Dynamic Time Varying Models for Predicting Patient Deterioration

by

Reuben Knowles McCreanor

Program in Statistical and Economic Modeling
Duke University

Date: ______________________

Approved:

____
Rebecca Steorts, Supervisor

____
Sayan Mukhejee

____
Emma Rasiel

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Program in Statistical and Economic Modeling
in the Graduate School of Duke University
2017

Abstract

Dynamic Time Varying Models for Predicting Patient Deterioration

by

Reuben Knowles McCreanor

Department of Statistical and Economic Modeling
Duke University

Date: ____________________________
Approved:

____
Rebecca Steorts, Supervisor

____
Sayan Mukhejee

____
Emma Rasiel

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in the Department of Statistical and Economic Modeling
in the Graduate School of Duke University
2017

Abstract

Streaming data are becoming more common in a variety of fields. One common data stream in clinical medicine is electronic health records (EHRs) which have been used to develop risk prediction models. Our motivating application considers the risk of patient deterioration, which is defined as in-hospital mortality or transfer to the Intensive Care Unit (ICU). Duke University Hospital recently implemented an alert risk score for acute care wards: the National Early Warning Score (NEWS). However, NEWS was designed to be hand-calculable from patient vital data rather than to optimize prediction. Our approach considers three further methods to use on real-time EHR data to predict patient deterioration. We propose a Cox model, a joint modeling approach, and a Gaussian process. By evaluating the implementation of these models on clinical EHR data from more than 51,000 patients, we are able to provide a comparison of the methods on real EHR data for patient deterioration. We evaluate the results on both performance and scalability and consider the feasibility of implementing each approach in a clinical environment. While the more complicated models may potentially offer a small gain in predictive performance, they do not scale to a full patient data set. Thus, within a clinical setting, the Cox model is clearly the best approach.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

DUH	Duke University Hospital.
EHR	Electronic Health Records.
ER	Emergency Room.
EWSs	Early Warning Scores.
GP	Gaussian Process.
JM	Joint Model.
LOCF	Last One Carried Forward.
MCMC	Markon Chain Monte Carlo.
NEWS	National Early Warning Score.
RBF	Radial Basis Function.
ROC Curve	Receiver Operating Characteristic Curve.
RRT	Rapid Response Team.
SVM	Support Vector Machines.

Symbols

x	A scalar (lowercase).
$\mathbf{x}$	A vector (lowercase, bold).
$\mathbf{X}$	A matrix (uppercase, bold).
$\mathbf{x}'$	Transpose of $\mathbf{x}$.

Acknowledgements

None of the work here would have been possible without the ongoing support and mentorship from Rebecca Steorts and Benjamin Goldstein. Beka first took me on as a research assistant in the spring of my first year and has been instrumental in teaching me how to approach a new and complicated problem and guiding me towards previous work and new areas to explore. Ben has also been invaluable as a resource whenever I got stuck and a profound influence in persuading me that the best method is often the simplest one. I would also like to acknowledge Angie Shen who I have worked with since we began this research back in 2016 and Matthew Phelan who has helped bridge the gap between bio-statistical and statistics practices. The clinical team at Duke Hospital, Cara O'Brien, Armando Bedoya, and Meredith Clement, has been incredible in adding insights from an implementation perspective and helping me understand and interpret results from patient vital measurements. Finally, I would like to acknowledge Priyanka Mehta for her ongoing support and suggestions without which I would not have had the motivation to finish this thesis.

1

Introduction

1.1 Motivations

Streaming data are becoming more common in a variety of fields. The most common data stream in clinical medicine is Electronic Health Records (EHRs). EHRs are characterized by having information on many individuals over multiple time points, with time varying feature information. These EHRs can be used to predict patient deterioration. The idea is to create a model that accurately predicts when a patient is likely to experience an event, which is defined as any of the following; patient mortality, a Rapid Response Team (RRT) triggered by an event such as cardiac arrest, or transfer to the Intensive Care Unit (ICU).

In a clinical setting, this type of prediction is of vital importance and inaccurate predictions bring a myriad of social and financial costs. Predicting events unnecessarily causes superfluous interventions which are not only financially costly for the patient and hospital, but also detract from the limited resources the hospital has to treat patients at any one time. Further, studies have shown higher rates of infection in the ICU and therefore transferring patients unnecessarily only serves to increase patient risk (Donowitz et al., 1982). Conversely, failing to predict an event means that a patient may not get the required medical interventions needed to save their life. Thus, the accurate prediction of patient deterioration by creating systems for Early Warning Scores (EWSs) is of vital importance.

These EWSs serve to notify clinicians of changes in patients' clinical status and the need for potential intervention. A commonly used EWS is the National Early Warning Score (NEWS), which is used for detecting patient deterioration. Recently, Duke University Hospital (DUH), under the management of one of our team members, Dr. O'Brien, integrated automated calculation and reporting of the NEWS into the Duke EHR system. While the NEWS has been shown to be efficacious (c-stat = 0.87), it was designed to be easily hand-calculated, and thus utilizes only a few feature variables. As such, it is not optimized to fully leverage all the data available through EHR systems. Further, as a general score, it is not calibrated for the Duke patient population. The goal of this study is to develop more robust risk models for patient deterioration. By providing a comparison of model performance on real EHR data, this work aims to make it easier to prioritize the most effective models when creating systems for predicting patient deterioration.

1.2 National Early Warning Score (NEWS)

The NEWS system currently in use at DUH was developed by the Royal College of Physicians in Scotland in order to design a standardized system for measuring patient risk (RCP, 2012). As can be seen in Figure 2.2, NEWS is measured across seven variables: respiration rate, oxygen saturations, supplemental oxygen, temperature, systolic blood pressure, heart rate, and level of consciousness. In order to calculate a patient's score, we go through each variable and based on its current value, assign it a score of between zero and three. For example, if a patient had a temperature of more than 38 degrees Celsius, this would give them a score of one. After adding together the scores from their six other vitals variables, we would be able to calculate their total NEWS score. A score of greater than or equal to seven across all categories, or a score of three in any single category, is considered high-risk and grounds for transfer to the ICU.

This scoring is arbitrary in the sense that rather than the scores of one, two, and three, being designed to maximize prediction, they were simply chosen as integer representations of the risk levels: low, medium, and high. In the case of a deteriorating patient, NEWS can be a useful score. We can imagine an admitted patient whose vitals reading get progressively worse. First, their systolic blood pressure begins to fall below 109, while their heart rate decreases to less than 51. This could be the beginning of Bradycardia which in extreme cases can cause cardiac arrest. Here, NEWS can be useful in triggering an alarm to alert the hospital staff. However, NEWS is also a deterministic score. Any attending nurse or physician is likely to already be paying close attention to a patient's heart rate and blood pressure and thus the alert may not be telling them anything new. Further, NEWS can be overly sensitive as an alarm system. A patient on supplementary oxygen would trigger an alarm as soon as any one of their other five vitals deviated even slightly from the normal range. This is not necessarily indicative of any further problems and can lead to NEWS alarms being so sensitive that they are ignored by the majority of hospital staff.

NEWS was designed as a general score rather than specifically tailored to the DUH population. Furthermore, it was designed to be hand calculable such that it could be calculated with a clipboard in the field or in an ambulance en route to the hospital. While these are valuable features, they also mean that NEWS is not optimized for patient prediction.

If we examine the NEWS score for a single patient as in Figure 1.2, we see that even within a single patient's stay, there are wide fluctuations in their score. In this patient's stay, they are elevated to a medium-risk level five times, and a high-risk

National Early Warning Score (NEWS)*

PHYSIOLOGICAL PARAMETERS	3	2	1	0	1	2	3
Respiration Rate	≤8		9 - 11	12 - 20		21 - 24	≥25
Oxygen Saturations	≤91	92 - 93	94 - 95	≥96			
Any Supplemental Oxygen		Yes		No			
Temperature	≤35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥39.1	
Systolic BP	≤90	91 - 100	101 - 110	111 - 219			≥220
Heart Rate	≤40		41 - 50	51 - 90	91 - 110	111 - 130	≥131
Level of Consciousness				A			V, P, or U

*The NEWS initiative flowed from the Royal College of Physicians' NEWS Development and Implementation Group (NEWSDIG) report, and was jointly developed and funded in collaboration with the Royal College of Physicians, Royal College of Nursing, National Outreach Forum and NHS Training for Innovation

FIGURE 1.1: Scoring matrix for NEWS showing potential scores for the range of each variable. Scores for all seven variables are summed together and a score of greater than or equal to seven is considered high-risk.

level one time, all within a single one week period.

In DUH, NEWS was implemented starting March 2015 in a rolling fashion across all wards. In principal, a high-risk NEWS score should trigger an RRT (Rapid Response Team) for that patient. In reality, however, according to Dr. O'Brien, in the clinical setting at DUH, these high-risk alerts are ignored 98 percent of the time. Our goal is to develop customized risk prediction models that predict patient deterioration better than NEWS by leveraging the full set of time series data available for patients through the DUH systems and eventually create an alert system that can be trusted in a clinical setting.

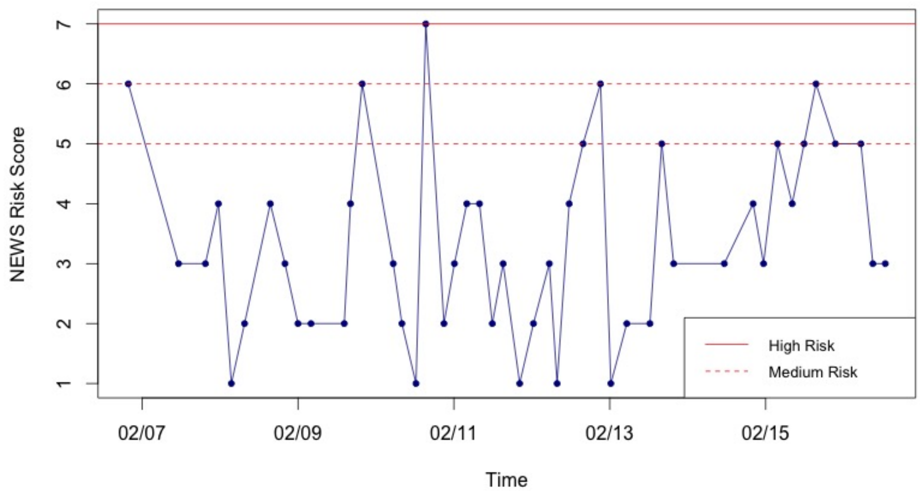

FIGURE 1.2: NEWS score for a single patient stay during their week long stay at Duke University Hospital.

1.3 Prior Work

Part of our research team recently performed a systematic review of 107 papers using EHR data to develop risk models. While these models have shown moderate to strong predictive performance, one shortcoming is incorporating time-varying features. Only 36 studies incorporated multiple measurements into the risk assessment model and only 9 of these studies used time-varying features, with most using summary metrics.

This paper focuses on three key models for predicting patient deterioration: A Cox model, a joint modeling approach, and a Gaussian process. Each of these areas builds upon previous work that is discussed in detail in each model section. This work aims to further the field is by providing a comparison of these methods on real EHR data from a large subset of patients from DUH.

While this work focuses on predicting patient deterioration, it is closely related to the prediction of sepsis in hospital patients which is a subset of potential deterioration. Prior work on patient prediction largely falls into three categories: classification, optimization, and survival analysis.

In the classification based approach, prior work by Shiao and Cherkassky (2016) uses Support Vector Machines (SVM) to apply a binary classification framework to survival data. In this manner, after training the model, each patient will receive a classification of either event or no event, based on the data from the inputted covariates. The key advantage of this SVM is its predictive performance. Given that the SVM algorithm maximizes the minimum margin between class labels, it often gives very high accuracy in prediction. Further, while SVM only provides a binary classification, a modified version, SVM+, can be employed to give probabilities for uncertain membership classes. However, one of the key issues with such an approach is the interpretation of the algorithm. As a classification method, it is often unclear which variables are driving the prediction and how changes in these variables change patient risk. In a hospital setting, is not necessarily useful to know that a patient is deteriorating without knowing which covariates are driving this increasing risk.

Further classification based approaches have also been used to predict patient mortality from sepsis using machine learning random forest models (Taylor et al., 2015). While these approaches have yielded good results in terms of predictive ability, they do not take into account the time-varying structure of vitals data. Further, although many patients can have different numbers of observations depending on the length of their stay, a random forest classification model can only be implemented when each patient data stream has the same number of observations.

Another interesting approach is predicting patient deterioration or sepsis is the idea of a risk score. Given a set of covariates, the idea is to design a risk scoring system through some combination of these covariates that can provide an alert before the patient beings to deteriorate. Given that any combination of these covariates is possible, this essential becomes a complex optimization problem where we want to pick the combination and scoring system that maximizes predictive accuracy. Much of the work in this optimization based approach has been undertaken by Dyagilev and Saria (2016) at John Hopkins University.

Finally, the last type of approach involves improving existing survival based approaches such as the Cox model by using more flexible methods such as Gaussian process or Exponential family distributions. Prior work by Barrett and Coolen (2012) uses Gaussian process regression to directly infer the relationship between event times and covariates. While still assuming a Cox-type hazard rate, this more flexible process means models can infer non-monotonic relationships between the covariates and event times. In its current state, this type of model has only been run on simulated data, however, it will eventually provide an interesting comparison with a simpler Cox model for survival analysis.

In using Exponential family distributions, prior work by Ranganath et al. (2016), uses a Bayesian framework with deep exponential families to model survival data. This approach allows for a non-linear latent structure that can capture complex dependencies between the covariates and the failure time. Early indications show that this method may offer improvements over traditional Cox models and there will be interesting future comparisons on both predictive performance and the feasibility of implementation for deep survival analysis on large patient data sets.

1.4 The Patient Data

The data for this work comes from EHR from DUH. In total, we have data on 51,673 patients admitted to inpatient care wards between November 1, 2013, and December 31, 2015. Each patient has data for up to a seven-day period giving more than 23 million vitals data points across respiration, pulse, consciousness, oxygen saturation, temperature, blood pressure, and supplementary oxygen.

The data also includes important patient demographics such as sex, age, time of admission, lab data giving information on any physician ordered labs during their stay such as blood counts, and data on rare events such as cardiac arrest. Aside from the vitals data, we also have access to lab results from 13,718,860 labs, and more than 98 percent of patients have at least one lab ordered during their stay. Further details on the break up of the patient data can be found in chapter 2.

2

The Patient Data

2.1 Data Summary

Our data is in the form EHRs from DUH. In total, we have data for 51,673 patients admitted to inpatient care wards between November 1, 2013 and December 31, 2015. For each patient, we have repeated measurements of key vitals for up to a seven day period for their respiration rate, oxygen saturations, temperature, systolic blood pressure, heart rate, level of consciousness, and the presence of supplemental oxygen. We also have demographics data and time-invariant information for each patient including factors such as sex, age, and indicators such as whether the patient suffers from diabetes. Finally, we have laboratory test results from any labs that were ordered for a patient during their stay. The most common of these is a White Blood Cell (WBC) count that is ordered for 93 percent of patients. For the purposes of this research, we define an event as patient morality, or transfer to the Intensive Care Unit (ICU).

Table 2.1: Summary of key patient data.

Data Type	Count
Patients	51,673
Demographic variables	28
Vitals measurements	23,263,683
Lab results	13,718,860

2.2 Data Cleaning

From our full data set of 51,673 patients, we subset the data to a one year period from July 1, 2014 to June 30, 2015, when the majority of patients were admitted. This leaves us with a total of 34,477 patients. The key issues in cleaning our data include missing information, out of range vitals, ascribing patients to an event, the presence of supplementary oxygen, and patients' level of consciousness.

From this subset, a total of eight patients have missing demographics information and are removed from the data set. According to the clinical team, out of range vitals are likely a result of measurement error. In consultation with the clinical team we created plausible ranges for each vital, and removed the less than 0.1 percent of measurements that did not fall into these ranges from the data. In our dataset, subjects have two variables, one binary variable specifying whether they are receiving supplementary oxygen, and one categorical variable that describes the type of supplementary oxygen device such as a face mask. These two variables are not always in accordance and it is possible for a patient have an indicator for no supplementary oxygen while listing some type of supplementary oxygen device. For simplicity, we assume that any type of supplementary oxygen device corresponds to receiving supplementary oxygen. This assumption turns our two variables into a single indicator variable for supplementary oxygen.

Finally, within the data, consciousness has five levels: alert, lethargic, responsive

Table 2.2: Proportions of demographics with and without events.

Category	Patients without Event (percent)	Patients with Event (percent)
Race (Black)	28.5	30.6
Race (Other)	5.6	5.5
Race (White)	65.8	63.9
Sex (Female)	48.6	43.5
Sex (Male)	51.4	56.5

to verbal stimuli, responsive to physical stimuli, and unresponsive to stimuli. Given that many of these levels are largely in indistinguishable and do not offer insights into predicting patient deterioration, we again transform these variables to a single indicator variable for patient responsiveness wherein a patient is either alert or not alert.

2.3 Demographics and Pre-existing Conditions

In considering the presence of events in the data, we see that approximately 5 percent of patients in the data experience an event. As can be seen in Table 2.2, among the demographics variables of sex, race and age, there is not a significant difference between patients with and without events.

Further, we also want to examine the number of events across indicators such as diabetes and HIV. As can we seen in Table 2.3, while there are marginally higher number of patients that experience events from the populations with Diabetes, this difference is not large enough to warrant concern for our predictive inference.

2.4 Survival Probability

The Kaplan-Meier curve estimates the survival function at time t in the j^{th} time interval, which is the probability of surviving through the j^{th} and all previous time intervals.

Table 2.3: Proportions of pre-existing conditions with and without events.

Category	Patients without Event (percent)	Patients with Event (percent)
CKD	18.8	28.7
COPD	10.8	16.8
Diabetes	29.2	34.8
HIV	1.2	1.3
Malignancy	30.1	33.3
MI	8.9	16.7
Post-op	7.0	7.3
Stroke	5.2	8.7
Surgery	26.2	21.4
Transplant	7.1	9.4

We plot the KM curve for a random sample of 30,000 patients. The data is collected for up to 7 days over each patients stay. As previously noted, events are not very common. Overall, 10 percent of patients have an event within this period. However, the intensity of events goes down as time passes and most events happen within the first two days.

2.5 Patient Similarity

Finally, we want to ask, given some subset of patients that have events in the same timeframe, do patients at time $t + n$, look like themselves at time t, or look like at patients at time $t+n$. Essentially we are asking a question about patient trajectories and as a patient progresses through their hospital stay, whether they continue to follow a unique trajectory by looking like themselves in the past, or whether they follow a dynamic trajectory by looking more like some other patient at the same future time point. This analysis has important implications for the type of models we fit, and the clustering for the GP progress in chapter 5.

In order to explore this idea, we randomly select 100 patients that were discharged without an event after four days. We want to ask the question: do patients on day

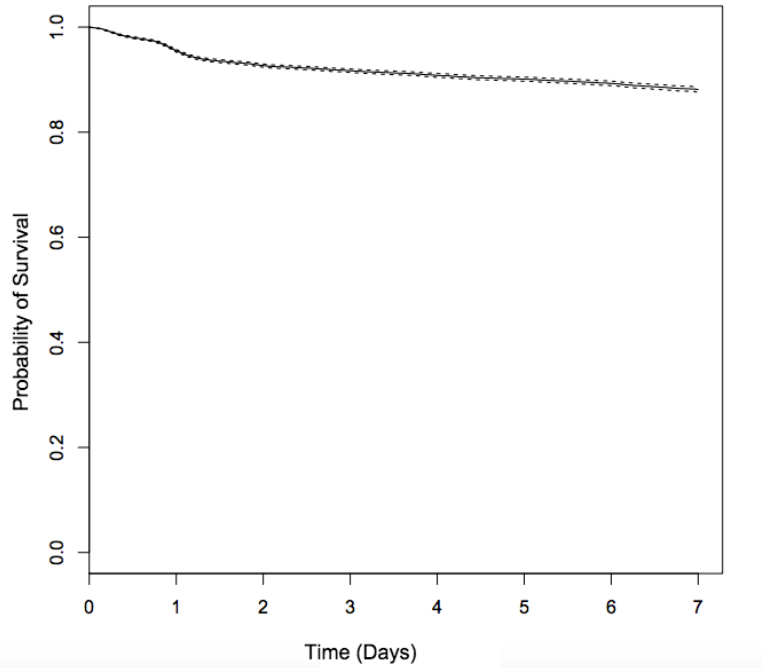

FIGURE 2.1: Kaplan-Meire Plot showing the probability at time t of surviving through all previous time periods (Shen, 2016).

four look like other day four patients, or do they look more like themselves on day one? To answer this we create two sets of observations for these 100 patients, their mean vitals on day one, and their mean vitals on day 4. We then use a k-means clustering approach to group these observations, and see which observations are most similar.

To find k, the number of clusters, we employ a simple empirical approach to examine the sum of squares for differing numbers of clusters where:

$$SS = \sum_{j=1}^{K} \sum_{n \in S_j} ||x_n - \mu_j||^2. \tag{2.1}$$

After clustering our patient data, we first find that patients look more like themselves than other patients. Even with four clusters, 57 percent of patients have their

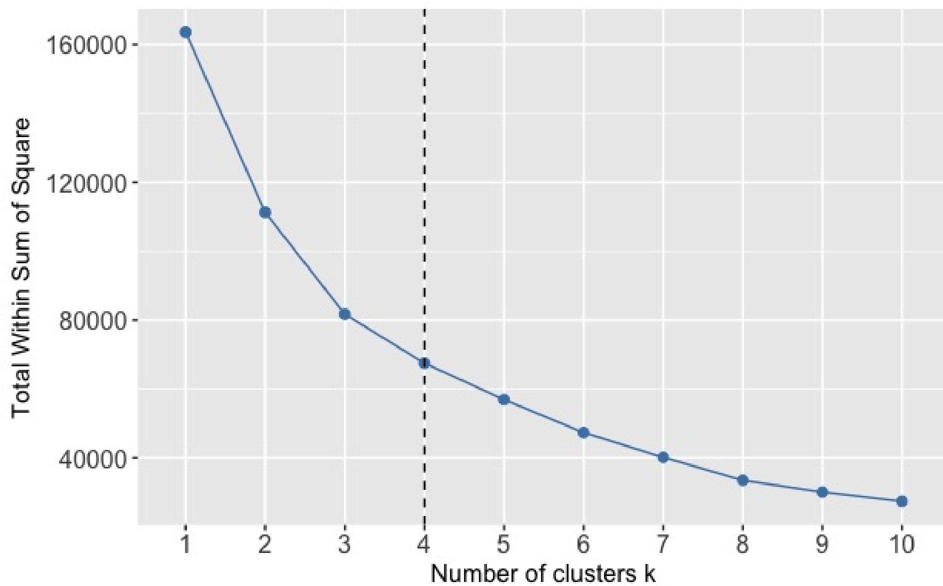

Figure 2.2: Optimal number of clusters for patient vitals.

observations from day one clustered to the same group as their observation from day four. This suggests that patient trajectories are relatively stable over a patient's hospital stay. As there is no significant change in many vitals measurements over this period, patients continue to look like themselves on day four as opposed to other patients with similar trajectories.

Next, we want to ask the same question for patient with events. As there are less patients with events, we instead take the 67 patients that had an event on day four. We again find four as the optimal number of clusters with the majority of methods giving an optimal cluster number of four. In a similar manner, we find that 44 percent of patients end with both sets of vitals measurements in the same cluster. This is interesting because it suggests that while some patients that experience an event may deteriorate in a similar manner to other patients with events, the individual nature of patient deterioration makes this a challenging process to model.

3

The Cox Model

3.1 Background

The Cox proportional hazards model is a semi-parametric regression model from the class of survival models. It is a popular model in survival analysis that assesses the importance of covariates in survival times through a hazard function (Harrell, 2001). The hazard rate is defined as the probability of suffering from an event, in this case, mortality, an RRT, or transfer to the ICU, given that the individual has survived up to a specific point in time. The model is semi-parametric because there are no assumptions about the shape of our baseline hazard function.

In general, the Cox model relates the time before the occurrence of an event to the covariates that are associated with this same timeframe. More specifically, in the case of our data, the Cox model relates the time before transfer to the ICU or another event, to the time-varying covariates developed from our patient variables such as systolic blood pressure and temperature.

3.2 The Model

For each patient, the model estimates a hazard function and assesses how the time varying covariates affect this hazard. For each individual i, we say that:

$$\begin{aligned} \lambda(t|X_i) &= \lambda_0(t)\exp(\beta_1 X_{i1} + ... + \beta_p X_{ip}) \\ \Rightarrow \lambda_i(t|x_i) &= \lambda_0(t)e^{x_i(t)'\beta} \end{aligned} \tag{3.1}$$

Where $\mathbf{x}$ is our vector of time-varying covariates such as temperature and blood pressure, $\boldsymbol{\beta}$ is a vector of coefficients, and $\lambda_0(t)$ is the baseline hazard.

3.3 Prior Work

With the prevalence of EHR records in hospital systems, streaming data and ways to develop risk models for evaluating this data, are becoming increasingly common within the medical field. Hug and Szolovits (2006), published work on predicting risk and trajectory of intensive care patients using survival models. In their work, they proposed a Cox based model to monitor patient outcome projections across the current stat and near-term trajectories as patients progress through the ICU. Using Boston-area hospital data from 3116 patient across 132 characteristics, they used a Cox model to access patient trajectories and find risk at any time during a patient's stay in the ICU.

Where our approach differs is the goal of not simply modeling patients in a single stay, but using ICU transfer as an outcome variable. This allows us to not only assess a patient during their stay but to create risk predictions that will help decide when intervention is necessary for a patient and to create a robust metric to aid ICU transfer decisions.

3.4 Data Setup

In order to run a Cox model, we have to transform our data from its original long format, into a wide format with a counting process set up. In order to do this, we take all of a patient's vital measurements for up to a seven-day period while they are admitted to hospital. In Duke Hospital, these vitals are updated approximately every three hours. Every time a vital update occurs, this corresponds to a new row in our counting process data frame where the start time of this row equals the end time of the previous row. A binary indicator is used for the presence of an event, and vitals measurements are included up to the occurrence of an event, the discharge of a patient, or the end of the seven-day window, whichever comes first.

In a hospital setting, vitals measurements are not always updated uniformly. It is easy to imagine a scenario where a nurse or machine might update a patient's blood pressure while not taking their temperature. In this case, a new row would be created in this counting process format, however, there would not be an updated value for temperature. To counter this problem, we use a Last One Carried Forward (LOCF) approach wherein previous values are carried forward in each update for any vital measurement that is not updated. A simulated version of this counting process format for our data can be seen in Figure 3.1.

3.5 Results

We fit the model to a random sample of 1,000 patients using the full set of vitals variables to obtain a recalibrated version of the NEWS. The coefficients from the model are displayed in Table 3.1.

Id	Start	Stop	Temp	Pulse	Resp	SpO2	SysBP	O2	Consciousness	Event
1	0	3	37.00000	70	16	100	120	0	Alert	0
1	3	6	37.00000	79	14	100	154	0	Alert	0
1	6	9	36.77778	77	16	97	154	0	Alert	0
1	9	12	36.77778	72	13	97	154	0	Alert	0
1	12	15	36.77778	73	14	95	154	0	Alert	0
1	15	18	36.77778	71	16	96	154	0	Alert	0
1	18	21	36.77778	80	16	95	154	0	Alert	0
1	21	24	36.50000	78	14	100	121	1	Alert	0
1	24	27	36.50000	78	14	100	121	1	Alert	0
1	27	30	36.50000	82	19	100	112	1	Alert	1

FIGURE 3.1: Simulated data in counting process format, note the event in the last row.

Table 3.1: Coefficients for patients features from the Cox Model fit

Features	Coefficients
Consciousness	1.19
O2	1.15
Respiration	0.09
Temperature	0.03
Pulse	0.03
Systolic BP	-0.01
SpO2	-0.05

3.6 Scalability and Implementation

In order to access the runtime of our model, we perform an empirical runtime analysis. The Cox model with demographic information is run on randomly selected subsets of 100, 1,000, 5,000, 10,000, and 27,000 patients (the full training set). Each run is repeated 50 times and the results are reported in the Figure 3.2.

We note that the runtime increases approximately linearly with the number of patients. However, even with the full set of more than 27,000 patients, the full run

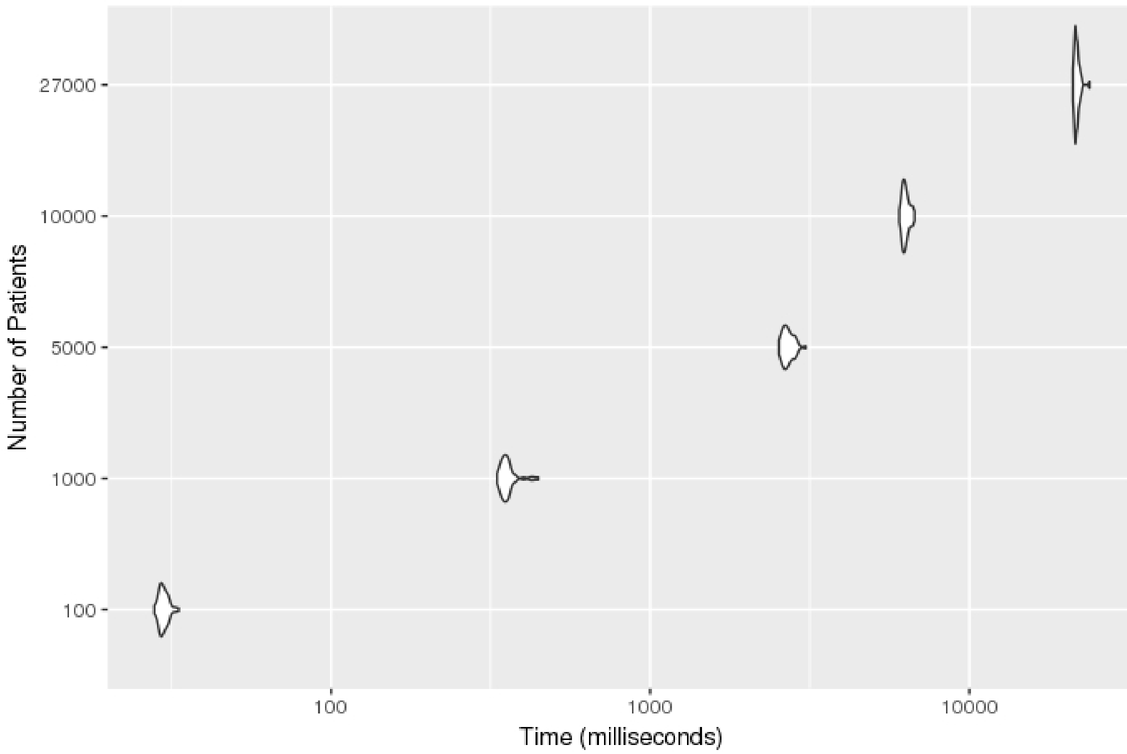

FIGURE 3.2: Run time for Cox model across increasing numbers of patients.

time is only 13 seconds. Thus, the Cox model proves to be easily scalable for large numbers of patients.

4

The Joint Model

4.1 Background

The advantage of the Cox model discussed in chapter 3 is its computational simplicity. However, the simpler Cox model fails to take into account a patient's changing clinical dynamics. In contrast, a joint model (JM) allows us to simultaneously model both survival and longitudinal data. In particular, JMs are useful when longitudinal data points are highly associated with the time to an event. The Cox proportional hazards model for time-to-event data does not consider dependencies between longitudinal and time-to-event data.

In contrast, a JM has two components, a longitudinal sub-model such as a linear mixed effects model and a time-to-event sub-model such as a Cox proportional hazards models (Sudell et al., 2016). These two sub-models are linked using an associate structure, the JM, that quantifies the relationship between these sub-models and the outcome variable. By taking into account this dependency, JMs have the potential to reduce parameter estimate bias and thus provide more efficient estimates of the treatment effects on the time to an event.

4.2 The Model

We use a two-stage joint model on both survival and longitudinal data.

In stage one, we fit a mixture model for each individual i:

$$y_i(t) = m_i(t) + \epsilon_i(t) = x_i'(t)\beta + z_i'(t)b_i + \epsilon_i(t). \tag{4.1}$$

Where x_i and z_i are the fixed and random longitudinal features with coefficients β and b_i, and $\epsilon_i(t) \overset{iid}{\sim} N(0, \Sigma)$.

In stage two, we use a survival model with a hazard function $h_0(t)$. We say that:

$$h_i(t|M_i(t), w_i) = h_0(t)e^{\gamma' w_I + \alpha m_i(t)}. \tag{4.2}$$

Where $m_i(t)$ is the longitudinal history from Equation 4.1 and α is the corresponding coefficients.

4.3 Prior Work

The advent of JMs to patient data is recent development with the earliest documented paper by Wulfsohn and Tsiatis (1997). The main applications for JMs are in clinical trials where the sparsity of measurement and time-to-event aspects make a great fit for the JM approach. A recent paper by Wang et al. (2016), used the multiple repeated measurements and event time data to predict the onset of Parkinson's Disease.

The key differences between these type of longitudinal studies and predicting patient deterioration lie in the time horizon, measurement frequency, and cohort size. In the Parkinson's study, data points span a year or more in time, taken every three months, for approximately 800 patients. This scale gives vast differences in terms of the computational complexity. For 800 patients, there may be 28,000 or more observations. In our case, however, with more than 34,000 patients and vitals

Table 4.1: Coefficients for patients' features from the Joint Model fit

Features	Coefficients (posterior mean)
Pulse	1.33
Consciousness	1.15
O2	1.11
Temperature	0.34
Respiration	0.30
Systolic BP	0.21
SpO2	-1.46

collected every 3 hours and up to a 7 day period, across 7 variables, there are more than 5 million observations.

In terms of fitting a JM, much of the pioneering computational work has been undertaken by Dimitris Rizopoulos, the author of the R package JM and JMBayes (Rizopoulos, 2014). In our work, we make use of the JMBayes implementation to fit a multivariate joint model using the Bayesian Markov Chain Monte Carlo (MCMC) approach.

4.4 Data Setup

The joint modeling approach uses the same counting process set up as in the Cox model. However, within a joint model, we fit both the random effects and survival models separately before running the joint process.

4.5 Results

We run the JM on a random sample of 1000 patients. The resulting coefficients from the JM are included in Table 4.1. Compared to the Cox model, in particular, we notice that once we incorporate the time varying nature of our coefficient variables, Pulse, Temperature, and Respiration become far more important in our model.

4.6 Scalability and Implementation

The key issue with the JM is that even in its ideal theoretical implementation, the JM algorithm is quadratic in runtime (Riedel and McCallumi, 2011). Given the number of patients in our sample, it is not possible to implement the joint model on our full training data set. Further, with the current state of resources for JMs, while it is possible to fit a JM in the multivariate case, there is no current method for moving from here to survival predictions as we can in the univariate case.

Much like in the case of the Cox model, we again perform an empirical runtime analysis. The Cox model with demographic information is run on randomly selected subsets of 100, 5,00, and 1,000 patients. Unlike in the Cox model, the JM is so computationally intensive that it is not possible to run it on the full training set and even running on 1,000 patients takes more than 22 hours. Each run is repeated 50 times and the results are reported in the Figure 3.2.

As we can see, in the case of just 1,000 patients, the JM already takes 17 hours to run. As we increase the number of patients to 10,000, we see that the joint model takes more than a day to fit, even running in parallel on a 24 core server. The inability to scale the JM to the full training set makes it computational infeasible in its current state within a clinical environment.

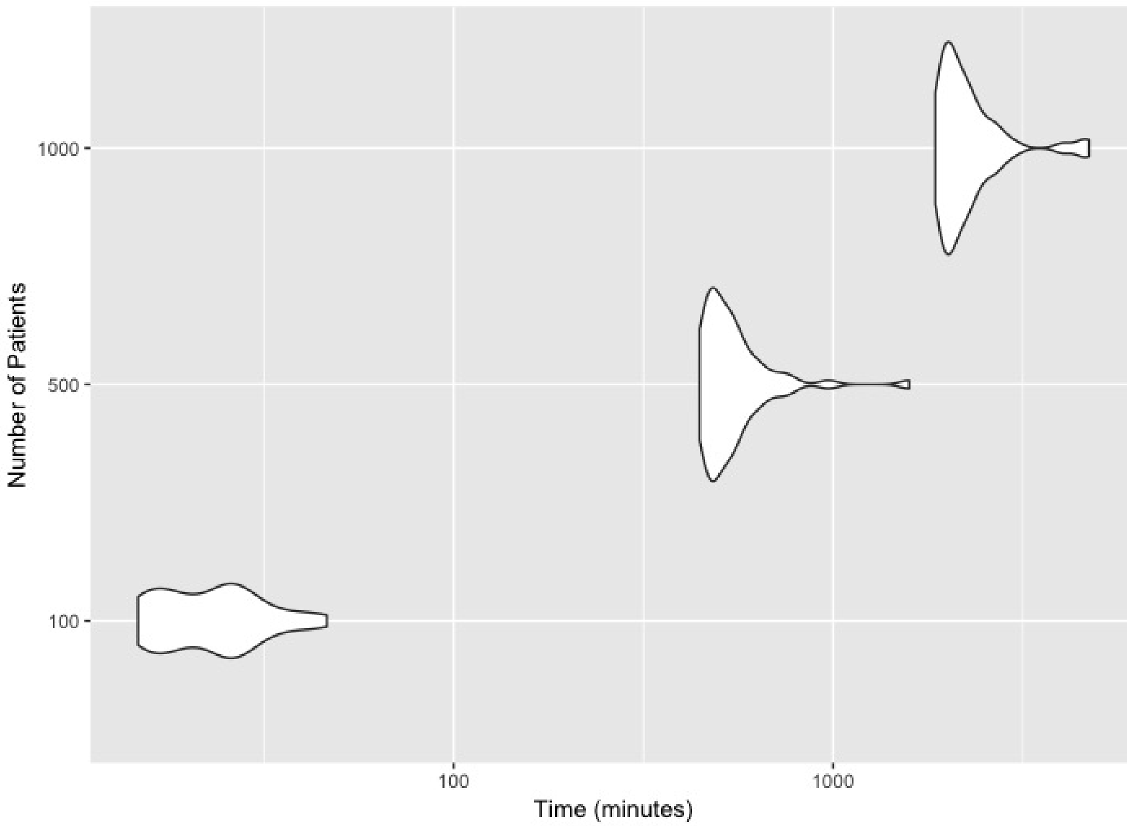

FIGURE 4.1: Run time for Joint model across increasing numbers of patients.

5

Gaussian Processes

5.1 Background

A Gaussian process (GP) is essentially a flexible form of regression in both time and space. As a non-parametric method, a GP finds a distribution over all possible functions $f(x)$ that fit the observed data. We begin with a prior distribution and update this prior as data points are observed, producing a posterior predictive distribution over all functions.

What makes a GP interesting for this type of medical application is the potential to forecast vitals measurements unique to each patient. Rather than comparing vitals measurements in aggregate as we do in the case of the Cox model and JM, a GP model is fit separately for each patient or cluster of patients.

To understand the process of the GP method, we take a single vital such as heart rate for a single patient. Given the heart rate measurements up to time t, we use a GP to a curve to the changes in the patient's heart rate. Taking this training window, we can then create a forecast window up to time $t+n$. When a new heart rate measurement comes in between time t and $t+n$, it is simply a matter of plugging this value into the posterior of our fitted GP. A high log-likelihood value indicates a large deviation from our forecast value. This is interpreted as a warning sign for the deterioration of a patient's vital and an alarm is triggered based on some predefined threshold.

5.2 The Model

A process, $\boldsymbol{Y} = \{Y_t : t \in T\}$, is said to be a Gaussian process if all possible finite dimensional vectors $\boldsymbol{y} = (y_{t_1}, y_{t_2}, ..., y_{t_n})^t$, for every collection of time points $t_1, t_2, \ldots, t_n$, and every positive integer n, have a multivariate normal distribution (Douc et al., 2014).

In general, we define our GP as:

$$\mathbf{y}_{nx1} \sim MVN(\boldsymbol{\mu}_{nx1}, \boldsymbol{\Sigma}_{nxn}). \tag{5.1}$$

Where $\boldsymbol{\mu}$ is the vector of means for each element in $\mathbf{y}$ and $\boldsymbol{\Sigma}$ is the covariance matrix for the pairwise correlation between each element in $\mathbf{y}$.

We can then use Gaussian Process Regression (GPR) to estimate some unknown regression function $y = f(x)$ for our given data $\{x_i, y_i\}_{i=1,\ldots,n}$ without needing to explicitly define $f(x)$.

Instead, we model f(x) such that:

$$f(x) \sim GP(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{5.2}$$

Where $k(\mathbf{x}, \mathbf{x}')$ is our covariance function and a valid kernel.

5.3 Prior Work

Prior work on using Bayesian GPs to identify deteriorating patients at the University of Oxford is currently in progress by Colopy et al. (2016). In their work, they fit a GP regression model as a method to continuously monitor patient time-series vital signs acquired in the step-down unit, a high acuity hospital environment. Using patient heart rate as the key vital they test their methods of data from the United of Pittsburg Medical Center, comprising of 333 patients, 59 of whom had a least one verified clinical emergency event.

What makes our application considerably more difficult is the size of our patient data set, structure and number of events, and multiple and often uneven vitals measurements. While Colopy et al. have data with only 333 patients, almost 18 percent of which have an event, it is a challenge to scale this approach to more than 27,000 patients with only a 5 percent rate of events. Further, even for Pulse rate, our vitals measurements are somewhat uneven in that they are ideally updated every three hours, but sometimes are not updated for longer periods. Finally, the biggest challenge remains the distribution of events in our data. While we have data up to the seven day period of admission, many events happen in the first 24 or 48 hours. Given the sparsity of data points, it is difficult to build a robust GP to forecast future event with so few previous data points.

5.4 Results

First, we fit a GPR for a single patient on their Pulse rate using a standard mean zero and RBF (Radial Basis Function) kernel formation. In Figure 5.1 GP we see the data points in blue, with the predicted fit from our GP and the 95 percent credible interval for our predictions given in green.

Further work will use this fit to produce a forecast. Given this GPR, we want to consider some time point t, and a new pulse rate coming in at time $t + n$. For each new data point, we compare it to the posterior distribution to obtain a log-likelihood value. A low log-likelihood within a forecast window suggests a change in patient status and an alarm can be triggered on this basis.

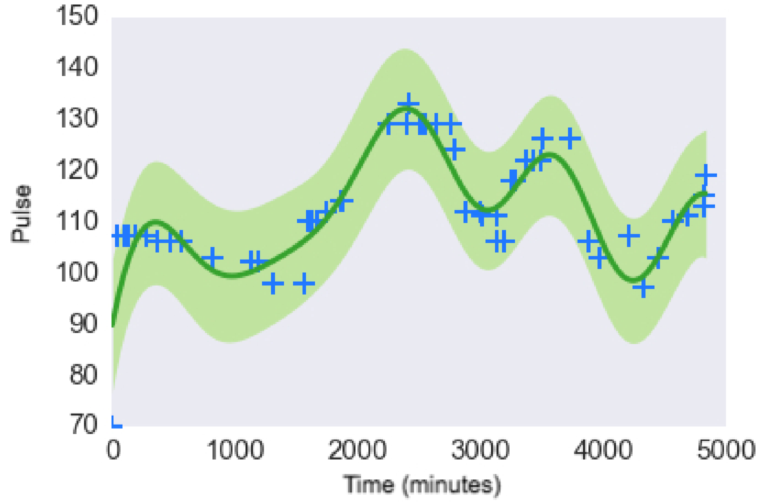

FIGURE 5.1: Gaussian Process for an individual fit to pulse rate over entire hospital stay.

5.5 Scalability and Implementation

The biggest asset of the GP is its customizability. Rather than comparing patients' vitals among other patients, the GP creates a unique forecast for each patient in the data set. However, the obvious downside is that these curves have to be fit on each patient individually.

While a single GP can be run in 10 seconds, the full set of 27,000 patients would take more than 3 days. In theory, this process is embarrassingly parallel. There is no reason that each GP could not be fit separately and simultaneously. In practice, however, this would require costly subsetting of the data for each individual patient, and only scale to the total number of machines available.

However, the key issue is implementing a GP model is not simply the computational complexity, but its performance across the entire patient data. While the model fits well with dense observations, the majority of patients with events have them in the first 24 hours in which many patients have as few as two total observations. Given that we cannot build a time series model on such sparse data, there is no reliable way to build a GP model on individual patients without first clustering similar patients which has proved unreliable as discussed in section 2.5. Further work plans to address both the clustering of this sparse data, and the possibility of building a cross-sectional GP model on the entire patient data set so that predictions could be based not only on individual patient observations, but also from the correlation structure between patients.

6

Conclusions

6.1 Model Comparisons

From the work across the Cox proportional hazards model, Joint Model, and Gaussian Processes, it is clear that the Cox model provides the most readily adaptable formulation for predicting patient deterioration. While the JM offers advantages with its more efficient coefficients, the computational complexity, lack of scalability, and the current absence of predictive probabilities in the multivariate case make it a difficult model to implement in its current state. In the GP model, the sparsity of individual patient measurements does not make it an ideal model for this type of data. Furthermore, the individual nature of the GP fits makes this a difficult model to build to scale.

Overall, given the ability of the Cox model to both fit the training data in a matter of seconds, and produce event probabilities over any time horizontal, the Cox model would be the clear choice to implement in a clinical setting. As per the principles of Occam's razor, we find that the best model is the simplest one.

Comparing the performance of the Cox model to NEWS within our hospital data,

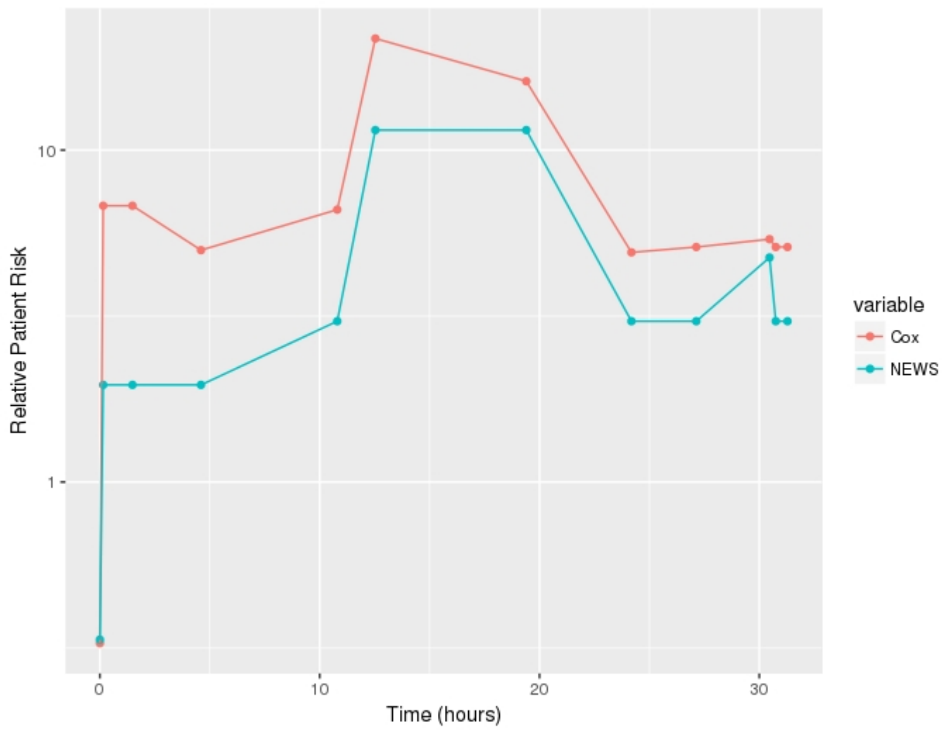

FIGURE 6.1: Relative patient risk for the Cox and NEWS models over a patient's hospital stay.

we find that while the NEWS gives an AUC value of 0.63, even fitting the Cox model on the same variables as NEWS gives an AUC value of more than 0.80. In Figure 6.1, we can see that even for a single patient with an event, our Cox model provides a far higher level of relative risk than the same patient under the NEWS score, and thus gives far better predictive performance.

6.2 An App for Dynamic Prediction

The key difference between a deterministic scoring system such as the NEWS and a dynamic modeling approach is the interpretation of coefficients for individual patients. In NEWS, all the vitals measurements being fed into the score are well known and understood. As there is no reweighing within the scoring system, there is additional output other than the overall NEWS score. At this point, it is up to the individual clinician to decide how to act on a NEWS score based on his or her best judgment.

However, in the case of a dynamic model, there is much that can be gleaned from further exploring the driving factors behind a prediction. In the Cox model, at every time point, we can get a prediction of risk over the next t hour period. At any given point where there is a change in the covariates, we can examine this change to understand the factors that are driving it. For example, a change in predictive probability driven by a demographic variable may not present nearly as much risk as a change in predictive probability driven by a slow but prolonged change in a patient's respiration rate. This type of deep patient level information could be invaluable in improving the decision making process with a hospital environment and helping to mitigate patient risk. Further, while we know, according to the clinical team, evaluated NEWS score are ignored 95 percent of the time, the hope is that providing further insights into the driver behind a new risk score will increase its utility in the hospital setting.

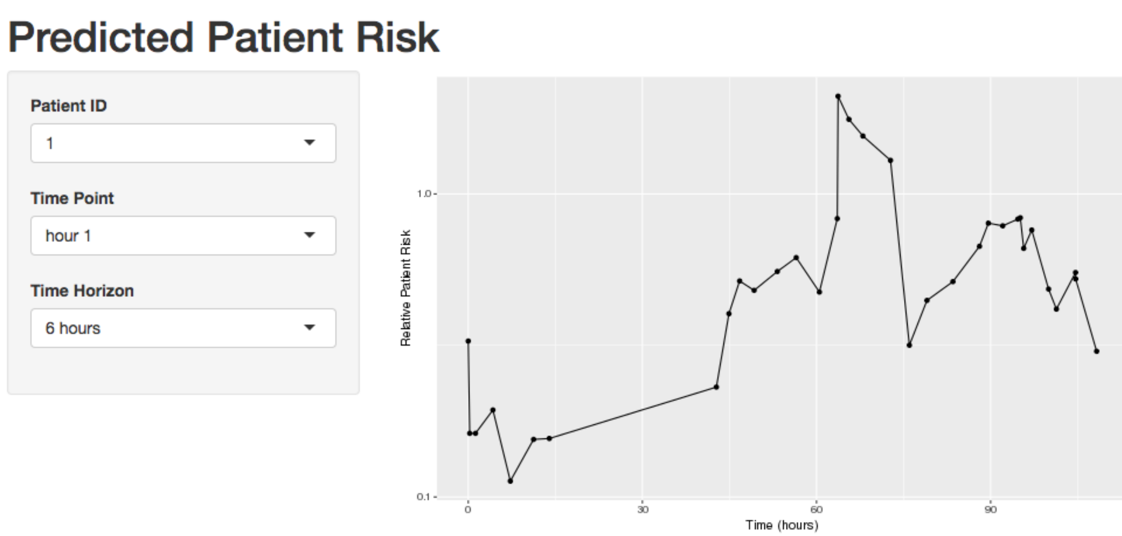

FIGURE 6.2: Prototype app showing dynamic prediction for a single patient during their hospital stay.

To provide this patient level detail, we are in the process of building a fully functioning prototype app as can be seen in Figure 6.2. From the data, this app will allow the user to select any patient that currently has vital information. They then select a time point and the risk forecast window. The app will display the predictive probabilities up until this time point, and at any point where this probability changes, the user can however over this change point for detailed covariate information.

6.3 Further Work

Creating dynamic predictions within a hospital environment is currently at the forefront of many statistics and medical projects. Due to the novelty of the field, there are still many areas that could be further explored. The Cox implementation outlined in this paper provides a great baseline against which to compare further models. Given its computational simplicity, it is also the most promising candidate for implementation within a hospital environment.

As work continues on multivariate joint models, it is likely that a joint model on a patient dataset of this size will become computationally feasible. When it does, it will be interesting to explore to predictive performance of different formations of joint models compared to a Cox proportional hazard model. Given the incorporation of dynamic longitudinal data with the JM scheme, it should be able to outperform a Cox model. The only trade off will be the computational stability and complexity.

Gaussian process models did not prove the best fit for this type of patient data given the sparsity of individual patient measurements. However, as technology continues to improve, we are reaching the tipping point for the advent of cheap and reliable wearable monitors. Even now, a wearable that recorded and transmitted a patient's pulse rate every 30 seconds is perfectly feasible. With this type of unique streaming data, GPs could be used to fit individual predictive curves for each patient based on their unique dynamics and alert a hospital system as soon as these begin to vary. The real advantage of this type of system is the personalization inherent in the GP model. Rather than having to rely on normal population vitals, a GP allows us to compare patient vitals to their own normal levels.

Finally, recent advances in the areas of statistics and machine learning also offer interesting potential models for classifying deteriorating patients. Even with uneven vitals data for each patient, a Markov Chain classification model could allow for a probabilistic outcome on the probability of undergoing an event at any time within a patient's vitals data stream. The key barrier to implementing this type of analysis is that the current literature for model comparison does not allow for a loss function outside of a deterministic ROC curve approach. Further work will focus on developing a robust loss function that could assess the efficacy of models to predict patient outcomes across both the predictive performance of the methods, and the implementation cost with regards to data collection and computation.

Bibliography

Barrett, J. E. and Coolen, A. C. C. (2012), "Gaussian process regression for survival analysis with interval censored data," *Biometrika*, 99, 1–11.

Colopy, G. W., Pimentel, M. A. F., Roberts, S. J., and Clifton, D. A. (2016), "Bayesian Gaussian Processes for Identifying the Deteriorating Patient," *Department of Engineering Science, University of Oxford*.

David W. Hosmer, J., Lemeshow, S., and May, S. (2011), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, John Wiley and Sons.

Diez, D. (2017), *Survival analysis in R*, available online.

Donowitz, L. G., Wenzel, R. P., and Hoyt, J. W. (1982), "High risk of hospital-acquired infection in the ICU patient," *Crit. Care Med.*, 10, 355–357.

Douc, R., Moulines, E., and Stoffer, D. (2014), *Nonlinear Time Series: Theory, Methods and Applications with R*, Chapman and Hall and CRC Press.

Dyagilev, K. and Saria, S. (2016), "Learning (predictive) risk scores in the presence of censoring due to interventions," *Machine Learning*, 102, 323–348.

Harrell, F. E. (2001), *Regression Modeling Strategies*, Springer New York.

Hug, C. W. and Szolovits, P. (2006), *Predicting the risk and trajectory of intensive care patients using survival models*, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science.

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010), "Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data," *Journal of Clinical Oncology*, 16, 2796–2801.

Ranganath, R., Perotte, A., Elhadad, N., and Blei, D. (2016), "Deep Survival Analysis," *JMLR*, 56.

RCP (2012), "National Early Warning Score (NEWS): Standardising the assessment of acute illness severity in the NHS," *Royal College of Physicians*, Report of a working party.

Riedel, S. and McCallumi, A. (2011), "Fast and Robust Joint Models for Biomedical Event Extraction," *Department of Statistics, North Carolina State University*.

Rizopoulos, D. (2012), *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*, Chapman and Hall/CRC.

Rizopoulos, D. (2014), "The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data using MCMC," *ArXiv e-prints*.

Shen, A. (2016), "Developing a prediction model with dynamic Electronic Health Records data," *Duke University, Department of Statistical Science*.

Shiao, H.-T. and Cherkassky, V. (2016), "SVM-Based Approaches for Predictive Modeling of Survival Data," *Department of Electrical and Computer Engineering, University of Minnesota*.

Sudell, M., Kolamunnage-Dona, R., and Tudur-Smith, C. (2016), "Joint Models for Longitudinal and Time-to-Event Data: A Review of Reporting Quality with a View to Meta-Analysis," *BMC Medical Research Methodology*, 16, 168.

Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., and Hall, M. K. (2015), "Emergency Department Patients With Sepsis: A Local Big Data?Driven, Machine Learning Approach," *Academic Emergency Medicine*, 0.1111/acem.12876.

Tsiatis, A. A. and Davidian, M. (2004), "Joint Modeling of Longitudinal and Time-to-Event Data: An Overview," *Statistica Sinica*, 14, 809–834.

Wang, J., Luo, S., and Li, L. (2016), "Dynamic Prediction for Multiple Repeated Measures and Event Time Data: An Application to Parkinson's Disease," *eprint arXiv:1603.06476*.

Wulfsohn, M. S. and Tsiatis, A. A. (1997), "A joint model for survival and longitudinal data measured with error," *Biometrics*, 1, 330–9.