient withdrawal from the ICU is generally a result of deterioration or improvement. A free *R-project* package, *cmprsk*, is available for creating these models [25].

There are many other techniques related to *competing risks.* Many of these attempt to infer the portion of patients deemed "cured" and model the dependency of the censoring on the covariates. Some recent work illustrating these techniques can be found in [37] and [13].

5.3 Intelligent Monitoring and Decision Support

Since the beginning of the field, researchers have been trying to apply artificial intelligence to assist medical practitioners. Numerous systems have been developed. Many have shown promise, but few have been utilized. The successful ones tend to be very focused in nature and fit easily within the existing health care work flow. Pople argues that much of the difficulty comes from the inherently ill-structured nature of medicine [58].

Medical decision support has followed recent interest in utilizing temporal data in intelligent systems. This allows systems to utilize the temporal information that is often included in definitions of various medical conditions. For example, [39] describes utilizing constraints specific to cardiology to enhance reasoning in the Heart Disease Program (HDP). Augusto's recent review article, *Temporal reasoning for decision support in medicine* [1], provides a nice overview of this area, and highlights work that he feels needs more attention. Specifically, he points out that the area of medical prognosis has received less attention than diagnosis and therapy planning.

5.3.1 Signal Artifact Detection and False Alarm Reduction

The proliferation of false alarms in the ICU continues to be a concern as the number of monitoring devices increases. In [62], Tsien identifies some of the more problematic alarms and explores some of the causes of these alarms. The specific nature and generally poor performance of these alarms has led many researchers to focus on trying to reduce them. While more specific than the work in this project, many of the same tools apply and similar problems arise.

Several researchers have worked to address problems with signal quality. One approach, suggested in [66], validates arterial blood pressure alarms by examining the signal quality and utilizing the relationship between the arterial blood pressure and the electrocardiogram.

Other approaches have built more complicated models for validating the signals. One type of model that shows particular potential in biomedical signal processing is the Hidden Markov Model (HMM) [15]. HMMs continue to be foundational to much of the recent progress in speech recognition applications, and much effort continues to be dedicated by the the speech recognition community to enhance these methods. Novak *et al*[43] compared an HMM and another common speech recognition method, Dynamic Time Warping (DTW), for classification in arrhythmia analysis, intracranial pressure monitoring, and electroencephalogram monitoring. They found that these models could potentially be used to expedite the analysis of these signals.

5.3.2 Patient State Identification

An important alarm ideally reflects instability or risk of instability in a patient. Most of the patient state-models medical practitioners consider are subordinate to specific disorders such as hemorrhagic shock. Looking more generally at patient state (stable vs unstable), however, has received less focus.

Hidden Markov models have been used to try to gain insight into the underlying hidden state of a patient. Using various sets of variables, Brause attempted to fit an HMM to a set of sepsis patients [3]. While he obtained reasonable prediction values using this model, the identification of underlying sepsis states was inconclusive. With a similar objective, other researchers have also aimed at characterizing patient state from time series data using a window-based decision-tree model that includes temporal trend information [6]. The results from this work have been limited.

Chapter 6

Conclusion

In this thesis, we have suggested a methodology for using survival models to gain insight into patient state and trajectory. To conclude this thesis, we first provide a summary of the thesis. We then highlight areas for future research.

6.1 Summary of Contributions

To begin with, in chapter 2 we discussed some of the relevant background for this thesis. This included an overview of survival analysis where we provided necessary definitions. This chapter also included a brief discussion of the SAPS mortality metric. The final piece of background information described in Chapter 2 was various classification algorithms used in this thesis.

In chapter 3, we described the data used for this research. The chapter begins with a description of the MIMIC II project and the various data sources included within that project. Next, we stepped through the preprocessing techniques utilized to prepare the final dataset. We also explained how annoyances such as missing values were handled. The chapter concluded with a summary of the final datasets.

Chapter 4 discussed the models we created. This chapter described modeling the entire set of patients available as well as modeling a smaller subset of the patients that are in the MICU. For each of these datasets, the method used for feature selection was described and the resulting models were discussed. While identification of underlying patient state was inconclusive from these models, interesting trends were observed. Next, utilizing these trends, several models were developed for predicting patient mortality. The chapter concluded by comparing the discriminatory power of these models to mortality prediction from the SAPS metric.

In chapter 5 we discussed some of the related work that other researchers have done. We started by providing a general overview of the various mortality prediction scores. Next, alternative survival analysis methods were discussed. The chapter concluded with a brief discussion of intelligent monitoring systems and decision support.

6.2 Future Work

There are several open questions raised by this work that warrant further investigation. First, a closer analysis of the survival predictions over the course of a patient's stay would be interesting. Secondly, additional features could be added in an effort to improve the predictiveness of the models. It would also be interesting to explore survival analysis methods that can account for informative censoring. Finally, it would be worthwhile to consider these results with a larger dataset. In this section we briefly discuss each of these ideas in order.

From the results shown in chapter 4, it was clear that there was quite a bit of variance in the survival predictions over the course of a patient's stay. At the time of this writing, many of the discharge summaries for these patients were unavailable as they were being deidentified. With these discharge summaries and the corresponding nursing notes, it would be interesting to attempt to identify if the significant highs and lows in these plots reflect meaningful information about the patient's state at corresponding times.

Additional features would likely allow for improved model performance. One of the major criteria for selecting features in the models in this thesis was frequent availability. Many of the most predictive features, however, such as lactate and other lab results were not used because they are only measured once per day. Extracting additional medications (non-intravenous) from the nursing notes might also prove useful to the outcome models. Another possibility for additional features would be to use wavelet features that capture trend dynamics as suggested by Saeed in [50].

Exploring more sophisticated survival models might also be interesting. A starting point would be to further explore models based on *competing risks*. Initial attempts to create *competing risks* models were unsuccessful due to extreme computational costs. These models can also be enhanced by using a mixture model that accounts for informative censoring.

Finally, using additional data would allow more specific models to be designed. The set of patients used in this thesis was a relatively small subset of the patients available in the MIMIC II database. For the purposes of this work, the data were specifically limited by a small set of a patient outcomes. Additional patient outcomes are expected to be available from the hospital shortly.

Additional data could also be obtained by using various imputation methods. Imputation is commonly used in statistics to replace missing data points with valid values. In preparing the datasets used in this thesis, instances that were missing a value for any of the selected features were omitted. This reduced the number of patients by more than half. Using imputation, many of these patients could be kept by filling in the *NA* data points with estimated values. Several techniques are available for estimating missing values. Some of the more common methods include mean substitution, simple regression, regression with an error term, and the expectation maximization (EM) algorithm. The size and sparsity of the data used in this project make applying these techniques challenging, but tackling this challenge could potentially result in more data being available for modeling and evaluation.

Appendix A

Trend Data Preprocessing Assumptions

A.1 Matching Case files to Patients

The trend data for each patient are contained in a separate case file. The case files are encoded using the *Physionet wfdb* format [22]. The starting times for the trend files were extracted using the *wfdbdesc* utility and the signals were extracted to free text using the *rdsamp* utility.

The first challenge in using the 1-minute trend data is to align the case file containing the monitor output with the correct patient. Each monitor assigns a case identifier (CaseID) to its output. An algorithm has been developed to match the CaseID to the patient identifier (PID) by looking for values in the trend files and the CareVue ChartEvents table that correspond. The main difficulty in this process arises in the common case where one PID has multiple corresponding CaseIDs. These individual cases, which sometimes overlap, need to be merged together.

The following strategy was used for merging multiple cases that match a given patient. First, the starting time and the duration were extracted for each case. Using the first segment—or in the case of a tie, the longest—the segments were merged together such that values from newer cases replaced older values. If the cases do not overlap, then missing values were inserted for the missing segment. Using this technique, one file containing all the appropriate trend values is made for each patient. This file has instances available at 1-minute intervals for the patient's entire ICU stay.

Additionally, there are some cases where multiple files exist for one CaseID. In this case, the longest file was used.

A.2 Filtering, Missing Values, and Derived Features

For preprocessing the 1-minute trend data, two methods were used. First, a median filter was applied to each feature. A simple method for handling missing values was also used.

For the median filter, a 3-minute window was used. This was applied to each of the features to reduce the noise. In many cases the signals still had problems. For example, it was necessary to apply the data validation rules described in section 3.2.1 in Chapter 3 to verify that the systolic blood pressure was greater than the mean blood pressure and that the mean blood pressure was greater than the diastolic blood pressure.

For instances with missing values, the last valid value within a 30-minute window was used for each feature. If no valid values were present in the preceding 30 minutes, then the value was left as missing. In other words, data values were held for 30 minutes after the signal dropped. This allowed small blocks of missing features to be filled in with the value that was last known to be good. In some cases, where the signal dropped for longer periods of time, this simply resulted in 30 repetitions of the last valid value before the signal contains missing values.

A number of derived features were calculated for signals. These included the *mean*, the *slope*, the *standard deviation*, the *minimum*, the *maximum*, and the *sum*. For each of these calculations, a window indicating the number of consecutive instances to use, was provided. Of the derived features, the *slope* and a short *mean* seemed to provide the most insight into patient outcome. While not discussed in this thesis, several of the other features were explored but ultimately ignored in order to keep the size of the feature set at a reasonable size.

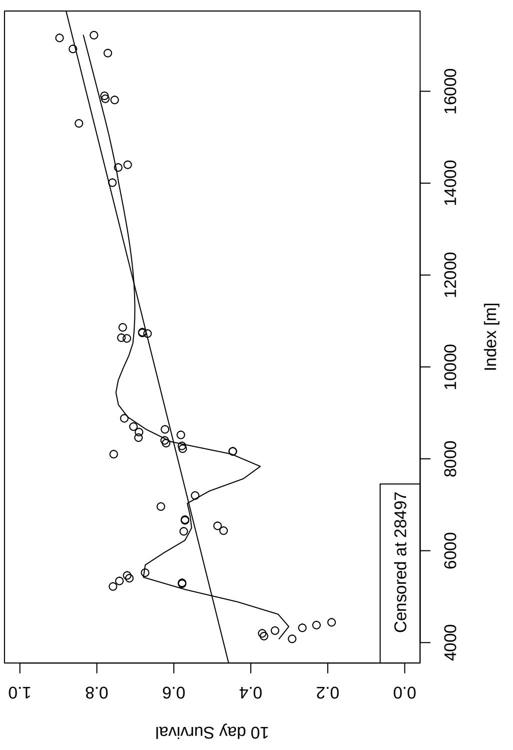
For calculating these features, the windows at the beginning and end of a patient's ICU stay were shortened as necessary. For the slope calculation, we used the coefficient of the least squares regression line over the given window. Missing values were ignored in all of the derived feature calculations.

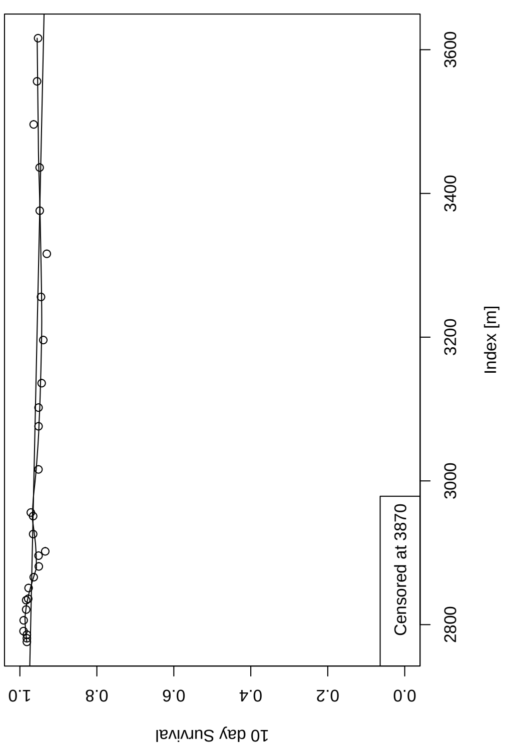
Appendix B

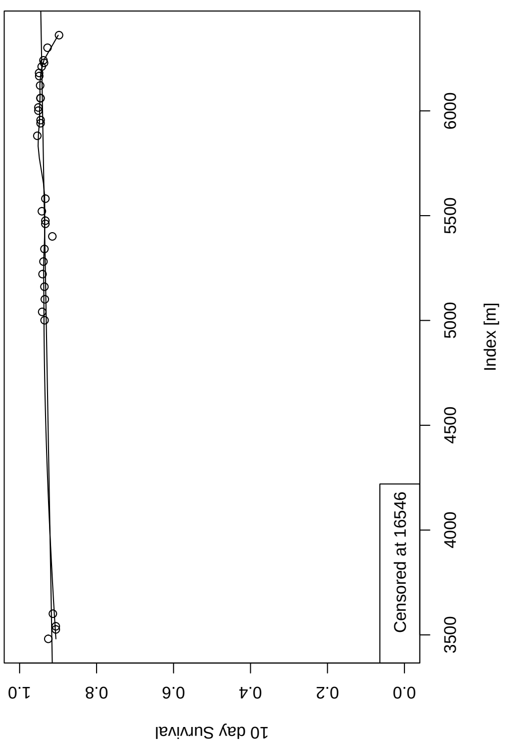
Additional Patient Time-varying Survival Estimates

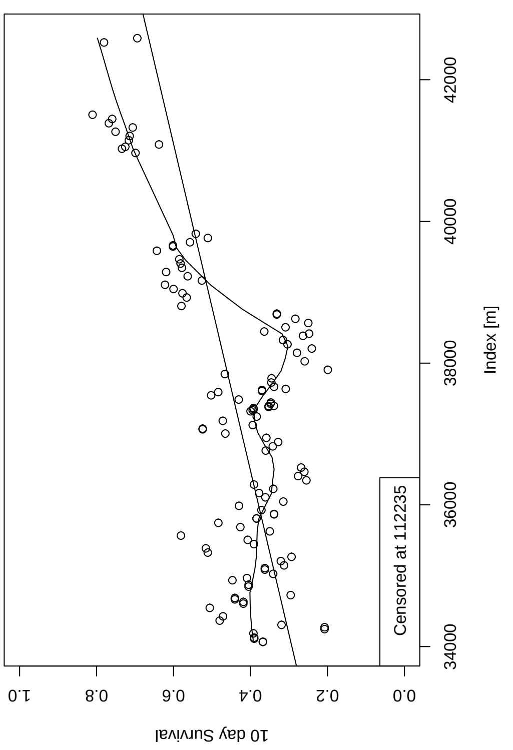
B.1 Censored Patients

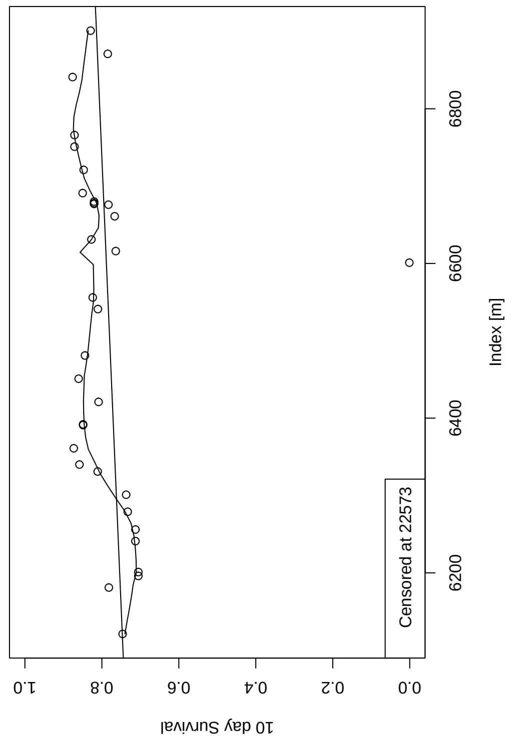
This section provides additional graphs showing the 10-day survival estimate as a function of time for 20 randomly selected censored patients. Each of these patients was selected from the test set. The plots include two fitted lines. One of these lines is fit using least squares regression and the other is a smooth line fit using local regression (the *loess.smooth* routine [45]). For the smoothing, a span value of 0.25 was used along with the default polynomial degree of 1.

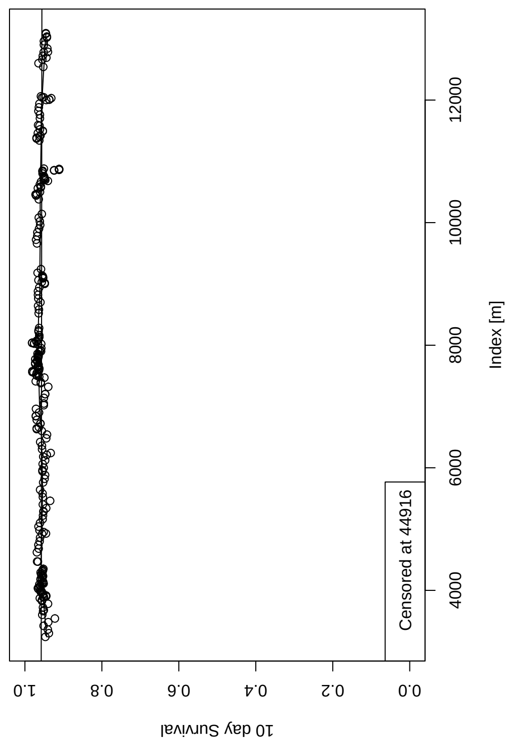

Censored at 28497
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
4000
6000
8000
10000
12000
14000
16000

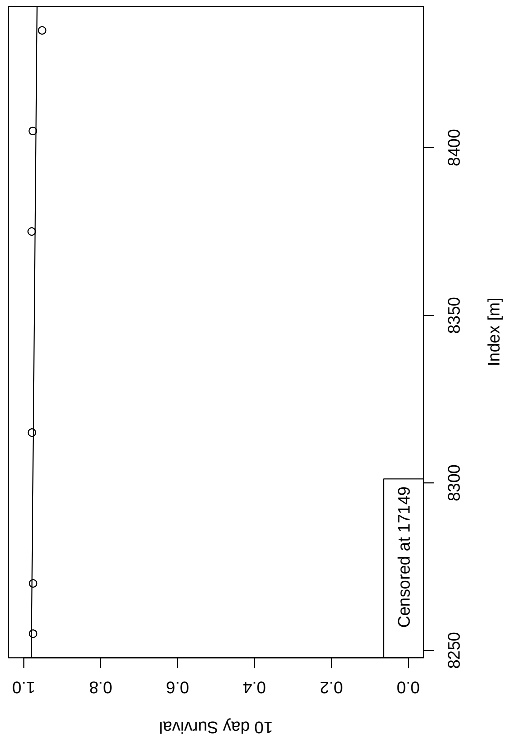

Censored at 3870
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
2800
3000
3200
3400
3600

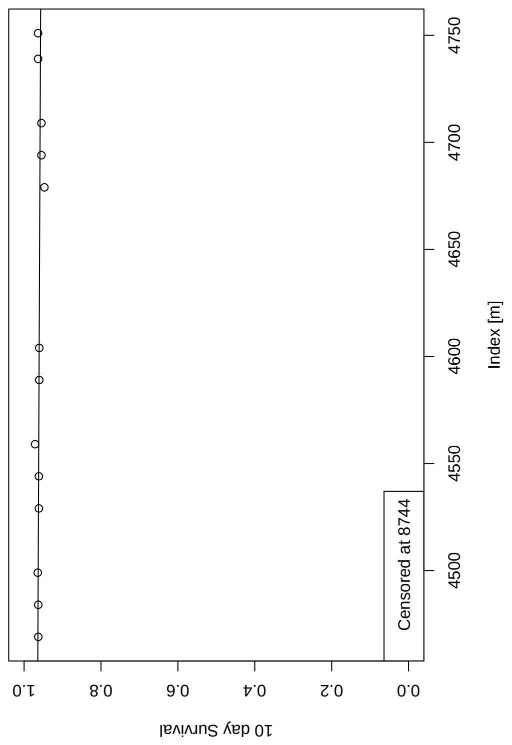

Censored at 16546
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
3500
4000
4500
5000
5500
6000

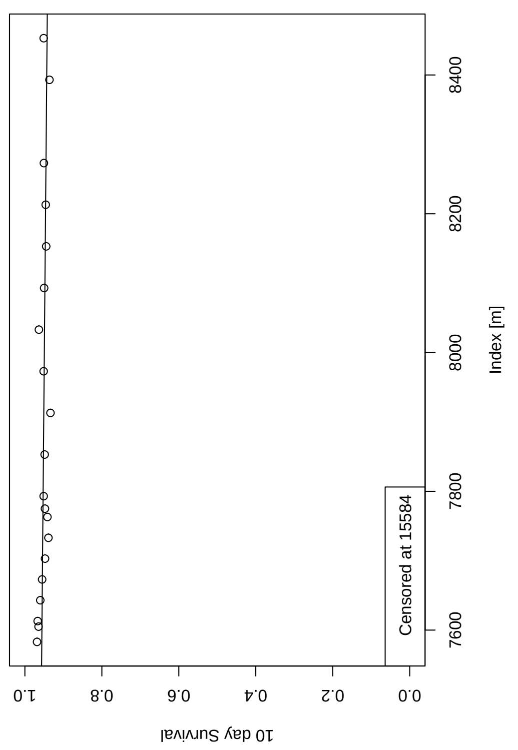

Censored at 112235
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
34000
36000
38000
40000
42000


10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
6200
6400
6600
6800
Censored at 22573

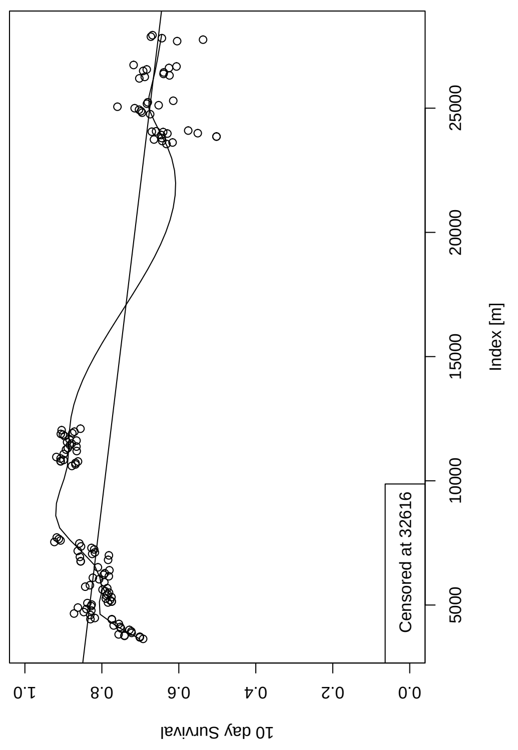

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
4000
6000
8000
10000
12000
Censored at 44916


10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
8250
8300
8350
8400
Censored at 17149

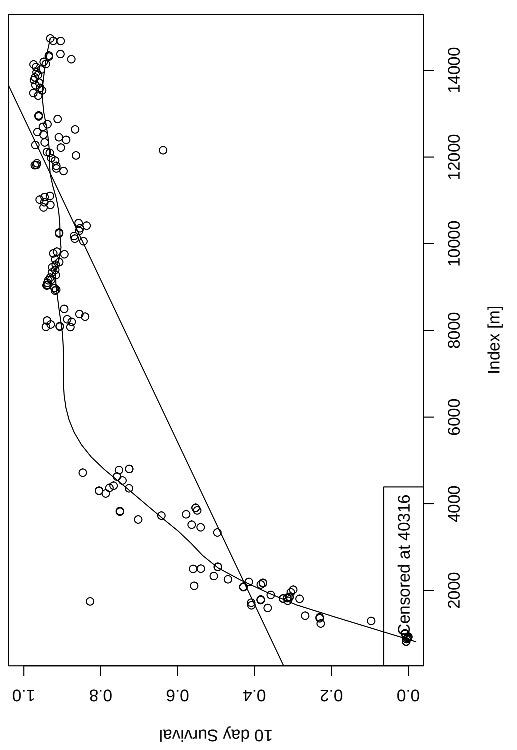

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
4500
4550
4600
4650
4700
4750
Censored at 8744



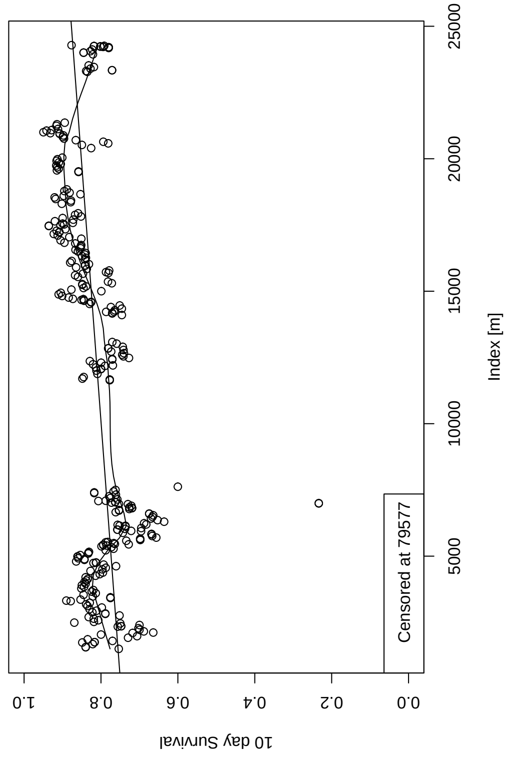
Censored at 15584
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
7600
7800
8000
8200
8400



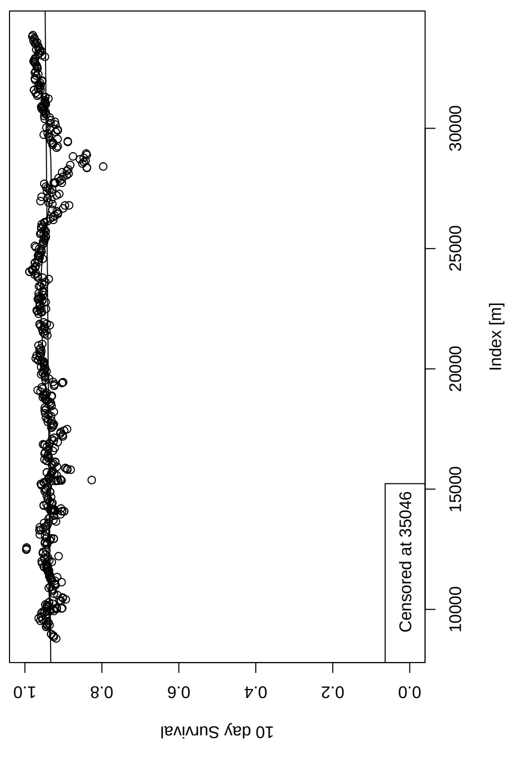
Censored at 32616
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
5000
10000
15000
20000
25000

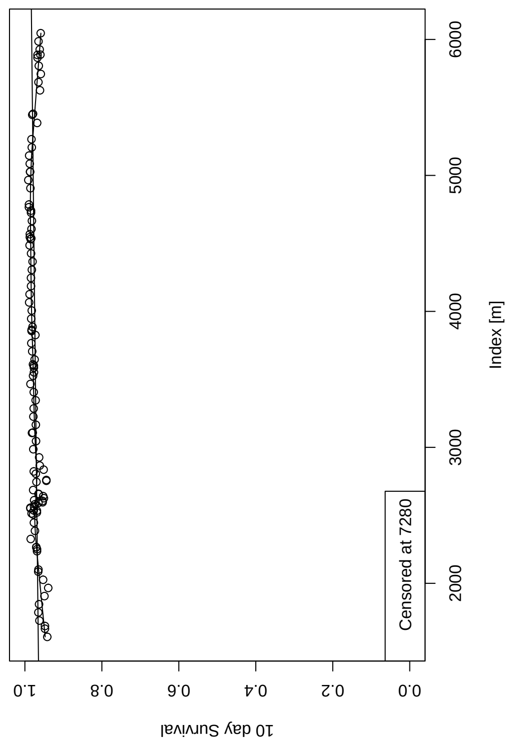


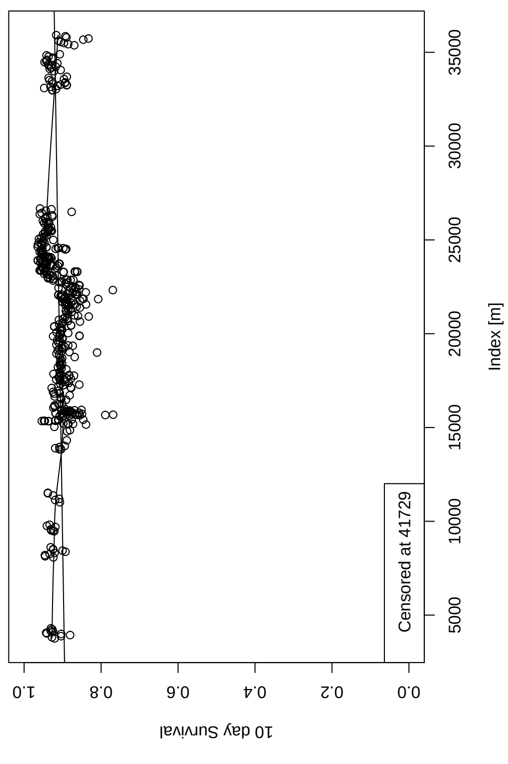
Censored at 40316
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
2000
4000
6000
8000
10000
12000
14000

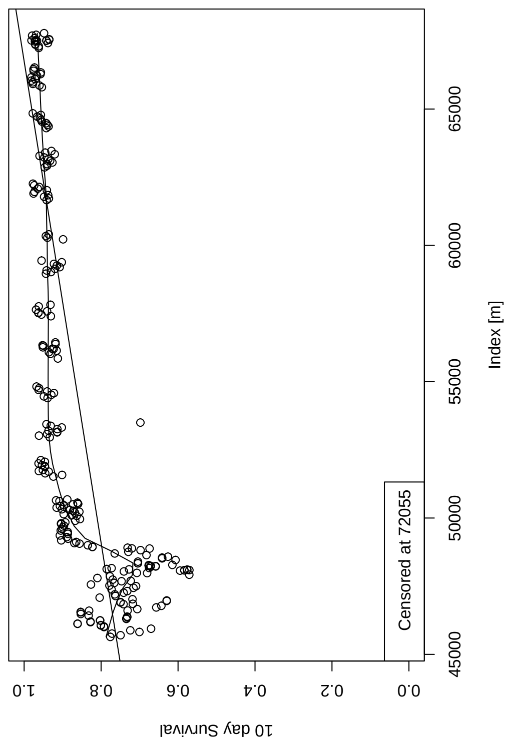


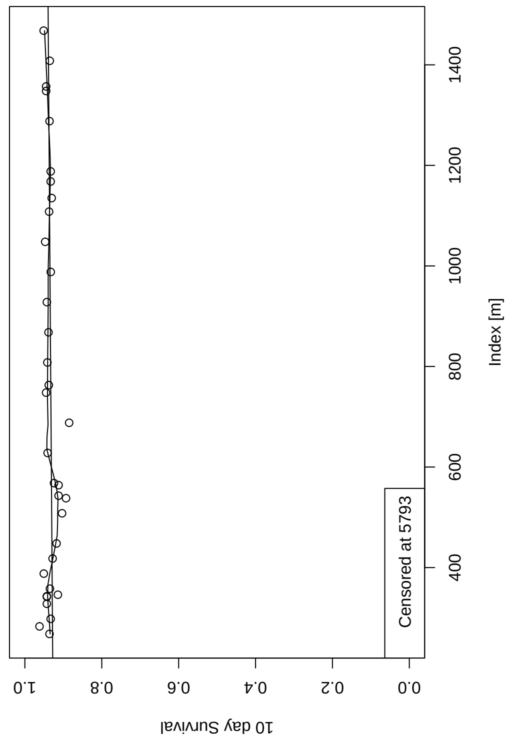
Censored at 79577
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
5000
10000
15000
20000
25000

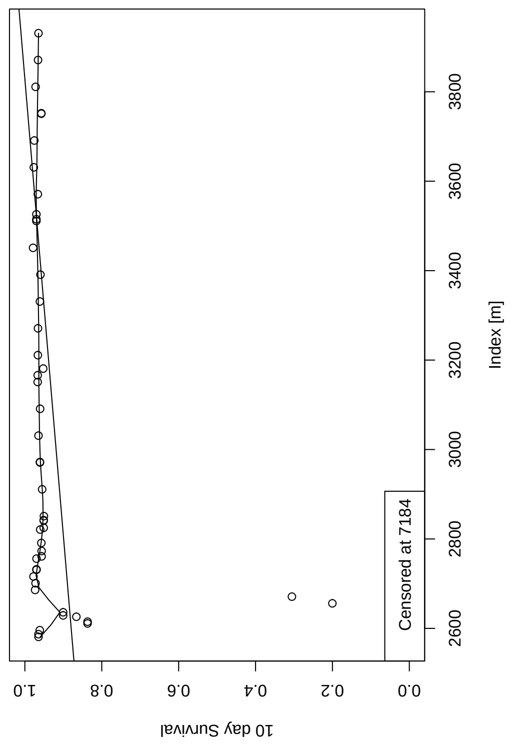

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
10000
15000
20000
25000
30000
Censored at 35046

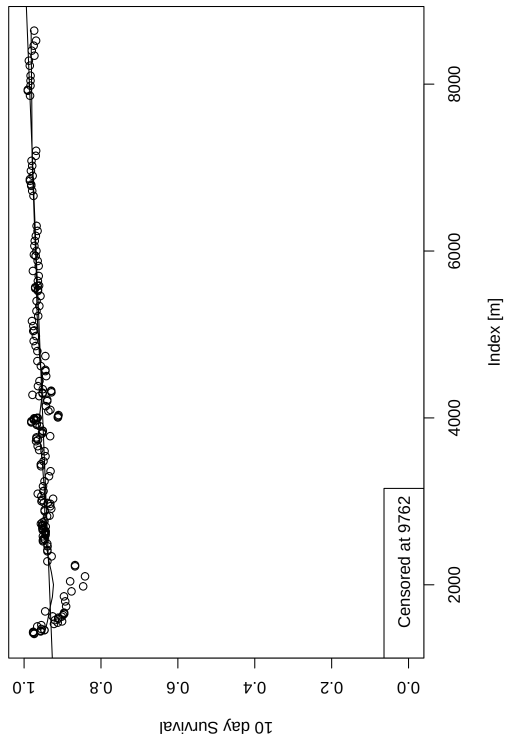

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
2000
3000
4000
5000
6000
Censored at 7280

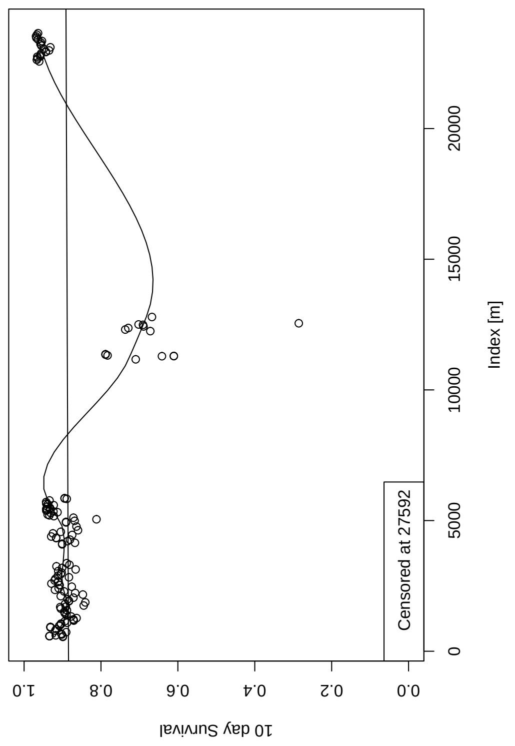

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
5000
10000
15000
20000
25000
30000
35000
Censored at 41729


10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
45000
50000
55000
60000
65000
Censored at 72055


Censored at 5793
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
400
600
800
1000
1200
1400

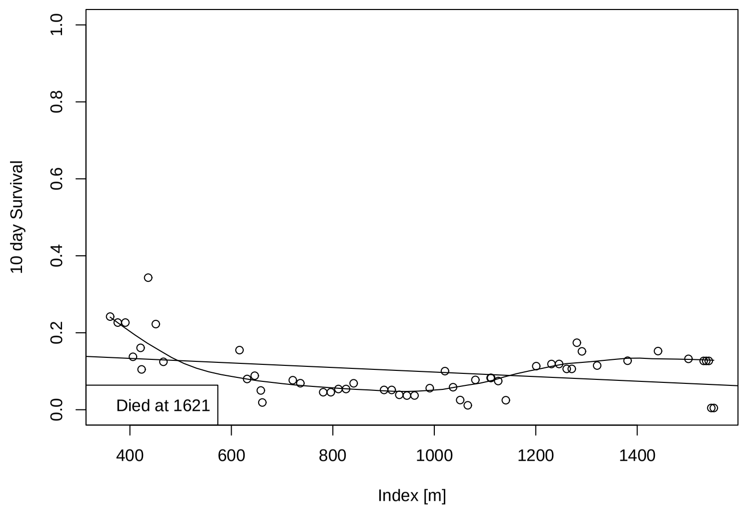

Censored at 7184
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
2600
2800
3000
3200
3400
3600
3800

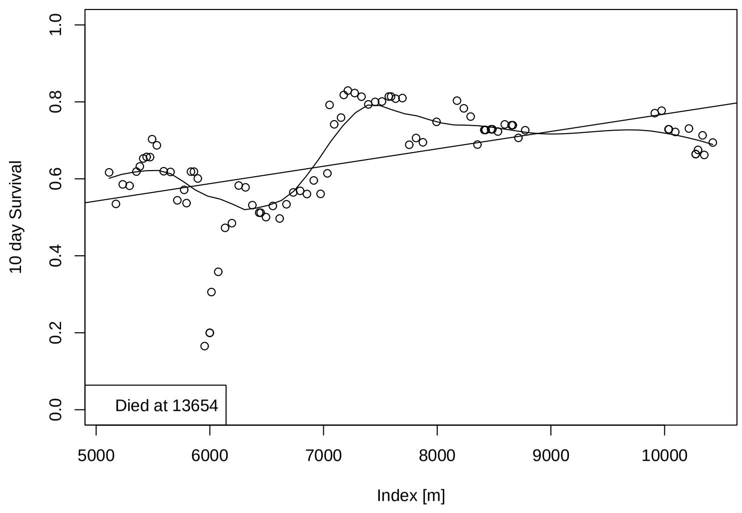

Censored at 9762
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
2000
4000
6000
8000

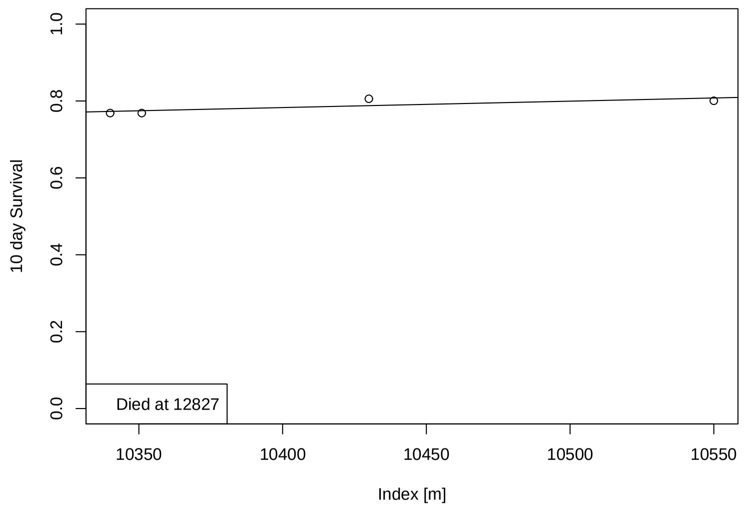

Censored at 27592
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
0
5000
10000
15000
20000

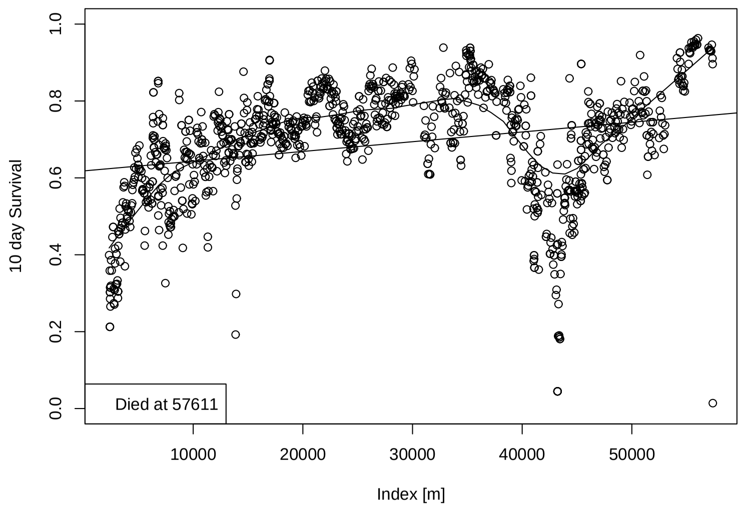
B.2 Uncensored Patients

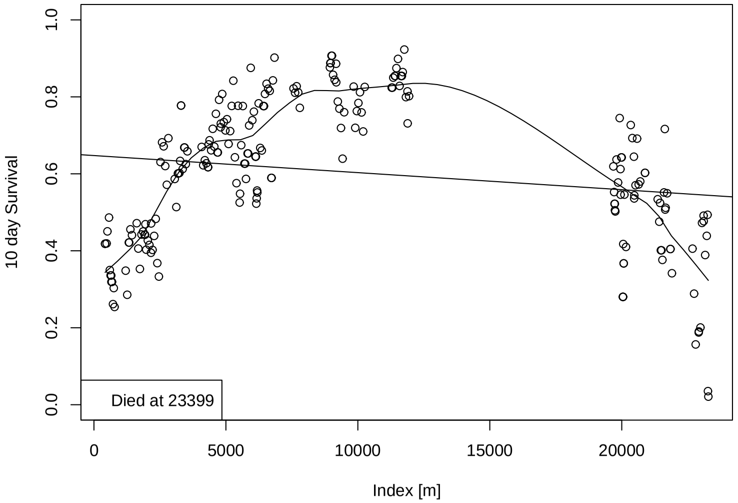
This section provides additional graphs showing the 10-day survival estimate as a function of time for 20 randomly selected uncensored patients. Each of these patients was selected from the test set. The plots include two fitted lines. One of these lines is fit using least squares regression and the other is a smooth line fit using local regression (the *loess.smooth* routine [45]). For the smoothing, a span value of 0.25 was used along with the default polynomial degree of 1.

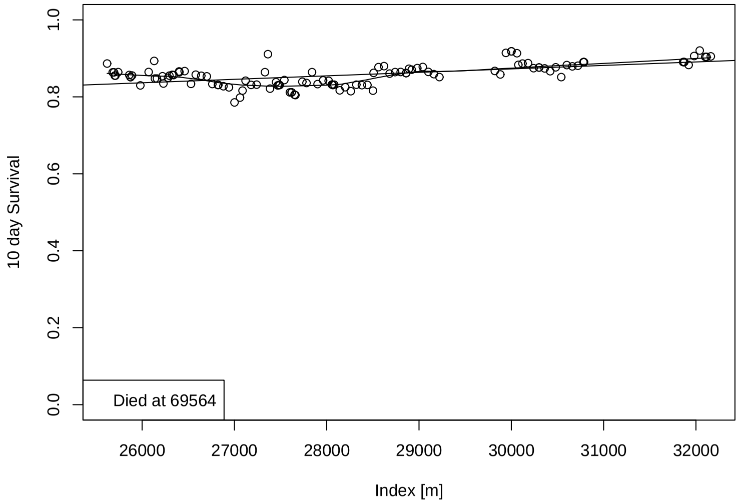

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
400
600
800
1000
1200
1400
Died at 1621

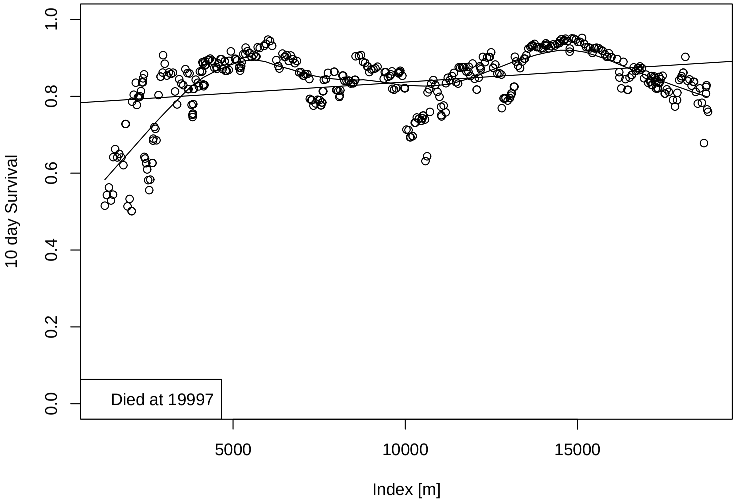

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
5000
6000
7000
8000
9000
10000
Died at 13654

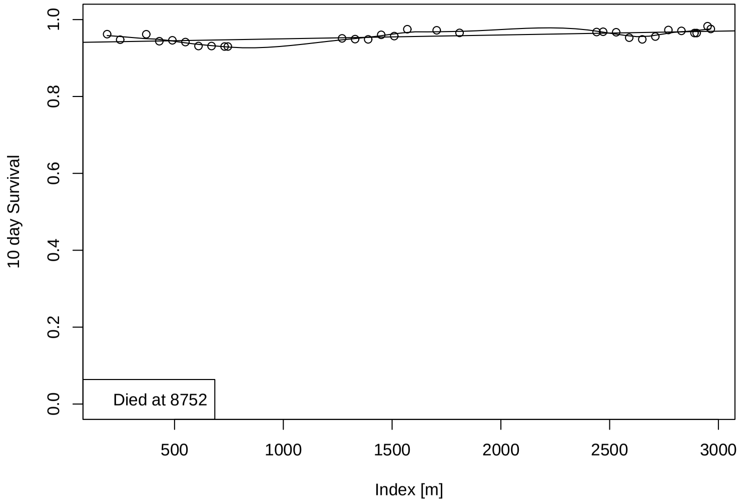

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
10350
10400
10450
10500
10550
Died at 12827

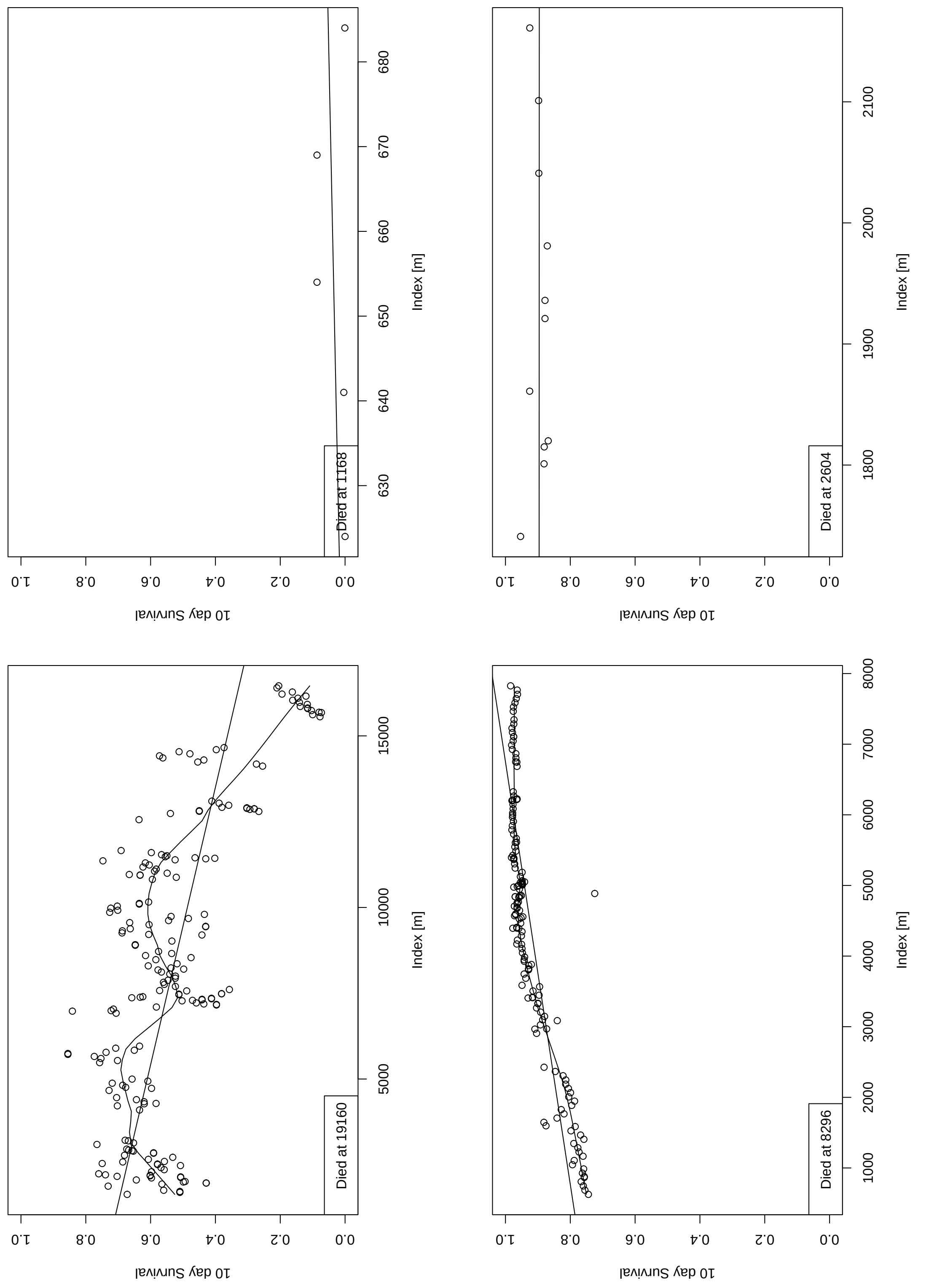

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
10000
20000
30000
40000
50000
Died at 57611

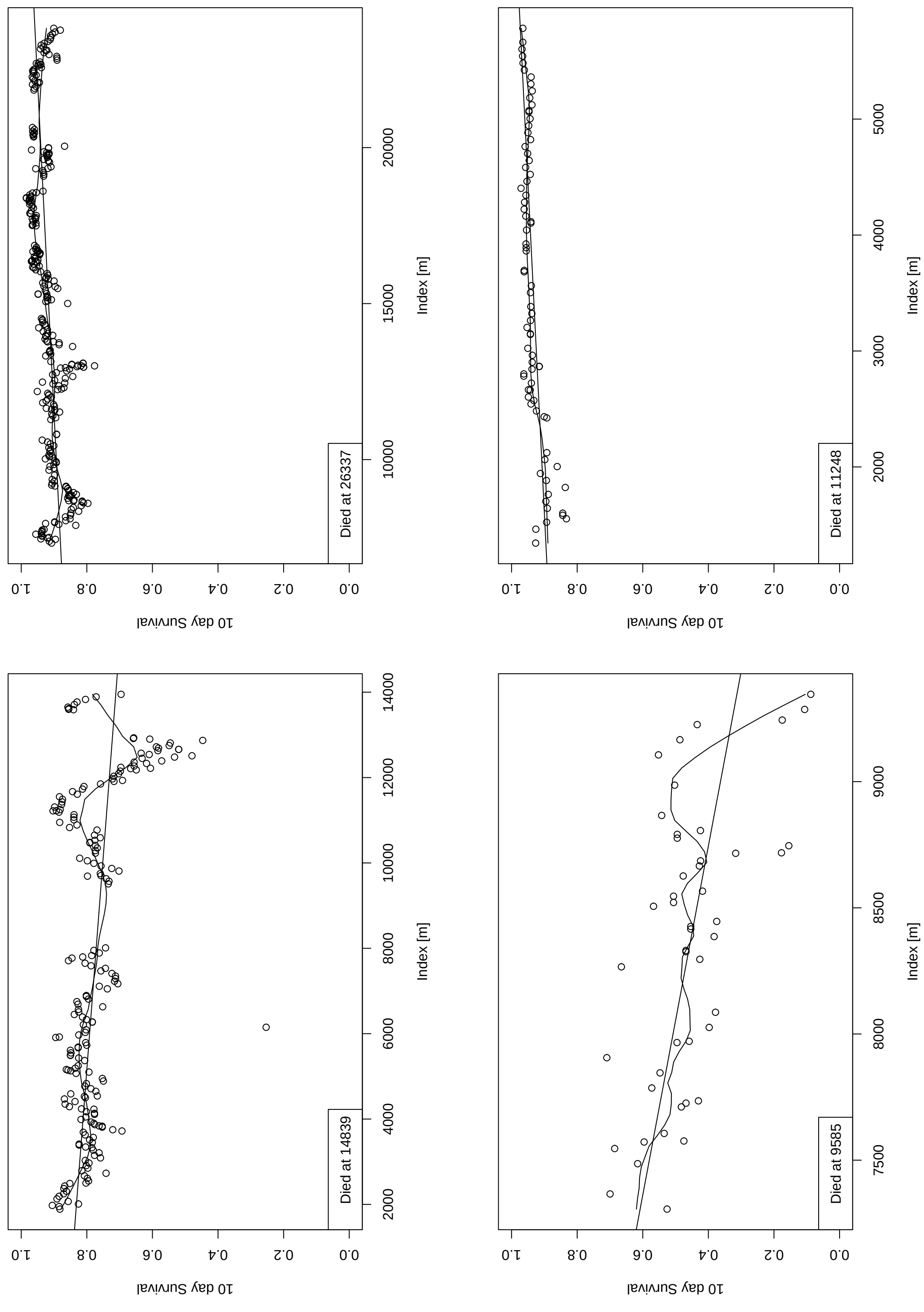

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000
Index [m]
Died at 23399

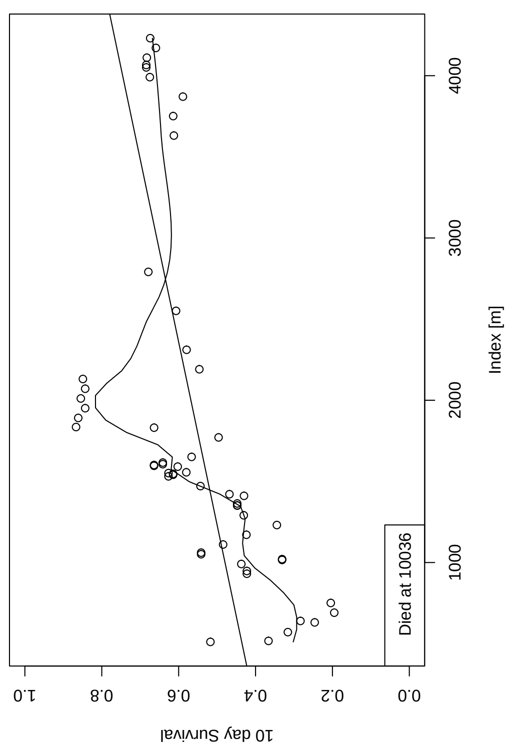

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
26000
27000
28000
29000
30000
31000
32000
Index [m]
Died at 69564

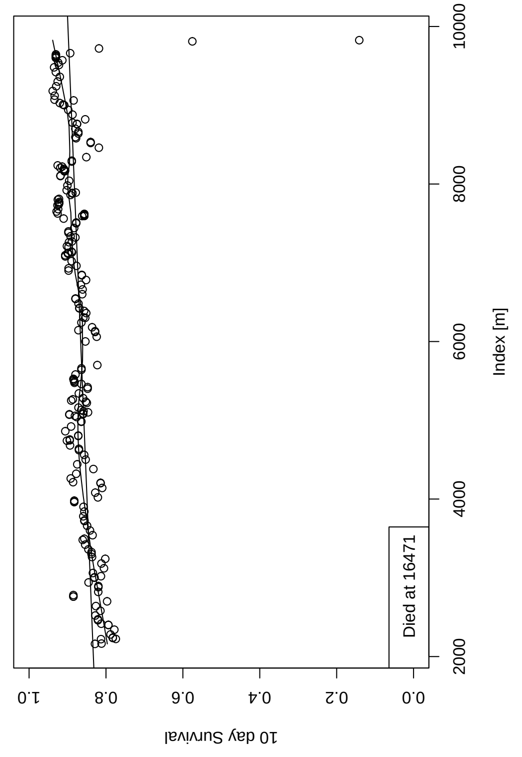

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
5000
10000
15000
Index [m]
Died at 19997

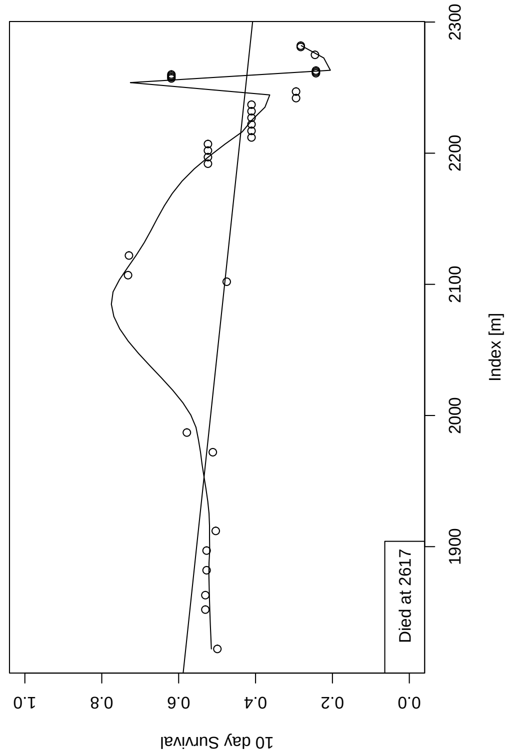

10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
2000
2500
3000
Index [m]
Died at 8752

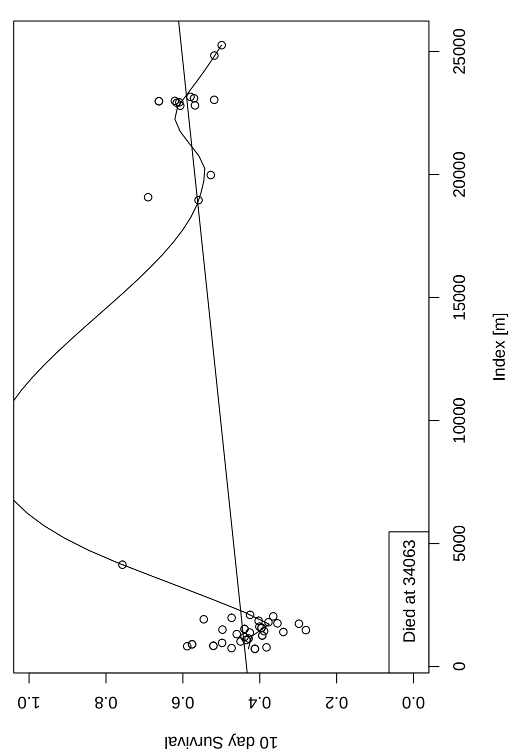

Died at 19160
Died at 1168
Died at 8296
Died at 2604
10 day Survival
Index [m]
0.0
0.2
0.4
0.6
0.8
1.0
5000
10000
15000
630
640
650
660
670
680
1000
2000
3000
4000
5000
6000
7000
8000
1800
1900
2000
2100


Died at 26337
Died at 14839
Died at 11248
Died at 9585
10 day Survival
Index [m]
1.0
0.8
0.6
0.4
0.2
0.0
10000
15000
20000
2000
4000
6000
8000
10000
12000
14000
2000
3000
4000
5000
7500
8000
8500
9000


Died at 10036
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
1000
2000
3000
4000


Died at 16471
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
2000
4000
6000
8000
10000


Died at 2617
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
1900
2000
2100
2200
2300


Died at 34063
10 day Survival
0.0
0.2
0.4
0.6
0.8
1.0
Index [m]
0
5000
10000
15000
20000
25000

Bibliography

[1] J.C. Augusto. Temporal reasoning for decision support in medicine. *Artificial Intelligence in Medicine*, 2005.

[2] R.J. Bosman, H.M. Oudermane van Straaten, and D.F. Zandstra. The use of intensive care information systems alters outcome prediction. *Intensive Care Medicine*, 24(9):953–8, Sep 1998.

[3] Rudiger W. Brause. About adaptive state knowledge extraction for septic shock mortality prediction. In *International Conference on Tools with Artificial Intelligence*, pages 3–8. IEEE, 2002.

[4] N.E. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(89-100), 1974.

[5] J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66(3):429–439, 1979.

[6] D. Calvelo, M.C. Chambrin, D. Pomorski, and P. Ravaux. Icu patient state characterization using machine learning in a time series framework. In *AIMDM'99: Lecture Notes in Computer Science*, volume 1620, pages 356–360. Springer, June 1999.

[7] X. Castella et al. A comparison of severity of illness scoring systems for intensive care unit patients: Results of a multicenter, multinational study. *Critical Care Medicine*, 23(8):1327–35, Aug 1995.

[8] Marie-Christine Chambrin. Alarms in the intensive care unit: how can the number of false alarms be reduced? *Critical Care*, 5(4):184–188qx, May 2001.

[9] M.C. Chambrin, P. Ravaux, D. Calvelo-Aros, A. Jaborska, C. Chopin, and B. Boniface. Multicentric study of monitoring alarms in the adult intensive care unit: a descriptive analysis. *Intensive Care Med*, 25:1360–1366, December 1999.

[10] C-C Chang and C-J Lin. Libsvm: a library for support vector machines. Technical report, Computer Science and Information Engineering, National Taiwan University, 2001-2006.

[11] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[12] Y.C. Chen, C.Y. Chen, H.H. Hsu, C.W. Yang, and J.T. Fang. Apache iii scoring system in critically ill patients with acute renal failure requiring dialysis. *Dialysis and Transplantation*, 31(4):222–233, April 2002.

[13] Y.Q. Chen and S.C. Cheng. Linear life expectancy regression with censored data. *Biometrika*, 2006.

[14] W. Cleveland and C. Loader. Smoothing by local regression: Principles and methods.

[15] A. Cohen. Hidden markov models in biomedical signal processing. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 20. IEEE, 1998.

[16] D.R. Cox. The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society*, 21(2):411–21, 1959.

[17] D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34(2):187–220, 1972.

[18] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, , and Andreas Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2006. R package version 1.5-13.

[19] E Fiaccadori et al. Predicting patient outcome from acute renal failure comparing three general severity of illness scoring systems. *Kidney International*, 58:283–92, 2000.

[20] J.P. Fine and R.J. Gray. A proportional hazards model for the subdistribution of a competing risk. *JASA*, 94:496–509, 1999.

[21] 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. *The Journal of the American Medical Association*, 270(24):2957–2963, December 1993.

[22] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[23] David R. Goldhill and Sumner Anne. Outcome of intensive care patients in a group of british intensive care units. *Critical Care Medicine*, 26(8):1337–1345, August 1998.

[24] E.P. Goss and H. Ramchandani. Survival prediction in the intensive care unit: a comparison of neural networks and binary logit regression. *Socio - Economy and Planning Science*, 1998.

[25] Bob Gray. *cmprsk: Subdistribution Analysis of Competing Risks*, 2004. R package version 2.1-5.

[26] G. Heller and J.S. Simonoff. Prediction in censored survival data: a comparison of the proportional hazards and linear regression models. *Biometrics*, 48:101–15, 1990.

[27] Frank E Harrell Jr. *Design: Design Package*, 2005. R package version 2.0-12.

[28] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[29] J.P. Klein and M.L Moeschberger. *Survival Analysis.* Springer-Verlag, 1997.

[30] W.A. Knaus, E.A. Draper, D.P. Wagner, and J.E. Zimmerman. Apache ii: a severity of disease classification system. *Critical Care Medicine*, 13(10):818–29, Oct 1985.

[31] W.A. Knaus, D.P. Wagner, E.A. Draper, et al. The apache iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–36, Dec 1991.

[32] 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. *Critical Care Medicine*, 9(8):591–7, Aug 1981.

[33] J.R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers. A simplified acute physiology score for icu patients. *Critical Care Medicine*, 12(11):975–977, 1984.

[34] J.R. Le Gall, A. Neumann, F. Hemery, J.P. Bleriot, J.P. Fulgencio, B. Garrigues, C. Gouzes, E. Lepage, P. Moine, and D. Villers. Mortality prediction using saps ii: an update for french intensive care units. *Critical Care Medicine*, 9:R645–R652, Oct 2005.

[35] S. Lemeshow, D. Teres, J.S. Avrunin, and R.W. Gage. Refining intensive care unit outcome prediction by using changing probabilities of mortality. *Critical Care Medicine*, 16(5):470–7, May 1988.

[36] 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. *JAMA*, 270(20):2478–86, Nov 1993.

[37] Y. Li, R.C. Tiwari, and S. Guha. *Mixture Cure Survival Models with Dependent Censoring*. PhD thesis, Harvard University, 2005.

[38] C.J. Lin and R.C. Weng. Simple probabilistic predictions for support vector regression. Technical report, National Taiwan University, 2004.

[39] William Long. Temporal reasoning for diagnosis in a causal probabilistic knowledge base. *Artificial Intelligence in Medicine*, 8:193–215, 1996.

[40] C. Meredith and J. Edworthy. Are there too many alarms in the intensive care unit? an overview of the problems. *Journal of Advanced Nursing*, 1995.

[41] R.G. Miller and J. Halpern. Regression with censored data. *Biometrika*, 69(3):521–31, 1982.

[42] R. Moreno and P. Morais. Outcome prediction in intensive care: results of a prospective, multicentre, portuguese study, intensive care medicine. *Intensive Care Medicine*, 23(2):177–186, Feb 1997.

[43] D. Novák, D. Cuesta-Frau, T. Al ani, M. Aboy, P. Mico, and L. Lhotská. Speech recognition methods applied to biomedical signals processing. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 26. IEEE, 2004.

[44] S original by Terry Therneau and ported by Thomas Lumley. *survival: Survival analysis, including penalised likelihood.* R package version 2.26.

[45] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[46] M. Resche-Rigon, Azoulay, and S. Chevret. Evaluating mortality in intensive care units: Contribution of competing risks analyses. *Critical Care*, 10(1), December 2005.

[47] G. Rocker, D. Cook, P. Sjokvist, B. Weaver, S. Finfer, E. McDonald, J. Marshall, A. Kirby, M. Levy, P. Dodek, D. Heyland, and G. Guyatt. Clinician predictions of intensive care unit mortality. *Critical Care Medicine*, 2004.

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

[49] M. Saeed, C. Lieu, and R.G. Mark. Mimic ii: A massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers In Cardiology*, volume 29, pages 641–644. IEEE, 2002.

[50] M. Saeed and R.G. Mark. Multiparameter trend monitoring and intelligent displays using wavelet analysis. *Computers in Cardiology*, 27:797–800, 2000.

[51] David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–41, 1982.

[52] H.P. Schuster, F.P. Schuster, P. Ritschel, S. Wilts, and K.F. Bodmann. The ability of the simplified acute physiology score (saps ii) to predict outcome in coronary care patients. *Intensive Care Medicine*, 1997.

[53] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. *ROCR: Visualizing the performance of scoring classifiers.*, 2005. R package version 1.0-1.

[54] T. Sinuff et al. Mortality predictions in the intensive care unit: Comparing physicians with scoring systems. *Critical Care Medicine*, 34(3):878–85, 2006.

[55] I. Smith, P. Kumar, S. Molloy, A. Rhodes, P.J. Newman, R.M. Grounds, and E.D. Bennett. Base excess and lactate as prognostic indicators for patients admitted to intensive care. *Intensive Care Med*, 27:74–83, 2001.

[56] Peter J. Smith. *Analysis of Failure and Survival Data*. Chapman & Hall/CRC, 2002.

[57] J. Stare, H. Heinz, and F. Harrell. On the use of buckley and james least squares regression for survival data. *New Approaches in Applied Statistics*, 2000.

[58] Peter Szolovits, editor. *Artificial Intelligence in Medicine*, chapter 5, pages 119–190. Westview Press, Boulder, Colorado, 1982.

[59] M. Tableman and J.S. Kim. *Survival Analysis Using S*. Chapman and Hall/CRC, 2004.

[60] T.M. Therneau and P.M. Grambsch. Martingale-base residuals for survival models. *Biometrika*, 77(1):147–60, 1990.

[61] Christine L. Tsien. *TrendFinder: Automated Detection of Alarmable Trends*. PhD thesis, MIT, 2000.

[62] C.L. Tsien and J.C. Fackler. Poor prognosis for existing monitors in the intensive care unit. *Critical Care Medicine*, 25(4):614–619, April 1997.

[63] Thomas Lumley using Fortran code by Alan Miller. *leaps: regression subset selection*. R package version 2.7.

[64] L. Yan, D. Verbel, and O. Saidi. Predicting prostate cancer recurrence via maximizing the concordance index. In *Proceedings of the tenth ACM SIGKDD International conference of Knowledge discovery and data mining*, pages 479–485. KDD, ACM Press, August 2004.

[65] Ying Zhang. Real-time analysis of physiological data and development of alarm algorithms for patient monitoring in the intensive care unit. Master's thesis, MIT, 2003.

[66] W. Zong, G.B. Moody, and R.G. Mark. Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure. *Med Biol Eng Comput*, 42(5):698–706, Sep 2004.