Predicting Intensive Care Unit Patient Outcomes through Patient Similarity

by

Joseph Seung Young Park

S.B., Computer Science and Molecular Biology, M.I.T. (2017)

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

Master of Engineering in Computer Science and Molecular Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Author .....
Department of Electrical Engineering and Computer Science
May 28, 2019

Certified by .....
Roger Mark, Professor, Thesis Supervisor
May 28, 2019

Certified by .....
Jesse Raffa, Ph.D., Thesis Supervisor
May 28, 2019

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Predicting Intensive Care Unit Patient Outcomes through Patient Similarity

by

Joseph Seung Young Park

Submitted to the Department of Electrical Engineering and Computer Science

on May 28, 2019, in partial fulfillment of the

requirements for the degree of

Master of Engineering in Computer Science and Molecular Biology

Abstract

An ICU stay involves invasive treatments, and frequently, the decision to continue therapy is made with limited information based on the physician's personal experience. This thesis proposal describes a tool to assist this decision by identifying similar patients and using their outcomes for prediction. We used the eICU Collaborative Research Database (eICU-CRD) v2.0 for the project. Different time varying and time constant features about the patient's demographics and clinical trajectory was used as input data, such as patient age and longitudinal blood pressure measurement. Using this information, a Cox Proportional Hazards model was built to map the multivariate time series of input data to a univariate time series, which was used to match the patient to a cohort of similar patients. Based on the cohort, this model predicted the probability of a healthy discharge by using the aggregate outcome of the cohort for prediction.

Thesis Supervisor: Roger Mark
Title: Professor

Thesis Supervisor: Jesse Raffa
Title: Ph.D.

THIS PAGE IS INTENTIONALLY LEFT BLANK

Acknowledgments

Thank you to Professor Roger Mark, Dr. Jesse Raffa, and all the members of the Laboratory for Computation Physiology at MIT for your help throughout the project. Especially to Jesse, thank you so much for all your help in every step of the project, from basic questions about R and databases to debugging buggy packages and writing the thesis. You have been an amazing mentor, and this thesis would not have been possible without you.

THIS PAGE IS INTENTIONALLY LEFT BLANK

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most difficult decisions that is made in the Intensive Care Unit (ICU) centers around end-of-life decisions, both for the patient family members and the care providers. Along with ethical, religious, and financial concerns regarding invasive treatments in the ICU, this decision is made even more difficult because the physician often only has their own personal experience to base their recommendation on. Therefore, development of a tool which augments physician experience with the vast troves of electronic health records would improve decision-making at this difficult time. This thesis aims to use the large dataset resources for this purpose by constructing groups of previously seen patients similar to the patient under consideration, using the historic patients' outcome to guide decision-making, and communicate the result effectively to both care providers and family members.

There have been several previous works that have been done to predict patient outcomes. One such method is the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) IV score [13]. This method takes in many different predictor variables (e.g., vital sign and laboratory data) collected during the first day of ICU stay and uses a multivariable logistic regression model to predict the patient mortality rate and the estimated length of stay. This method has good discrimination for the mortality rate (area under the receiver operating characteristic curve, or AUROC, of 0.88). However, there are several limitations of this method.

This method only looks at data collected from the first 24 hours of ICU. Therefore, the prediction cannot be updated for patients whose stay is longer than 24 hours, based on new measurements that are taken during their stay, and has not been validated beyond the 24-hour time point. Also, this model only looks at the "worst" measurement for each input variable. Because the condition of critically ill patients is dynamic and can change greatly within their hospital stay, the trajectory of a patient often cannot be captured by using a single value and may be better represented by the change of the variable through time. It is for these and other reasons that APACHE IV and similar scales are often not used in clinical decision making, but rather for benchmarking hospital performance long after the patient has left the ICU.

A more contemporary example of outcome prediction in the ICU was recently published by Rajkomar et al. using deep neural networks to predict various patient outcomes [7]. This method performed very well overall, achieving high discrimination (e.g. AUROC of in-hospital mortality was 0.93-0.94). The primary outcome prediction was made after 24 hours of admission, but the authors also made predictions ever 12 hours starting 24 hours before admission until 24 hours after admission, showing that the prediction could be updated based on a longer hospital stay. However, a problem with using deep learning methods is that despite its performance, it is can be very hard to interpret the result. For example, if the algorithm predicts that a person has a high risk of dying during their stay, it can be difficult to find out exactly why the algorithm predicted this result. When making decisions which involve life and death, care providers are cautious and want to ensure the prediction has some face validity prior to making any recommendation. This may be difficult to do when predictions are generated from a black box.

Although training a general model on a population could have good performance, these studies typically only provide "the average best choice," and the result may not be appropriate for patients that deviate from that of an average patient. Therefore, it may be useful to instead identify a cohort of similar patients and use their outcome to make a prediction, instead of using a general population-based method. The idea behind using a similar cohort to make a prediction on an index individual, also known as collaborative filtering [9], is widely used in many other fields. Collaborative filtering has also been used in a hospital context. In a review article published in 2017, Sharafoddini, Dubin, and Lee identified 22 such articles, most of which used a neighborhood-based approach for prediction, and in two studies, the similarity-based prediction models outperformed the general population-based models [8].

Most of these approaches will generate predictions as a probability of an outcome, and it has been shown that people have difficulties understanding probability. In a recent study, Hoffrage et al. showed that people understand statistical information better when it is presented in terms of natural frequencies instead of probabilities [2]. That is, study participants understood a statement such as "20 out of 100 people similar to you will not survive" better than saying "20% will not survive." Identification of a similar cohort of patients may also have benefits for the care provider. For a physician, by showing similar matches, they will be able to evaluate whether the matches make sense and whether the prediction is valid or not. This is supported by Eric Topol, who stated in a recent *Nature Medicine* article, "Perhaps the greatest long-term potential of AI in health systems is the development of a massive data infrastructure to support nearest-neighbor analysis, another application of AI used to identify 'digital twins'" [12]. However, this has not been fully implemented yet because of various factors. For example, lack of data, both in terms of the appropriate covariates that are needed to accurately model a patient

and the number of patients in the database to match to, is a large concern, especially when some patient data may be incomplete.

With this in mind, this thesis describes a method to predict patient outcome in ICU that attempts to address these issues. This approach aims to find patients with similar characteristics and clinical trajectories, and use the outcome of these patients to help inform decision making. The overview of the method is shown in Figure 1-1. First, after selecting for patients that meet certain inclusion criteria, different covariates are extracted, which is described in chapter 2. Using a Cox Proportional Hazards (Cox PH or Cox) model, the complex clinical trajectories of many different covariates, such as blood pressure or nurse assessment, are mapped to a univariate time series of the linear predictor, which models the instantaneous likelihood of a patient being successfully discharged at a given timepoint. This process of using the Cox PH model is described in chapter 3. Patients are matched to a cohort of similar patients using the values of the linear predictor, which is much easier due to smaller dimensionality than the initial multivariate measurements, and the outcome of the patients in the similar cohort is used to make a prediction about the patient, described in chapter 4. This method is more interpretable to both physicians and families as it identifies a cohort of similar historical patients and their outcomes. This can help the physician interpret the output by allowing them to evaluate whether the cohort is actually similar to the index patient or not, and it will help the family members understand the risk of the patient better by using natural frequencies instead of probabilities.

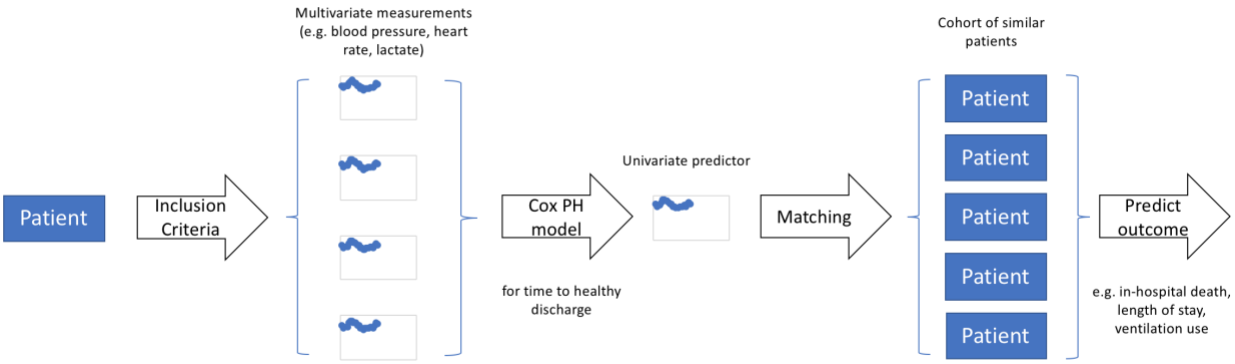

Figure 1-1. Flow chart that outlines the approach, which includes extracting appropriate patient data, mapping to a univariate time series using the Cox Proportional Hazards model, and using the outcome of a similar cohort to make a prediction.

THIS PAGE IS INTENTIONALLY LEFT BLANK

Chapter 2

Data Extraction

2.1 Introduction

In chapter 1, we proposed a multiple outcome patient prediction approach which identifies a cohort of similar patients. In order to create this patient prediction tool, a dataset needed to be chosen first to have a population of patients to match to. In the domain of critical care, there are two open freely available databases of ICU patient stays: Medical Information Mart for Intensive Care (MIMIC) III [3] and eICU Collaborative Research Database (eICU-CRD) [6] datasets from Laboratory for Computational Physiology (LCP) at the Massachusetts Institute of Technology (MIT). MIMIC III was collected from admissions to the Beth Israel Deaconess Medical Center in Boston, Massachusetts in 2002-2011 while the eICU-CRD was collected from 2014-2015 from over 200 hospitals across the United States. Although MIMIC III includes some modalities that do not exist in eICU-CRD, such as waveforms or progress notes, the eICU-CRD was chosen for this project because eICU-CRD contained data from more total patient stays, where the dataset contains data from more than 200,000 patient stays, compared to the ~60,000 patient stays in the MIMIC III dataset. Because the prediction depends on direct matching to patients, having more patient stays for training the model is advantageous since it can increase the specificity of the matches by providing a larger pool of patients that could be

matched. Additionally, even though having the data come from one source means that there is less variation in data collection, having data from a larger number of hospitals could also be beneficial to the model, since it may help identify and remove any bias from specific hospitals and help the prediction tool generalize to more settings.

In the eICU-CRD v2.0, 31 different tables exist that contain different types of data about the patients' ICU stays. Data can be linked across tables using the patient stay identifier, `patientunitstayid`. Time-dependent observations that are collected in eICU-CRD, such as physiological measurements or observations by hospital staff, have an offset attached to the observations, which is when the measurement was taken or recorded in terms of minutes from ICU admission.

Using the unit admission time as a reference, we define offsets as shown in Figure 2-1. t_0 is the offset at which the patient was admitted to the ICU unit, which was at offset 0 for all patients. t_e is the offset at which the relevant ICU stay for this patient was over, either from this patient being discharged from the unit (e.g. to floor or death) or by a code status change for the patient (e.g. comfort measures only), whichever occurred first. t_b is the burn-in offset, which is the period prior to data collection to allow enough time for patients to be properly set up for data collection, and t_c is the collection offset, which is the end of the data collection for the model. t_p was the offset to which prediction was being made to (t_p will be discussed further in chapter 4). For example, consider an index patient that survives the first 48 hours in the ICU. If $t_b = 180$ (3 hours) and $t_c = 2880$ (2 days) and the outcome that is being considered is "healthy discharge" (discussed later in the chapter) at $t_p = 4320$ (3 days), we would collect patient data from $t_b = 180$ to $t_c = 2880$. From this data alone, we then identify a similar cohort of patients who are a close match to the index based on the data from t_b to t_c, and using the outcome of the similar cohort

occurring before time t_p, which would be healthy discharge before $t_p = 4320$ in this example, we would then predict the outcome of the index patient. Out of these four offsets, t_0 and t_e were given by the data, while t_b, t_c, and t_p could be chosen as appropriate. The order of the milestone offsets must be $t_0 < t_b < t_c < t_p$. However, t_e is only required to be greater than t_0. For a specific patient, if $t_e < t_b$ or $t_e < t_c$, the patient would be removed from one of the inclusion/exclusion criteria described either this chapter or chapter 4, respectively. If $t_e < t_p$ and the desired outcome occurred, then the patient would be marked has having the outcome by t_p, and if not, the patient would be marked as not having the outcome by t_p.

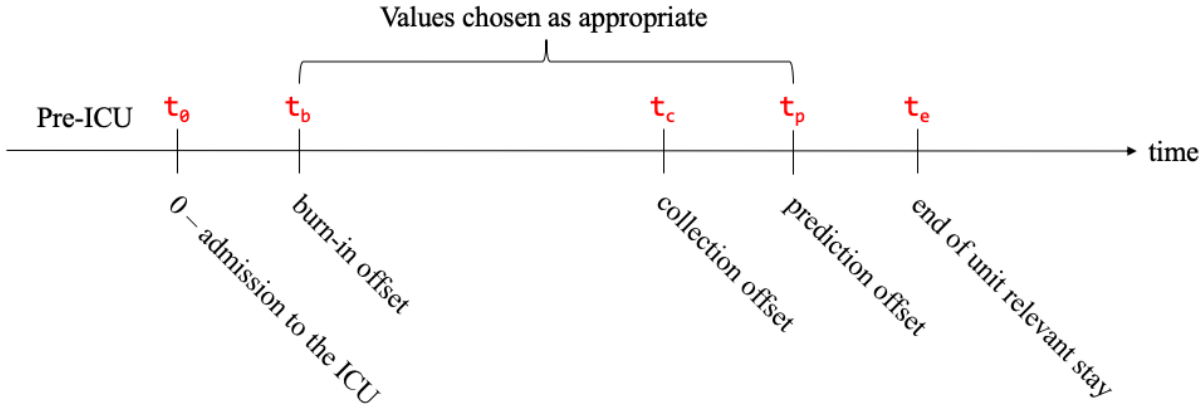

Figure 2-1. Milestone offsets for each patient.

For this thesis, eight distinct eICU-CRD tables were used. The tables and the type of information used from each table are listed in Table 2-1.

Table 2-1. Summary of eight eICU-CRD v2.0 tables that were used for building the model.

Table name	**Information contained / Usage in thesis**
apachePatientResult	APACHE score of each patient, as well as the predicted and actual mortality probability of the patient. The APACHE score (`apachescore`) and the associated version (`apacheversion`) were used as an inclusion criterion to eliminate unit stays which were readmissions or not ICU stays.
apachePredVar	Variables that were used to calculate the APACHE score. The columns that indicated comorbidities (`aids`, `hepaticfailure`, `lymphoma`, `metastaticcancer`, `leukemia`, `immunosuppression`, `cirrhosis`) were used as covariates.
carePlanGeneral	General care plan notes that were inputted by the patient caregivers. Specific notes were selected to gather information about changes to code status, which is the extent to which treatment can be given to patients, by observing the `cplitemvalue` column, and `cplitemoffset` is the offset that the note was taken.
diagnosis	Diagnosis made by the patient caregivers. The `diagnosisstring` column was used to gather information about different groups of diagnoses, and `diagnosisoffset` is the offset when that diagnosis was made.
nurseCharting	Observations about the patient charted by the nurse. This table was used to augment the vital signs data from the vitalAperiodic and vitalPeriodic tables, as well as to obtain the patient Glasgow Coma Scale (GCS) scores. The type of measurement was determined by values in `nursingchartcelltypevalname`, measurements were found in `nursingchartvalue`, and the `nursingchartoffset` is the offset when that charting was made. This table may contain overlap with the vitalPeriodic or vitalAperiodic tables.
patient	General information about the patients. This table was used to extract information about how the patient was discharged from the ICU unit and the hospital (`unitdischargeoffset`, `unitdischargelocation`,

	`unitdischargestatus`, `hospitaldischargeoffset`, `hospitaldischargestatus`), where offset is the number of minutes from the ICU admission, and basic time constant patient demographic variables (`gender`, `age`, `ethnicity`, `apacheadmissiondx`, `unitadmitsource`).
vitalAperiodic	Aperiodic measurements that are directly taken from the monitor without needing to be verified by a caregiver. The non-invasive blood pressure (`noninvasivemean`) was used, and `observationoffset` is the offset when the measurements were taken.
vitalPeriodic	Periodic measurements that are directly taken from the monitor without needing to be approved by a caregiver. The heart rate (`heartrate`), respiration rate (`respiration`), body temperature (`temperature`), and oxygen saturation (`sao2`) were used, and `observationoffset` is the offset when the measurements were taken.

2.2 Methods

The relevant data from the tables were assembled and processed in a format which could be later fit using the Cox Proportional Hazards model discussed in chapter 3, and the overview of the data extraction process is shown in Figure 2-2. First, before extracting the covariates, t_e was determined for all patients. t_e is the end of relevant ICU unit stay of a patient, and it was defined to be the lowest offset of either discharge (i.e. `unitdischargeoffset` from the patient table), or code status change, where a change to code status was defined as the first offset with `cplitemvalue` being one of: “Do not resuscitate,” “No CPR,” “No intubation,” “Comfort measures only,” “No cardioversion,” “No vasopressors/inotropes,” “No augmentation of care,” “End of life,” “No blood products,” “No blood draws,” and “Advance directives.” For all the

time-varying features, data was only collected up to t_e, because any measurements taken after t_e did not affect the outcome of the patient, which was determined at t_e. In some cases, data may be collected on patients prior to t_0, which could come from a patient's stay in the hospital before admission to the ICU (e.g. in the Emergency Department). These pre-ICU data were summarized so that only the last pre-ICU observation of each covariate was included, allowing us to carry forward the pre-ICU data into the ICU for those with data missing early in their ICU stay. Then, the different covariates were extracted from appropriate tables and merged to a single table that contains all covariates for a specific time interval of a patient. Patient intervals are defined by changes in covariate values for time-dependent variables, or the occurrence of the outcome. Covariate values are assumed to be constant within a patient interval. Rolling statistics of some time-varying covariates were calculated and added as additional covariates to capture additional aspects of the patient's clinical trajectory. For example, the rolling median could show a patient's general condition and filter noise, while the minimum/maximum could capture the worst measurements in a time period and serve a different function.

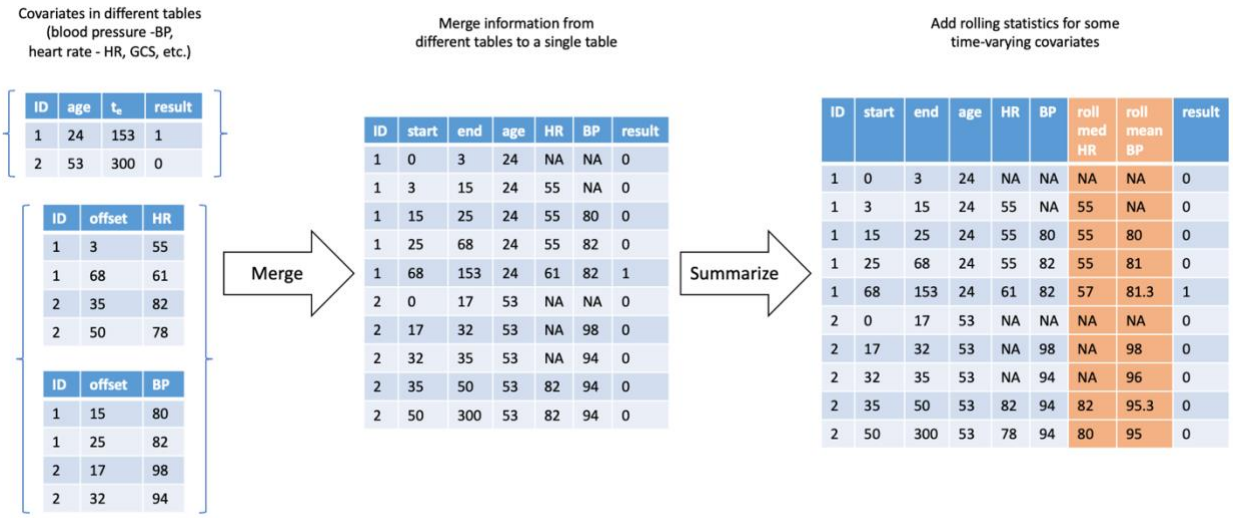

Figure 2-2. Overview of data extraction after initial inclusion criteria with example data.

In the example shown in Figure 2-2, first, each covariate, both time-constant (e.g. age, time of discharge, and the end result of ICU stay) and time-constant (e.g. heart rate, HR, and blood pressure, BP), was extracted from appropriate tables and formatted. Then, all the information in various tables were merged into a single table that showed the value of each covariate for some interval of time, where the new observations of the interval was made at the start, as well as the result at the end of each interval. After merging into a single table, rolling statistics (e.g. mean) of some time-varying covariates (e.g. HR and BP) were calculated to be used as additional covariates.

From the patient table, five time-constant features were extracted: gender, age, ethnicity, the admission diagnosis, and the source of admission. For the gender, the patients were grouped into two categories according to the `gender` column, based on whether or not they had the string "Male" for the value. For age, the value from the `age` column was directly taken, unless it had the value "> 89," in which case the age was set to 90, or an empty string, in which case the age was set to missing. For the ethnicity, the value from the `ethnicity` column was directly taken if it was one of "African American," "Asian," "Caucasian," "Hispanic," or "Native American", and all other values were grouped into "Other/Unknown." For the admission diagnosis, we grouped the patients to 13 groups based on the admission diagnosis used for APACHE (`apacheadmissiondx`), using the diagnosis hierarchy shown in appendix D and E of the ANZICS-APD data dictionary [1], where the 13 groups were: cardiovascular, gastrointestinal, gynecological, hematological, metabolic, musculoskeletal/skin disease, neurological, renal/genitourinary, respiratory, sepsis, trauma, other medical disorders, or undefined. For the source of admission, the value from the `unitadmitsource` column was directly taken if it was

one of “Emergency Department,” “Floor,” or “Operating Room”, and all other values were grouped into “Other.”

From the vitalAperiodic table, the non-invasive mean blood pressure data was collected from the `noninvasivemean` column. This was augmented by data from the nurseCharting table by taking data from rows with the column `nursingchartcelltypevalname` having the value “Non-Invasive BP Mean.” If there were multiple measurements for a specific patient at the same offset (e.g. because of the possible redundancy between the vitalAperiodic and nurseCharting tables), the median of those values was taken. For measurements taken before t_0, the value associated with the highest offset was taken, and the offset was set to be at 0. This was done so that we could populate pre-ICU data when ICU data was not readily available early in the ICU stay, and this is similar to how we used a burn-in period (t_b), described later in the thesis.

From the vitalPeriodic table, data about heart rate, respiration rate, body temperature, and oxygen saturation (SaO_2) were taken from columns `heartrate`, `respiration`, `temperature`, and `sao2`, respectively. The processing of this data was done in a similar fashion to that of mean blood pressure. First, this data was augmented by data from the nurseCharting table by taking data from rows with the column `nursingchartcelltypevalname` having the values “Heart Rate,” “Respiratory Rate,” “O2 Saturation,” “Temperature (C),” or “Temperature (F).” While most variables had consistent units across hospitals and patients, temperature was recorded in both Fahrenheit and Celsius. To achieve consistency, the temperatures from the vitalPeriodic table was converted to Celsius by treating any values ≥ 50 as Fahrenheit, and any below as Celsius. This conversion was more straight-forward for data from the nurseCharting, since the column “nursingchartcelltypevalname” identified whether the measurement was Celsius or Fahrenheit. After combining these sources of data, if there were multiple measurements for a specific patient

at the same offset, the median of those values was taken. For measurements taken before t_0, the value associated with the highest offset was taken, and the offset was set to be at 0.

From the nurseCharting table, in addition to using it to supplement the data from vitalAperiodic and vitalPeriodic tables, Glasgow Coma Scale (GCS) scores were taken by looking for observations with the column `nursingchartcelltypevallabel` having the value "Glasgow coma score." The GCS contains three components (eyes, motor, verbal), and higher score corresponds to better health of patients, with the maximum scores being 4 for eyes, 5 for verbal, and 6 for motor, and this serves as an indicator of how critically ill a patient is [10]. Instead of looking at the raw score of each component, the measurements were grouped into four categories: all low (each component having a score of 1, indicating no response in the eye, motor, or verbal scores), all high (each component having the highest scores, indicating the highest level of response), some high (some of the components having the highest score, but not all), or intermediate (not any of the above). As done in other tables above, for measurements taken before t_0, the value associated with the highest offset was taken, and the offset was set to be at 0. However, for this table, until the patients' first GCS assessment, a new category of missing was introduced, which was used until the first GCS assessment was done, for a total of five categories (all high, some high, intermediate, all low, or missing). This was done to make sure that all patients had some observation for GCS.

The diagnosis table contains information about the presence and onset of a diagnosis. From this table, data about different diagnoses were taken by looking at the column `diagnosisstring`, which was formatted from the most general to the specific, separated by a vertical bar ("|"). First, all diagnoses were grouped by looking at the first component (the most general) of the `diagnosisstring` column, which was the string before the first vertical bar. For

example, for a patient whose `diagnosisstring` contained the value "cardiovascular|arrhythmias|atrial fibrillation|with hemodynamic compromise.," they would be put into the "cardiovascular" group. This divided the diagnosis into 17 different groups, which were: "burns/trauma," "cardiovascular," "endocrine," "gastrointestinal," "general," "genitourinary," "hematology," "infectious diseases," "musculoskeletal," "neurologic," "obstetrics/gynecology," "oncology," "pulmonary," "renal," "surgery," "toxicology," and "transplant." Additionally, three other groups of diagnosis were constructed for 3 common ICU diagnoses: sepsis, pneumonia, and arrhythmia. For these three groups, the value contained in the column `diagnosisstring` had to match specific values: matching either "sepsis" or "septic" while not matching "aseptic" for sepsis; matching "pneumonia" for pneumonia; and matching one of "arrhythmia," "fibrillation," "tachycardia," "bradycardia," "flutter," or "av block" for arrhythmia. All patients started as not having the diagnosis for each of the 20 groups at offset 0, unless there was a diagnosis for a specific group before t_0, in which case they started as having the diagnosis at offset 0. Then, if a patient had a diagnosis for any of the groups after t_0, they were marked as having the diagnosis for the rest of their ICU stay, starting at the lowest offset at which the diagnosis was recorded. Therefore, a patient could have more than one diagnosis, and additional diagnoses of the same group did not change this covariate.

From the apachePredVars table, data about common comorbidities were taken by looking at the columns `aids`, `hepaticfailure`, `lymphoma`, `metastaticcancer`, `leukemia`, `immunosuppression`, and `cirrhosis`. These comorbidities are used in APACHE because they are known to affect outcomes in the ICU. Because these were comorbidities that the patients already had before being admitted into the ICU, the presence or absence of these each comorbidity was recorded at offset 0 for each patient.

After processing all the covariates from the different tables described above, the data were combined using the tmerge function from the R survival package [11], setting the covariates as time-dependent covariates (tdc) at the offset at which the observation was made. As shown in Figure 2-2, this function makes a new row for each interval from the time that an observation was made to the time that the next observation was made (or t_e, where the observation is the end of the patient stay) for any of the covariates discussed above. Also, for each interval, the status of the patient at this timepoint was recorded, where the status of the patient was a binary factor that depended on whether the patient had a healthy discharge at the end of the interval. Here, healthy discharge was defined as:

1. Unit discharge to home, or
2. Unit discharge to floor, with
    a. Live discharge from hospital, or
    b. Staying alive for at least 6 hours following the unit discharge.

This outcome was chosen over others, such as death or discharge, because the purpose of the thesis was to build a tool that assisted in decision making for doctors and patient families, and healthy discharge from the ICU may be of more interest than other outcomes. From the figure, we can see that patient 1 had a healthy discharge at offset 153 while patient 2 did not have a healthy discharge and had a censoring event at offset 300, either from having a discharge that was not considered a healthy discharge or from a code status change.

Before getting the statistical summaries of some measurements, the combined data containing intervals of stay for each patient was cleaned using the burn-in offset (t_b), set at t_b = 180 (3 hours), and collection offset (t_c). First, patients for whom the end of the relevant patient stay ended before the burn-in (i.e. $t_e < t_b$) were removed. Thus, patients with short stays, who

contribute little data related to their trajectory and play little significance to our main use case, were not used. For the remaining patients, several steps were taken to restrict their patient stay data to between t_b and t_c. From the burn-in period (between offsets t_0 and $t_b + 1$, inclusive), only the last interval was taken and the start of this interval was set at $t_b + 1$. The burn-in period was put in place because patient data will generally not be immediately collected or available upon admission to the ICU unit. Some time is needed in order for the patient to be ready for data collection, such as being attached to monitors and having assessments completed by hospital staff, and only taking the last measurement from the burn-in allows us to have the last-known data while allowing sufficient time for data to be gathered. All intervals where start $> t_b + 1$ and end $\leq t_c$, the period that we are matching to, were taken. For intervals where start $< t_c$ but end $\geq t_c$ and t_e, the end of the interval was set to t_c and status of the patient at this time was set to not having a healthy discharge at this time.

Then, for three of the vital signs measurements (blood pressure, heart rate, and respiration), three different rolling statistical summaries (minimum, maximum, and median) were collected as covariates. The rolling statistics were collected by looking back to recent measurements that are within a certain amount of time, and applying the summary statistic to all data within this interval. Specifically, for each time interval ($t_{start,1}$, $t_{end,1}$) of a patient, the rolling statistics was calculated using other intervals of the same patient ($t_{start,2}$, $t_{end,2}$), where the $t_{start,1} - t_r < t_{start,2} < t_{start,1}$ for $t_{r,min}$, $t_{r,max}$, or $t_{r,median}$ (for the rolling minimum, maximum, and median, respectively). t_{start} was used instead of t_{end} because new observations were made at t_{start}. For this thesis, $t_{r,min} = t_{r,max} = t_{r,median} = 60$ (1 hour).

Finally, before fitting different models to the constructed dataset, other steps were done including splitting the patient data into the training (80% of patient unit stays) and testing (20%)

sets. The 80% of the patients that were included in the training set were further split into five folds for cross-validation. For each cross-validation fold, one fold was chosen as the validation set (while all others were used as the training set) for the model for each iteration. After computing the result for each fold, the set of tuning parameters that had the best mean AUROC across the five folds was chosen, which will be explained in chapter 4. This was done in order to ensure that our model did not overfit and performs well on patients that it has not seen previously.

2.3 Results

In several steps of the data extraction process, the patients were filtered to only keep patients that met certain criteria. These inclusion/exclusion criteria, and the number of patients remaining in the study population after each step, are shown in Figure 2-3. This figure was calculated using t_b =180, and the third filtering step, where patients whose $t_e < t_b$ were excluded, could vary depending on the chosen t_b. We used t_b = 180 (3 hours) for the remaining analysis, and at the end, the patients who met the inclusion criteria were divided into the training and the testing set, where the training set contained 101,551 patients, and the testing set contained 25,388 patients.

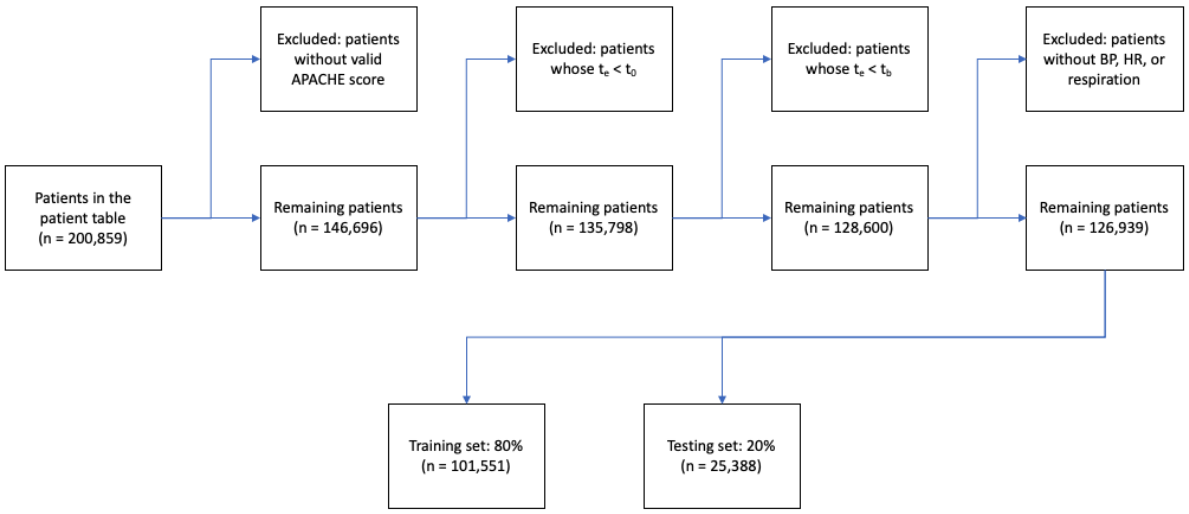

Figure 2-3. Various filtering steps taken during patient data extraction for t_b = 180 (3 hours).

A summary of the patients who met the inclusion criteria is shown in Table 2-2.

Table 2-2. Characteristics of patients that met the inclusion criteria, n = 126,939.

Sex	**Number of patients (n)**	**Proportion of patients (%)**
Female	57,678	45.44
Male	69,227	54.54
Other	34	0.02
Ethnicity	**Number of patients (n)**	**Proportion of patients (%)**
African American	14,843	11.69
Asian	1,814	1.43
Caucasian	96,993	76.41
Hispanic	5,075	4.00
Native American	894	0.70
Other	7,320	5.77
Types of units		
Cardiac ICU	8,842	6.97

CCU-CTICU	11,007	8.67
CSICU	3,935	3.10
CTICU	4,047	3.19
Med-Surg ICU	69,910	55.07
MICU	10,860	8.56
Neuro ICU	10,003	7.88
SICU	8,335	6.57
	Median	**Q1 (25%) – Q3 (75%)**
Age	64	52 – 75
	Mean	**Standard deviation**
APACHE IV Score	54.2	25.2

A plot of the number of patients for whom an observation was collected during a certain hour after being admitted to the ICU can be shown in Figure 2-4. Here, we can see there is a small bump for the very first hour after the burn-in (at hour 3), showing that using the burn-in was effective in capturing more patients. We can see the number of measurements steadily decrease over time, which is due in part to patients being discharged from the ICU unit or because many patients do not need to be monitored as closely after the first several hours.

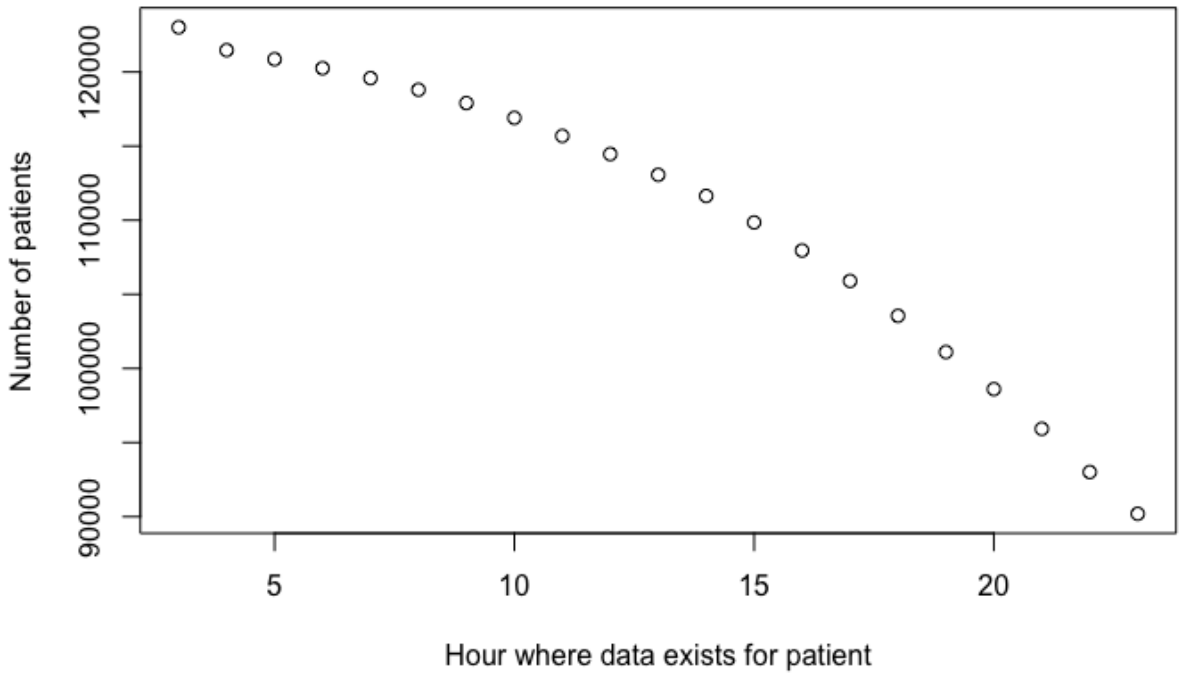

Figure 2-4. The distribution of the number of patients that met the inclusion criteria, for whom a measurement was taken during the hour over time after admission to the ICU.

A Kaplan-Meier [4] plot examining the time until healthy discharge for the patients who met the inclusion criteria is shown in Figure 2-5. However, one thing to note is that the outcome that was being observed was healthy discharge instead of the usual outcome of death, which means that the proportion that is shown on the y-axis is the proportion of patients who did not have a healthy discharge, not the proportion that have survived. The curve shows drops at times that correspond to full days after admission to the ICU – at around offsets 1440, 2880, and 4320. This could be because patients who are admitted to the ICU and not immediately discharged (who would have not met the criteria with the burn-in) are generally admitted during the day and both the patients and the hospital staff may prefer having the patient being discharged after confirming their condition the next day.

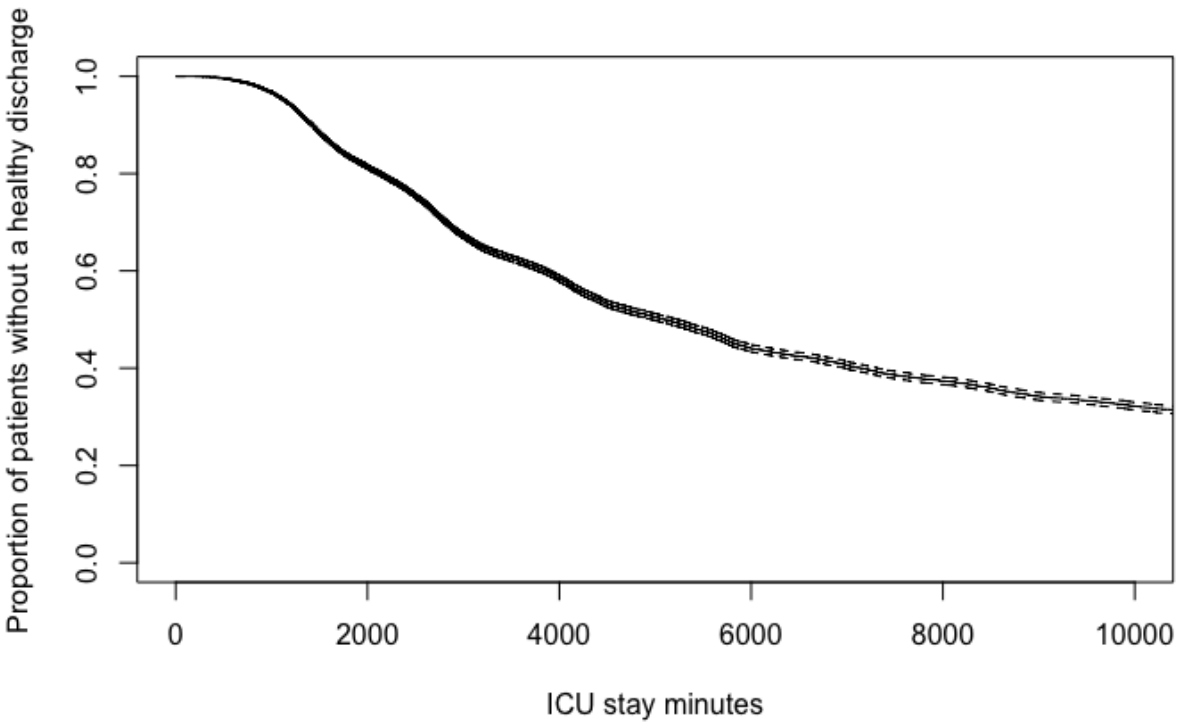

Figure 2-5. Kaplan-Meier plot that shows the proportion of patients who have not had a healthy discharge after admission to the unit.

2.4 Conclusion

Because of both time and computational constraints, not all features were able to be used. One aspect of this was not being able to use all the available and possibly useful covariates. For example, the lab table has many informative lab test results, such as serum lactate levels, which could be indicative of patient health and useful for modeling the patient outcome. However, they were not used because many patients were missing these results, and requiring patients to have these lab results would limit the usability of the approach. Similarly, another concern was the dimensionality of the data frame. Because each new observation added a row, while new covariates that were introduced added a column, the data frame became both tall and wide,

which limited in the number of observations that we could collect for a patient, as well as the number of covariates that we could collect. Despite this, we were able to extract the data from all eligible patients in eICU-CRD using the variables described in this chapter.

Chapter 3

Modeling ICU Discharge

3.1 Introduction

As seen in chapter 2, the ICU is a data-rich environment, containing important time-constant (such as age or gender) and time-varying variables. The time-varying covariates were features that changed from one interval to the other, which included: 20 groups of diagnosis, GCS score, mean non-invasive blood pressure, heart rate, respiration rate, temperature, SaO_2, and the rolling statistical summaries (minimum, maximum, median) of the three measurements (blood pressure, heart rate, and respiration rate). The time-constant covariates were features that remained constant for all intervals of a patient's ICU stay, which included: age, gender, ethnicity, source of admission, groups based on the admission diagnosis, and the seven comorbidities. As described in chapter 1, the goal of this thesis was to match an index patient to a cohort of similar patients whose outcome could be used to predict the outcome of the index patient. In order to achieve this goal, the distance between patients need to be calculated. However, using the data extracted as described in chapter 2 to calculate the distances would be difficult. For example, calculating the distance by weighing each covariate equally does not seem correct, especially given that the variables are found in different formats (binary, numeric, or categorical), and attempting to calculate the weights separately seems infeasible. Another factor that could make

this difficult is that, for the time-varying covariates, measurements were taken at different time for different patients. Therefore, the multivariate time series of a patient need to be reformatted to be used in calculating patient distances.

A good approach to address these issues may be to map these collected features to the incidence of important clinical outcomes, which would provide us with a lower dimensional time series. A low dimensional time series can then be used to match patients with similar trajectories. This approach addresses the above issues: first the mapping can have common support, and dealing with discrete and numeric data need not be a concern during matching; second, it weights the variables in terms of which are important for the clinical outcome, which is at least better than using equal weights; and finally, common observation times is less of a concern as interpolation/smoothing can be used on the single time series, instead of trying to figure it out for each variable. With this in mind, we developed a model which models the risk of healthy discharge as a function of important variables so that the model's output can be used to match trajectories of patients and use the matches for prediction.

3.2 Cox proportional-hazards (PH) model

3.2.1 Overview of the Cox PH model

Considering these problems, the Cox proportional-hazards (Cox PH) model was chosen to model the patient discharge from the ICU [5]. The Cox PH model is frequently used to model time-to-event outcomes and can accommodate time-varying covariates. A Cox PH approach models the hazard of an event as follows:

$$h(t, \boldsymbol{X}) = h_0(t) e^{\sum_{i=0}^{p} \beta_i X_i}$$

where h(t,X) is the hazard of a certain outcome at a time *t*, **X** is a vector of possibly time-varying explanatory variables, and β are coefficients associated with the different explanatory variables that will be fitted to the data. This approach models the hazard, h(t,X), the instantaneous probability of a certain outcome at time *t*, given that the outcome has not happened prior to time *t*. The hazard, h(t, **X**), is split up into two components: a baseline hazard, $h_0(t)$, that only depends on the time and is often considered a nuisance, and a component that only depends the explanatory variables:

$$e^{\sum_{i=0}^{p} \beta_i X_{i,t}}$$

This time-independent component is the key part in the Cox PH model, and using the appropriate weights, $\boldsymbol{\beta}$, which are calculated when fitting the model, this component maps the explanatory variables, **X**, to a single predictor. It is important to note that **X** is a function of *t*, in a sense that the values of the explanatory variables change for different times, but the coefficient $\boldsymbol{\beta}$ must be completely time-independent and must remain constant for a given *i*. The part that is exponentiated

$$\sum_{i=0}^{p} \beta_i X_{i,t}$$

is known as the linear predictor, or lp. Usually, the outcome that the Cox PH model models is the time of a negative outcome, such as death, which is why the instantaneous probability is referred to as the hazard. However, in our case, the outcome is the time of healthy discharge from the ICU, as defined in chapter 2, and correspondingly, the hazard is the instantaneous risk of healthy discharge, a good outcome. Therefore, a high value of the lp would indicate that the instantaneous probability of healthy discharge at the given *t* is high, while a low value of the lp would indicate that this probability is low. At the end of the relevant ICU unit stay of a patient,

t_e, the patients who did not have a healthy discharge were censored, at which point the outcome of the patient, such as death or code status change, was determined.

An important assumption that Cox PH model makes is the Proportional Hazards assumption: the hazard ratio (HR) must remain constant over time, where HR is defined as:

$$HR = \frac{h(t, \boldsymbol{X}^*)}{h(t, \boldsymbol{X})} = \frac{h_0(t)e^{\left(\sum_{i=0, i\neq k}^{p} \beta_i X_{i,t}\right) + \beta_k (X_{k,t}+1)}}{h_0(t)e^{\sum_{i=0}^{p} \beta_i X_{i,t}}} = e^{\beta_k}$$

and $\mathbf{X}$ and $\mathbf{X}^*$ are the same for every variable except one, $X_{k,t}$, for which $X^*_{k,t} = X_{k,t} + 1$. However, in this thesis, the Cox PH model was not used as the algorithm itself to predict the outcome of patients, but rather as a conduit to both reduce dimensionality of the problem and weigh the importance of each covariate to later allow matching for patients, so this assumption was not validated. If we identified problems with this approach, where performance deteriorated in a subset of patients, particularly as a function of time, violation of PH may indicate an alternative model maybe indicated to resolve this issue.

Using the Cox PH model solves many of the issues that were identified in the introduction. For the time-constant features, we could model them in the same way as time-varying features, while keeping the values constant. This model also calculates the weights separately for each covariate, so the model is able to take in many different kinds of variables, which may have different scales as well. Importantly, the Cox PH model is able to take the multivariate measurements as input and output a single predictor for each patient, which effectively reduces the dimensionality of the problem.

In the context of the overall pipeline, a Cox model was trained on each out-of-fold training set, and after calculating the linear predictor using the Cox model and reformatting so that the time series of the linear predictor is at the same time points for all patients, patients were matched to a cohort of similar patients, which will be discussed in chapter 4. During training,

several different Cox PH models were fitted, including training on all patients together or training on ten APACHE admission diagnosis groups separately. These ten admission diagnosis groups were the groups described in chapter 2, excluding three groups: gynecological, other medical disorders, or undefined. The ten admission diagnosis groups used were: cardiovascular, gastrointestinal, hematological, metabolic, musculoskeletal/skin disease (referred as just musculoskeletal), neurological, renal/genitourinary (referred as just renal), respiratory, sepsis, and trauma. When training the Cox PH model on both the complete patient set and the on the specific admission diagnosis groups, two different values of t_p, 1440 (1 day) and 2880 (2 days), were used.

3.2.2 Cox PH model trained on complete patient set

When training on all groups together, the covariates (formatted as described in chapter 2) that were used were: age, gender, ethnicity, source of admission, admission diagnosis group, comorbidities (`aids`, `hepaticfailure`, `lymphoma`, `metastaticcancer`, `leukemia`, `immunosuppression`, and `cirrhosis`), 20 groups of time-varying diagnoses ("burns/trauma," "cardiovascular," "endocrine," "gastrointestinal," "general," "genitourinary," "hematology," "infectious diseases," "musculoskeletal," "neurologic," "obstetrics/gynecology," "oncology," "pulmonary," "renal," "surgery," "toxicology," "transplant," "sepsis," "pneumonia," and "arrhythmia"), an interaction between GCS group and the admission diagnosis group "Neurological", temperature, oxygen saturation, non-invasive mean blood pressure, heart rate, respiration rate, as well as the rolling statistical summaries (median, minimum, maximum) of the last three vital signs measurements (blood pressure, heart rate, and respiration rate). For the rolling median of the three measurements, we used a natural cubic spline with degree of freedom

of 2 and looked at the interaction between the spline functions (for each of blood pressure, heart rate, and respiratory rate) and the diagnosis group.

3.2.3 Admission diagnosis group-specific Cox PH models

When training on the admission diagnosis groups separately, the covariates that were used were: age, gender, ethnicity, source of admission, seven comorbidities (`aids`, `hepaticfailure`, `lymphoma`, `metastaticcancer`, `leukemia`, `immunosuppression`, and `cirrhosis`), 20 groups of time-varying diagnoses ("burns/trauma," "cardiovascular," "endocrine," "gastrointestinal," "general," "genitourinary," "hematology," "infectious diseases," "musculoskeletal," "neurologic," "obstetrics/gynecology," "oncology," "pulmonary," "renal," "surgery," "toxicology," "transplant," "sepsis," "pneumonia," and "arrhythmia"), GCS group, temperature, oxygen saturation, non-invasive mean blood pressure, heart rate, respiration rate, as well as the rolling statistical summaries (median, minimum, maximum) of the last three measurements (mean blood pressure, heart rate, and respiration rate). For the rolling median of the three measurements, we used a natural cubic spline with degree of freedom of 2. The difference between training on all groups together and on the admission diagnosis groups separately was that admission diagnosis group, and any interactions involving the admission diagnosis group, were not used as a covariate. However, because a different Cox model was fitted for each admission diagnosis group, every model parameter could vary between different admission diagnosis groups. This means that the relative importance of any of the covariates could be different between different models, which was not the case with the model fit on the complete patient set (described in section 3.2.2). For example, for the trauma admission

diagnosis group, having a diagnosis related to a clotting disorder could have a much higher weight in Cox model fit to this group than for the metabolic admission diagnosis group.

After training on each out-of-fold training set, we chose the best set of parameters for patient matching, described in chapter 4. A final set of Cox models, one for each admission diagnosis group, were trained on the overall training set, which was evaluated on the testing set.

3.3 Results

Several different Cox PH models were fit, including training on all patients together at t_c = 1440 and 2880 (1 day, and 2 days). Even though there were some differences, there were some notable similarities between the two models. One of the similarities was, from the time-constant admission diagnosis groups, when compared against the endocrine group, neurologic, and toxicology showed up as statistically significant at $p < 0.001$ with a positive log(HR), indicating that being diagnosed to one of these groups was a positive indicator for healthy discharge. On the other hand, the groups pulmonary, renal, and pneumonia showed as statistically significant at $p < 0.001$ with a negative log(HR), indicating that being diagnosed to one of these groups was a negative indicator for healthy discharge.

For the three physiological measurements with the rolling statistics, which were mean blood pressure, heartrate, and respiration rate, the rolling maximum had a negative log(HR) while the rolling minimum had a positive log(HR). This result also seems intuitive because there is an ideal range for all of these measurements. Therefore, having a higher minimum and lower maximum indicates that the patient is closer to that ideal range and more likely to be healthier than other patients. However, one exception was the heartrate, where the minimum was not statistically significant.

In order to examine the functional form of the natural cubic spline terms, we took a random example interval for a patient, and while keeping all other covariates the same, varied the rolling median values for the three measurements (blood pressure, heart rate, respiratory rate) and tested calculated the lp for all the Cox models that were fitted (the 10 models trained on specific admission diagnosis groups, as well as the model trained on the complete patient set). Two different t_c were used in training the Cox models, which were also examined. The results are shown in Figures 3-1 to 3-3. We expected the effect of the rolling median measurements to be similar across the different admission diagnosis groups and have a concave down shape, since the likelihood of discharge should be the highest at a normal value but lower when the value deviates, but this was not the case. For example, in Figure 3-3, the effect of median respiratory rate on the Hematological group with $t_c = 1440$ (24 hours) seems to be opposite to that on the Respiratory group with $t_c = 2880$ (48 hours). For the plots that show the "all patients" group, which was the Cox model trained on the complete patient set and not any particular admission diagnosis group, it is not a linear plot because the Cox model was fitted on the interaction between the median measurements and the admission diagnosis groups, and the admission diagnosis group was varied when creating this plot. The shaded area shows the range of effects of the median vital signs measurements on the lp, across the different admission diagnosis groups, for the Cox model fitted on the complete patient set. In other words, the shaded area shows the effect profiles of median vital sign measurements on the lp across the 13 different admission diagnosis groups.

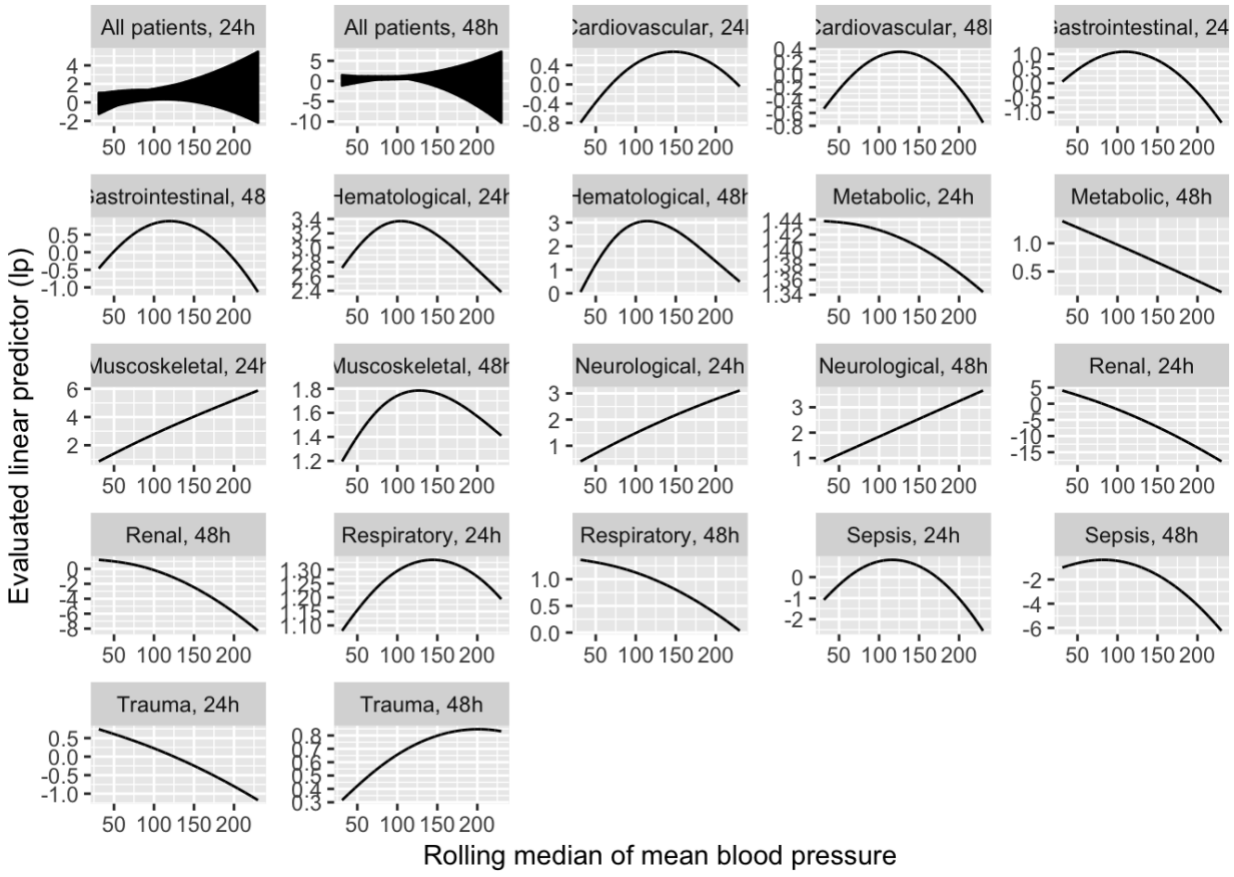

Figure 3-1. Effect of changing median blood pressure on the linear predictor term across different Cox models fit on different admission diagnosis groups, where “all patients” was fit on the complete patient set instead.

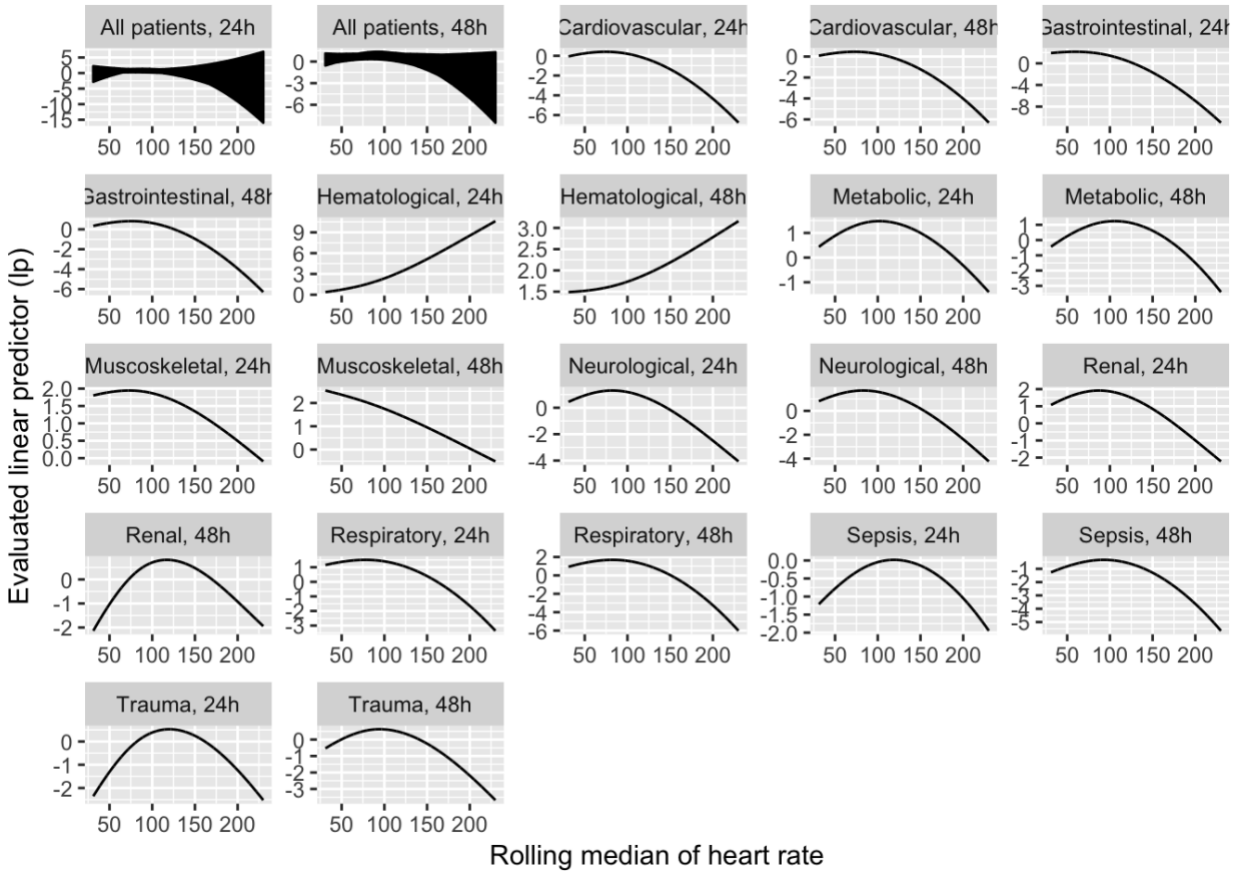

Figure 3-2. Effect of changing median heart rate on the linear predictor term across different Cox models fit on different admission diagnosis groups, where “all patients” was fit on the complete patient set instead.

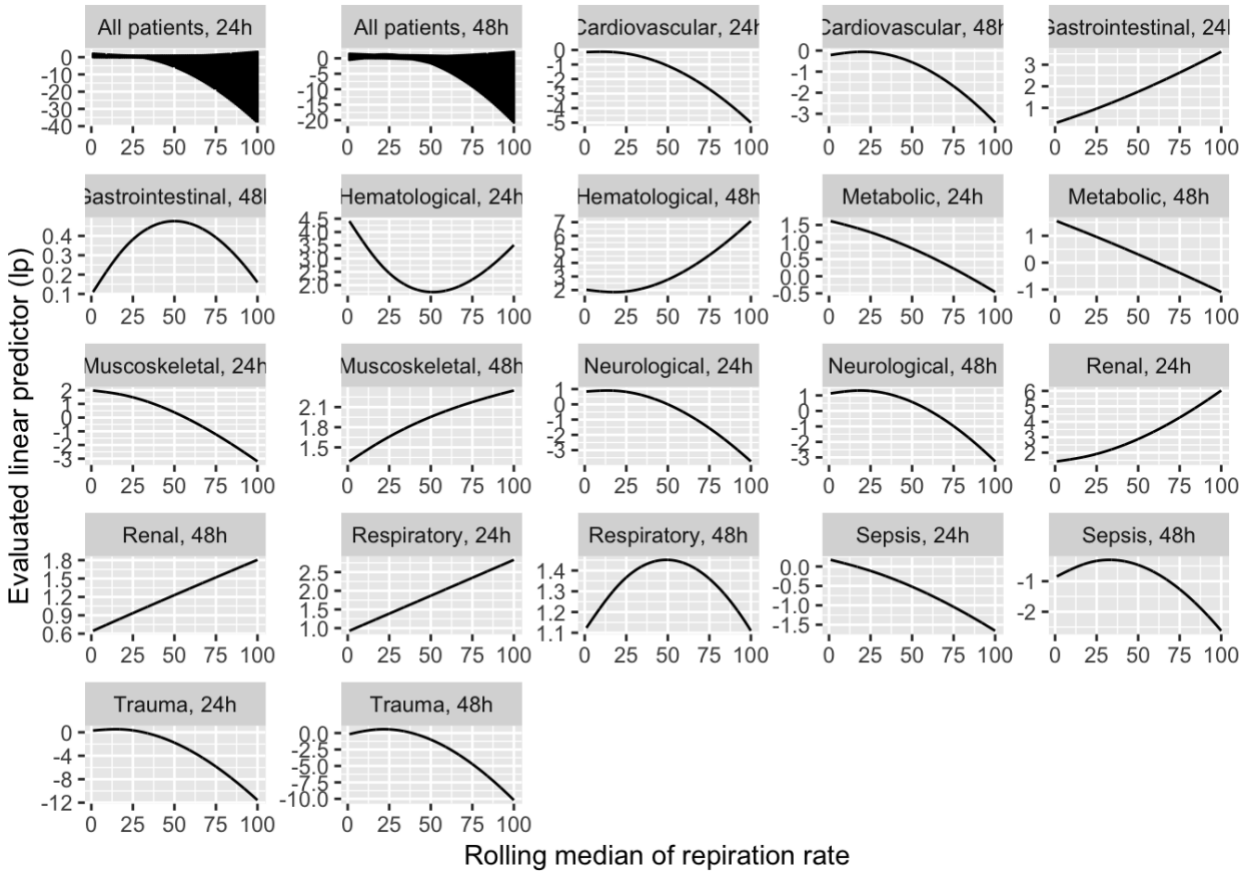

Figure 3-3. Effect of changing median respiratory rate on the linear predictor term across different Cox models fit on different admission diagnosis groups, where "all patients" was fit on the complete patient set instead.

We also trained on the different admission diagnosis group (discussed in chapter 2) at t_c = 1440, except for the groups gynecological, other medical disorders, and undefined. However, some groups were too small to have any useful results, such as gastrointestinal and hematological. However, it could still be observed that the relative weights in the Cox PH model was very different from one group to another, which may mean that, if we have enough patients to train on, restricting the patients to certain groups, such as the admission diagnosis groups, could be beneficial, instead of having all patients in the same training model.

The final model that was chosen was training on $t_c = 2880$ (2 days), based on training results discussed in chapter 4, and a separate Cox model was fit on each admission diagnosis group. The coefficient of each covariate in the models are shown in appendix A.

3.4 Conclusion

When inspecting the result of the Cox PH models that were constructed, there parts of the model that did not make sense. For example, some of the covariates had a 95% confidence interval from 0 to infinity, which could signal some problem with the model, such as collinearity or small sample size. From Figure 3-1 to 3-3, we expected the plots to look mostly similar across the different groups, but a lot of variations were found in the relationship between the median measurements and the linear predictor term. This could be because some of the diagnosis specific groups had a very small number of patients to train on.

Chapter 4

Patient Outcome Prediction

4.1 Introduction

Having mapped the multi-dimensional time series of covariates to a single-dimensional time series using the Cox PH model, we could use the time series of the predictor term, the linear predictor (lp):

$$\widehat{lp} = \sum_{i=0}^{p} \widehat{\beta}_i X_i$$

in order to calculate the distance between patients, which we can then use to find a cohort of similar patients and make a prediction on the patient outcome. There are many different types of matching and clustering algorithms available, but out of the possibilities, we used a k-nearest neighbor approach for this thesis.

There are several advantages of a k-nearest neighbor approach which makes it ideal for this thesis. First, a k-nearest neighbor prediction works best when the training set is large since there will be more example patients that the index patient can try to match to, leading to higher likelihood of finding a good match. Because the public database of eICU-CRD contains more than 200,000 patients, with the private database containing more than 3 million patients, an approach that benefits from having a large testing set seems appropriate. However, the most

important reason for using k-nearest neighbor was because of the interpretability of results. As discussed in chapter 1, the purpose of this thesis was to develop a tool that was more easily understandable for both the hospital staff and the patient's family. By identifying similar patients, the physician will be able to interpret the results more easily and choose to either accept the output, or decide that the tool may not be appropriate for specific patients. Moreover, presenting the prediction in terms of similar patients in a statement, such as "20 out of 100 people similar to you will not survive," will also help the patient and the family understand the prediction better.

4.2 Methods

Training was done on each cross-validation fold separately. First, a Cox PH model was trained on the fold training set (the patients in the training set who are not in the selected fold). The model was then used to predict the lp for each interval (t_{start}, t_{end}) from the data structure obtained at the end of chapter 2. From this point, instead of using the entire interval, only t_{start} was used, since new observations were made at t_{start} and the covariates stayed constant for a given interval. However, because the predictor values were calculated at different offsets for each patient (one predictor value being generated for each t_{start}, where patients do not have common t_{start} values), the predictor value had to be processed first before it could be used to calculate the distance between patients. To achieve this, the mean predictor value for each hour from t_b to t_c (i.e. all hourly intervals from (t_b, t_b+60] to (t_c-60, t_c]) was determined. If a patient did not have any data from any specific hour interval, then the mean predictor value was imputed by carrying forward the last existing hourly mean predictor value. An example of this process is shown in Figure 4-1.

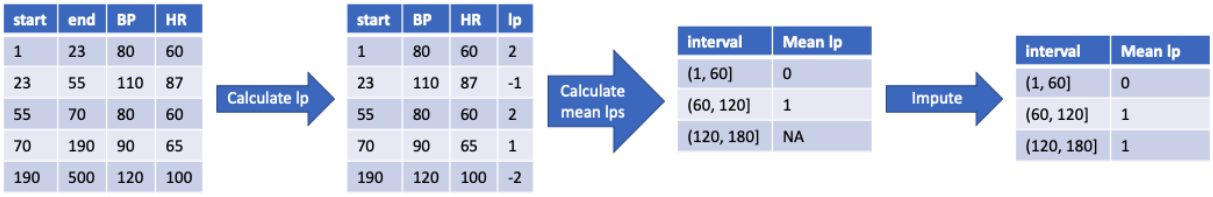

Figure 4-1. Example of using data from chapter 2 to obtain mean predictor values, using $t_b = 0$ and $t_c = 180$.

The distance between each pair of patients was determined using the calculated hourly mean predictor values. Then, for each patient, the pairwise distance between all other patients was determined and a cohort of similar patient was then chosen by choosing patients with the smallest distance to the index patient. Using the hourly mean predictor values from $t_b + 1$ to t_c, the outcome of the individuals in the similar cohort was used to make a prediction about the outcome of the index patient at a specific time, t_p (the offset after t_c at which the prediction is being made, shown in Figure 2-1). The predictor value for the index patient was the number of patients in the similar cohort who had a defined outcome at time t_p divided by the number of patients in the similar cohort. In this thesis, the defined outcome was healthy discharge, but this can be varied as appropriate.

During training, two different methods were used, described in chapter 3. In the first approach, a Cox model using all patients together was trained and the outcome of the index patient was determined using all patients that met the inclusion criteria. The second approach divided the patients by their APACHE admission diagnosis groups, trained a separate Cox PH model for each admission diagnosis group, and predicted the patient outcome based on matches only to other patients within the same admission diagnosis group.

The final approach that was chosen for predicting patient outcomes for the thesis, based on the results shown in the next section, was using the second approach and was evaulated against the testing set. The hyperparameters used for matching, dist and *k*, were chosen separately for each diagnosis model at $t_c = 2880$.

4.3 Results

4.3.1 Inclusion criteria, diagnosis groups, and parameters

At this point in the pipeline, after the initial data extraction, there were two additional match eligibility criteria that the patients needed to meet to be able to make a prediction, and be selected as a similar patient. First, for each patient, t_e needed to be greater than t_c, that is, the patient needed to still be in the ICU before the end of the data collection window, since there would be no prediction to be made as the outcome would already be known. Second, all patients needed to have some valid lp value in the first hour interval, the interval $(t_b, t_b + 60]$. Although the hourly mean predictor value could be imputed using the carry-forward method, the first hour could not be since there were no previous values that could be carried forward and used instead. Following from Figure 2-3, a similar analysis was done with these two criteria, shown in Figure 4-2. This figure was made using the values $t_b = 180$ (from Figure 2-3, where one of the steps excluded patients with $t_e < t_b$) and $t_c = 1440$. However, $t_c = 2880$ was also tested, and the number of patients that were included, both for the model using the entire patient dataset and the diagnosis-specific models, are shown in Table 4-1.

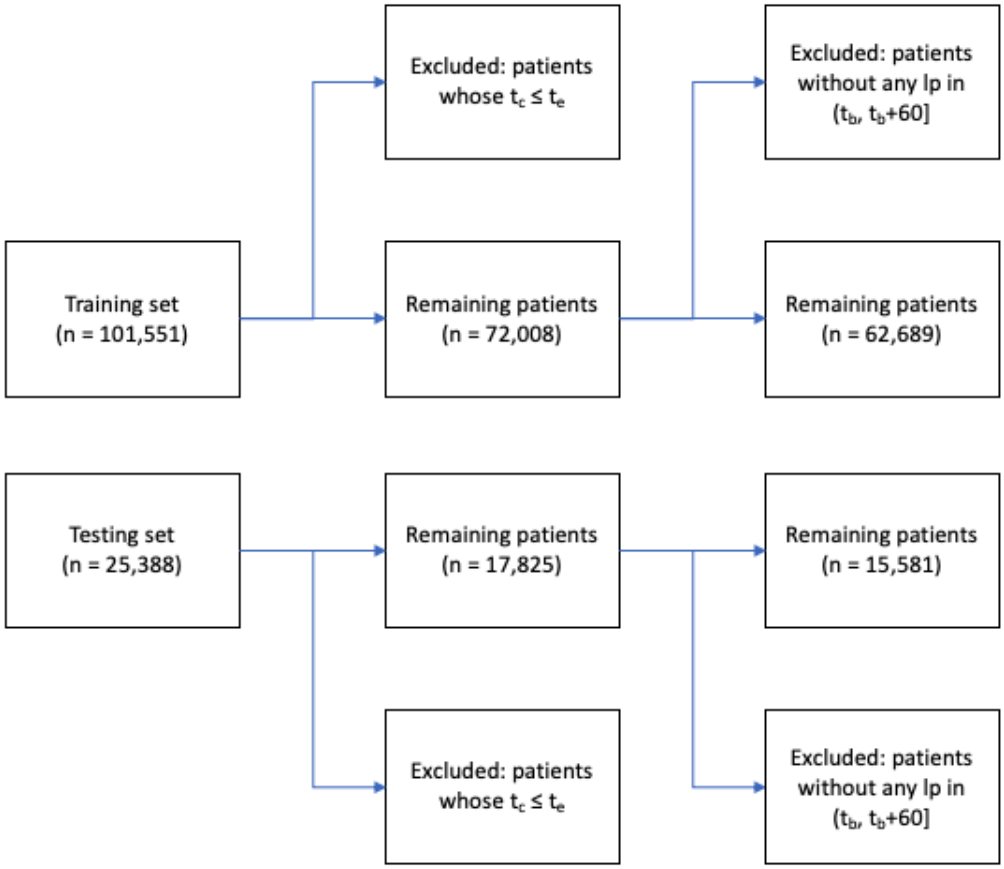

Figure 4-2. Filtering steps taken while calculating patient distances for $t_b = 180$ (3 hours) and $t_c = 1440$ (24 hours).

Table 4-1. Number of patients used for each Cox PH model and t_c. Note that the number of patients in diagnosis specific models do not add up to the number of patients in the model with all patients because three groups were excluded (gynecological, other medical disorders, and undefined)

t_c / Diagnosis model	1440 (1 day)		2880 (2 days)	
	Training	**Testing**	**Training**	**Testing**
All patients	62,689	15,581	35,811	8,924
Cardiovascular	18,698	4,573	9,842	2,504
Gastrointestinal	6,524	1,644	3,719	940
Hematological	395	113	209	55
Metabolic	5,185	1,305	2,285	592
Musculoskeletal	719	145	385	101
Neurological	8,709	2,184	4,966	1,218
Renal	1,385	367	804	181
Respiratory	9,004	2,294	5,879	1,468
Sepsis	8,576	2,084	5,672	1,370
Trauma	2,904	740	1,787	420

Two matching hyperparameters could be tuned for this project, which were the distance metric (dist) and *k* (the size of the cohort of similar patients used to make prediction from) . The time for which the prediction was being made, t_p, could also be varied. The burn-in interval, t_b, was set at 180 and not varied. A summary of the values that were tested are shown in Table 4-2. All possible combinations of these three parameters were tested in the admission diagnosis-specific and all patient methods.

Table 4-2. Values of tunable parameters (dist and *k*) and prediction time (t_p) tested in the model.

Distance metric (dist)	**Name**	**Formula**
	Euclidean (***L₂***)	$\sqrt{\sum_i (x_i - y_i)^2}$
	Manhattan (***L₁***)	$\sum_i \lvert x_i - y_i \rvert$
	Maximum (supremum norm)	$\max_i \lvert x_i - y_i \rvert$
k	**Number of patients used in similar cohort**	
	5	
	10	
	25	
	50	
	100	
	200	
	500	
$\mathbf{t_p}$	**Offset (minutes)**	**Hours**
	2160	36
	2880	48
	3660	60
	4320	72
	5040	84
	5760	96

4.3.2 Training results using complete patient set

The mean area under the receiver operating characteristic curve (AUROC) was calculated for each combination of hyperparameters over the five cross-validation folds, and the results are shown in Table 4-3. The mean AUROC is the mean across the five cross-validation folds, and only the best AUROC, and the corresponding dist and *k*, is shown for each t_p, and t_c. For t_c = 1440, t_p of 6480 and 7200 were not trained because the prediction time was too far in the future compared to the data collected, and for t_c = 2880, t_p of 2160 and 2880 could not be tested because t_p should be greater than t_c.

Table 4-3. The highest mean AUROC across the five cross-validation folds and the corresponding hyperparameters from training the model containing all patients across different t_c and t_p.

		t_c = 1440 (1 day)			**t_c = 2880 (2 days)**		
Diagnosis model	**t_p**	**dist**	***k***	**Mean AUROC**	**dist**	***k***	**Mean AUROC**
All patients	2160	maximum	500	0.665			
	2880	Euclidean	500	0.664			
	3600	maximum	500	0.657	Euclidean	500	0.667
	4320	maximum	500	0.651	Euclidean	500	0.669
	5040	maximum	500	0.645	Euclidean	500	0.667
	5760	maximum	500	0.641	Euclidean	500	0.665
	6480				Euclidean	500	0.661
	7200				Euclidean	500	0.659

From the table, we see that all models preferred larger similar cohort sizes, since the cohort size of k = 500 performed best across all sets of hyperparameters. Also, we see a difference in the best distance metric for t_c = 1440 (1 day) as opposed to t_c = 2880 (2 days), where the best distance metric for t_c =1440 was mostly maximum distance while the best distance metric for t_c = 2880 (2 days) was exclusively Euclidean. This could be because t_c = 2880 (2 days) allows enough time for patients to stabilize and for outlying intervals to make less of an impact, while for t_c = 1440 (1 day), a low hourly mean lp for any given interval could be more indicative of poor outcome.

4.3.3 Training results on specific APACHE admission diagnosis groups

Similar to section 4.3.3, the mean AUROC over the five cross-validation folds was calculated for each set of hyperparameter, and the results are shown in Table 4-4, where only the best mean AUROC, and the corresponding dist and k, is shown for each diagnosis model, t_p, and t_c.

Table 4-4. The highest mean AUROC across the five cross-validation folds and the corresponding hyperparameters from training the admission diagnosis-specific models across different t_c and t_p.

		t_c = 1440 (1 day)			**t_c = 2880 (2 days)**		
Diagnosis model	**t_p**	**dist**	***k***	**Mean AUROC**	**dist**	***k***	**Mean AUROC**
Cardiovascular	2160	Euclidean	500	0.635			
	2880	maximum	500	0.634			
	3600	maximum	500	0.629	Euclidean	500	0.652

	4320	maximum	500	0.626	Euclidean	500	0.644
	5040	maximum	500	0.620	Euclidean	500	0.643
	5760	maximum	500	0.615	Euclidean	500	0.636
	6480				Euclidean	500	0.631
	7200				Euclidean	500	0.631
Gastrointestinal	2160	maximum	500	0.653			
	2880	Manhattan	500	0.656			
	3600	Euclidean	500	0.652	Euclidean	500	0.671
	4320	Euclidean	500	0.651	Euclidean	100	0.686
	5040	Euclidean	500	0.650	Euclidean	100	0.687
	5760	maximum	500	0.644	Euclidean	200	0.689
	6480				Euclidean	500	0.686
	7200				Euclidean	500	0.684
Hematological	2160	Manhattan	200	0.602			
	2880	maximum	200	0.598			
	3600	Euclidean	25	0.589	Manhattan	10	0.626
	4320	Euclidean	100	0.608	Manhattan	10	0.630
	5040	Euclidean	100	0.598	Manhattan	10	0.652
	5760	Euclidean	100	0.575	maximum	100	0.670
	6480				Manhattan	25	0.693
	7200				maximum	100	0.689
Metabolic	2160	maximum	500	0.625			
	2880	maximum	500	0.641			
	3600	maximum	500	0.626	maximum	500	0.621
	4320	maximum	500	0.622	Euclidean	500	0.645
	5040	maximum	500	0.616	maximum	100	0.648
	5760	maximum	500	0.609	maximum	100	0.642
	6480				maximum	200	0.648
	7200				maximum	100	0.641
Musculoskeletal	2160	maximum	100	0.622			

	2880	Manhattan	200	0.624			
	3600	Manhattan	500	0.621	maximum	25	0.671
	4320	Manhattan	500	0.615	maximum	100	0.631
	5040	Manhattan	200	0.620	Manhattan	100	0.663
	5760	Manhattan	200	0.612	Euclidean	100	0.638
	6480				Euclidean	100	0.650
	7200				Euclidean	100	0.644
Neurological	2160	maximum	500	0.654			
	2880	Euclidean	500	0.645			
	3600	maximum	500	0.640	Euclidean	500	0.658
	4320	maximum	500	0.627	Euclidean	200	0.646
	5040	maximum	500	0.626	Euclidean	500	0.647
	5760	Euclidean	500	0.626	Euclidean	500	0.645
	6480				Euclidean	500	0.643
	7200				Euclidean	500	0.639
Renal	2160	Euclidean	500	0.642			
	2880	Euclidean	500	0.618			
	3600	Manhattan	500	0.633	Euclidean	500	0.682
	4320	Manhattan	200	0.646	maximum	200	0.663
	5040	Manhattan	200	0.646	Manhattan	500	0.674
	5760	Manhattan	200	0.648	Manhattan	500	0.677
	6480				Manhattan	500	0.668
	7200				Manhattan	500	0.664
Respiratory	2160	maximum	500	0.655			
	2880	maximum	500	0.660			
	3600	Euclidean	500	0.645	Manhattan	500	0.651
	4320	maximum	500	0.640	Euclidean	200	0.662
	5040	maximum	500	0.631	Euclidean	200	0.656
	5760	Euclidean	500	0.631	Euclidean	500	0.652
	6480				Euclidean	200	0.652

	7200				Euclidean	200	0.646
Sepsis	2160	Euclidean	500	0.705			
	2880	Euclidean	500	0.703			
	3600	Euclidean	500	0.698	Euclidean	500	0.711
	4320	Euclidean	500	0.699	Manhattan	200	0.722
	5040	Manhattan	500	0.692	Euclidean	500	0.718
	5760	Euclidean	500	0.688	Euclidean	500	0.716
	6480				Euclidean	500	0.712
	7200				Euclidean	500	0.709
Trauma	2160	Euclidean	500	0.649			
	2880	Euclidean	500	0.667			
	3600	Euclidean	500	0.659	Euclidean	200	0.672
	4320	Euclidean	500	0.655	maximum	100	0.685
	5040	Euclidean	500	0.654	maximum	100	0.677
	5760	Euclidean	500	0.653	maximum	100	0.678
	6480				maximum	100	0.672
	7200				maximum	200	0.662

There were significant differences in the performance of different admission diagnosis groups, and the selected best set of hyperparameters. Even though this difference could largely be attributed to the groups having different number of patients, differences could still be seen among groups that had similar sizes. In particular, the model using only sepsis patients performed better than all other models, including the model using all patients. The models for groups gastrointestinal, renal, and trauma patients performed better than the model using all patients for $t_c = 2880$, but not $t_c = 1440$.

In general, we observe that predictions using $t_c = 2880$ performed better than models using $t_c = 1440$. This was especially true for the admission diagnosis-specific models, where four models trained on different admission diagnosis groups performed better than the model trained on all patients at $t_c = 2880$, as opposed to only one (sepsis) at $t_c = 1440$. Therefore, we trained the final model using $t_c = 2880$ on admission diagnosis-specific models.

4.3.4 Testing results using specific admission diagnosis groups

Based on these results, the final model that was chosen for the thesis matched separately for each admission diagnosis group, using $t_c = 2880$. The parameters dist and *k* that performed the best at each t_p was used, and these parameters are shown in Figure 4-3. The results on using the model on the testing set is shown in Table 4-5. Similar to the training results, we see the sepsis group performing the best out of all the admission diagnosis groups. The combined results, which was calculated by collecting predictions of patients using the ten admission diagnosis-specific models, were similar to those using all patients together without separating the pipeline for different admission diagnosis groups.

Table 4-5. Testing with selected the hyperparameters, which split each admission diagnosis group into a separate group and used $t_c = 2880$ (2 days).

Diagnosis group	**t_p**	**AUROC (test set outcome)**
Combined	3600	0.6608
	4320	0.6696
	5040	0.6653
	5760	0.6677
	6480	0.6605
	7200	0.6569

Cardiovascular	3600	0.6489
	4320	0.6525
	5040	0.6464
	5760	0.6451
	6480	0.6303
	7200	0.6318
Gastrointestinal	3600	0.7065
	4320	0.6657
	5040	0.6584
	5760	0.6511
	6480	0.6603
	7200	0.6561
Hematological	3600	0.676
	4320	0.5599
	5040	0.6134
	5760	0.5943
	6480	0.6186
	7200	0.5289
Metabolic	3600	0.6206
	4320	0.6639
	5040	0.6589
	5760	0.6537
	6480	0.6329
	7200	0.6349
Musculoskeletal	3600	0.5884
	4320	0.7055
	5040	0.6355
	5760	0.6396
	6480	0.6059
	7200	0.6134

Neurological	3600	0.6183
	4320	0.6300
	5040	0.6325
	5760	0.6317
	6480	0.6349
	7200	0.6361
Renal	3600	0.6641
	4320	0.6315
	5040	0.6396
	5760	0.6456
	6480	0.6275
	7200	0.6176
Respiratory	3600	0.6566
	4320	0.6819
	5040	0.6769
	5760	0.6847
	6480	0.6720
	7200	0.6625
Sepsis	3600	0.708
	4320	0.7152
	5040	0.7001
	5760	0.7107
	6480	0.7023
	7200	0.6693
Trauma	3600	0.7259
	4320	0.7091
	5040	0.6926
	5760	0.6854
	6480	0.6870
	7200	0.6787

4.3.5 Example test outcomes

As an example of the output, we can observe two different index patients with a sepsis admission diagnosis group. In this hypothetical example, we have observed the patients for the first two days ($t_c = 2880$), and are considering what the next steps might be. A family member asks if they were to stay in the ICU for another two days ($t_p = 5760$), what are the likely outcomes. Using the outlined approach, for the first index patient whose patientunitstayid was 141515, the predictor value (similar to a probability estimate but may not be calibrated correctly using the current calculation method of unweighted mean outcome of the similar cohort) is calculated as 0.19, where the low score means that this patient is not likely to have a healthy discharge. The trajectory of the lp for the index patient, 141515, as well as the three most similar patients, 1566888, 2798447, and 3008354, are shown in Figure 4-3. Out of the three patients, only 3008354 had a healthy discharge. We see that the trajectories look mostly similar to each other, especially in the middle of the collection window. The outcome of the patient is later observed to be no healthy discharge by time $t_p = 5770$.

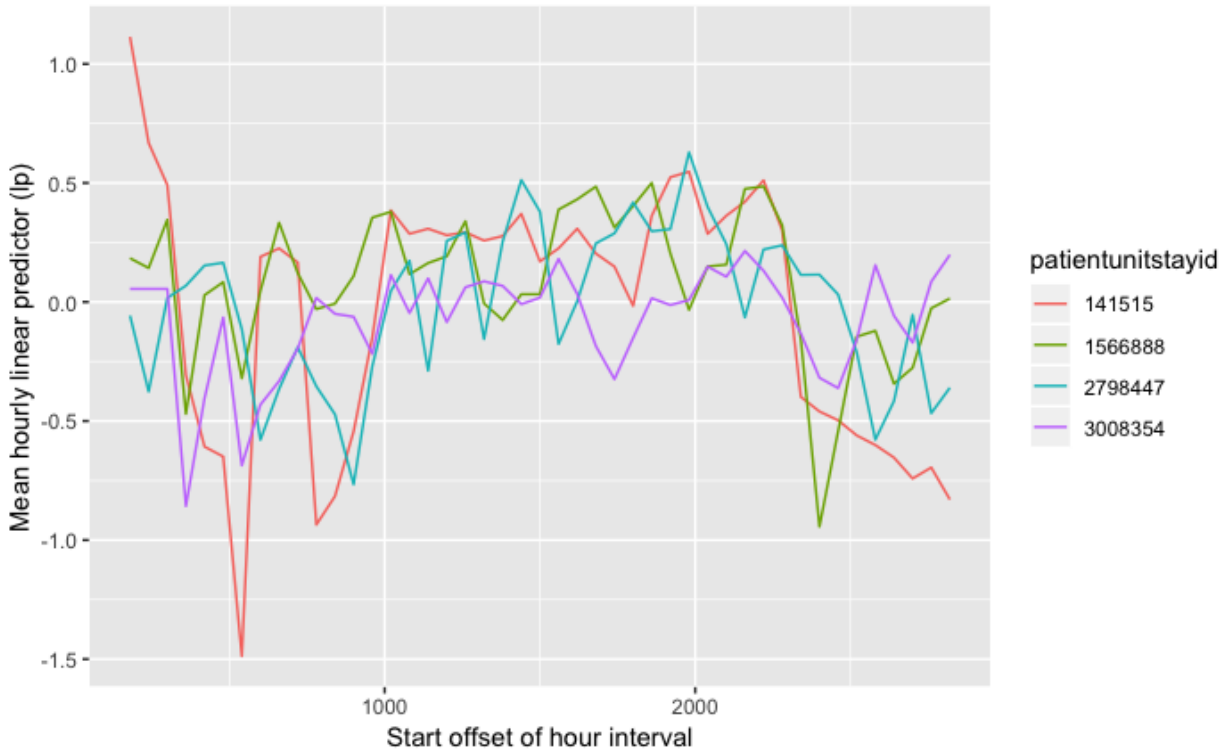

Figure 4-3. Time series of the mean hourly lp for index patient 141515, and three most similar patients.

The predictor value of the second index patient, whose patientunitstayid was 147633, is calculated to be 0.346, where the relatively higher score means that this patient is likely to have a healthy discharge. The trajectory of the lp for the index patient, 147633, as well as the three most similar patients, 2859270, 2854883, and 404987, are shown in Figure 4-4. Out of the three patients, only 404987 did not have a healthy discharge. Here again, we see mostly similar intervals, but one of the patients, 404987, is noticeable because their time series of the lp is mostly flat with minimal changes. However, they may have been matched because, on average, the trajectory of 404987 is similar to 147633 without noise. The outcome of the patient is later observed to be healthy discharge by time $t_p = 5770$.

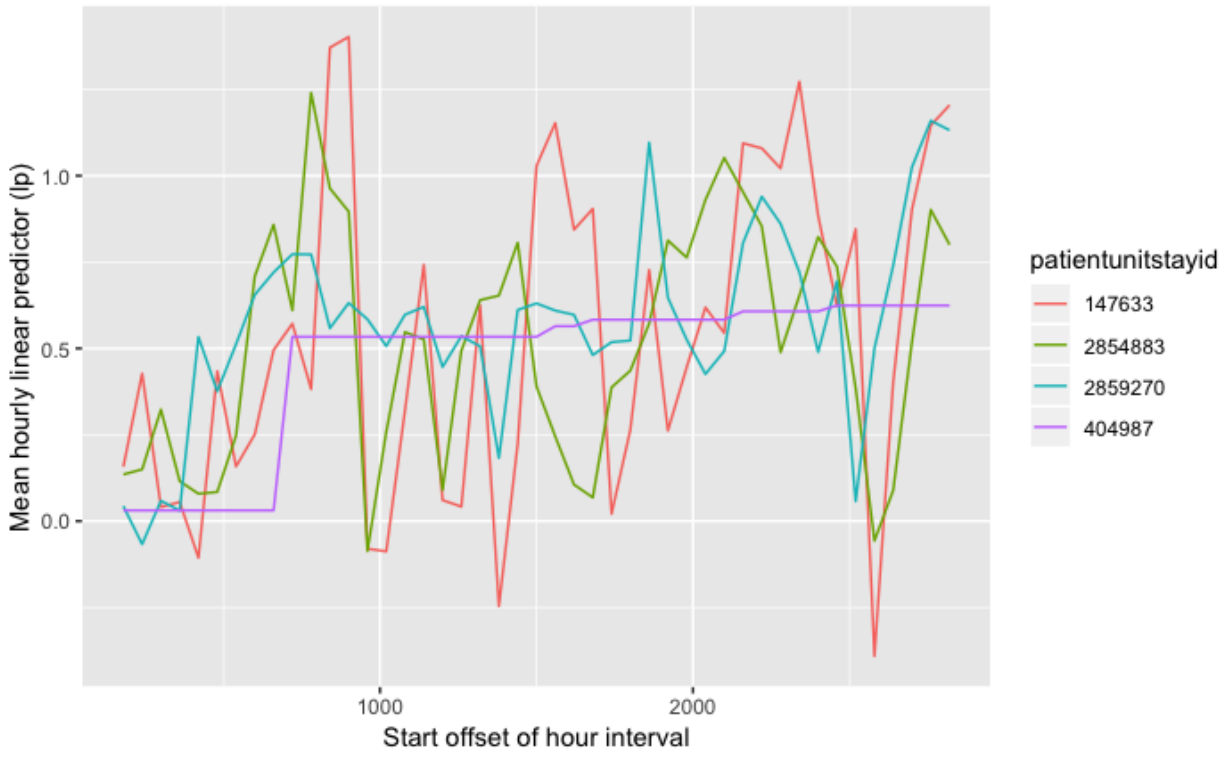

Figure 4-4. Time series of the mean hourly lp for index patient 147633, and three most similar patients.

4.4 Conclusion

Even though the overall number of patients that was used in the model seemed sufficient, there were several admission diagnosis groups for which there were very few patients, such as hematological and musculoskeletal. Because of these small sample sizes for training, these groups performed worse than the other diagnosis groups. Also, even for the larger groups, the performance varied considerably between the different groups. This may have been because the set of covariates that were chosen to model the patient and used to build the Cox PH model did not change from one admission diagnosis group to another. However, these covariates may have

only been appropriate for some groups, such as sepsis, while not appropriate for others, such as metabolic.

THIS PAGE IS INTENTIONALLY LEFT BLANK

Chapter 5

Conclusion and Future Works

Currently, there is a need for a tool that could assist a patient's family and care providers in making end-of-life decisions that provides a prediction that is easy to interpret and understand for both of the parties involved. Although the idea of identifying a ' digital twin' is often viewed as the panacea for clinical decision support, this thesis exemplifies that barriers exist in all aspects of the pipeline which can make this goal formidable. This work describes a method, using the eICU-CRD v2.0 database, that maps a multivariate time series of a patient stay data to a univariate time series using a Cox Proportional Hazards model. The method predicts the outcome of an index patient based on the outcomes of a similar cohort of patients based on the univariate time series. Using a short burn-in period, $t_b = 180$ (3 hours), and a collection window $t_c = 2880$ (2 days), the prediction tool was able to achieve up to AUROC of 0.7152 for a specific admission diagnosis group, sepsis, and 0.6696 across the ten admission diagnosis groups at $t_p = 4320$ (3 days) in the test set. Although the discrimination of the final model was not very impressive, the model was still able to clearly show the reason behind the prediction and the cohort of similar patients that were matched, which the physician can then use to either accept or reject matches to improve on the model.

Using this approach, we were able to train on a large database that contains over 200 hospitals, using over 100,000 patients from these hospitals. We were also able to use many

different types of covariates to train the model, including both time-varying and time-constant features, and different types of features (categorical, numerical, binary). However, the scope of the features that were used were limited, both by frequent missing data in some features as well as computational memory while training. Also, we were only able to train on one type of outcome, healthy discharge, although the approach could be used to test other outcomes, such as in-hospital mortality.

There are many future directions in which this model could be improved, one of which is during data extraction. As mentioned in chapter 2, one of the concerns that we faced was the size of the data frame, which was both too tall from having too many observations and too wide from having too many different types of covariates. In the future, lowering the resolution of the data, by lowering the resolution of the collected values or binning the time points when the data was recorded and lowering the time resolution, could allow the model to be trained with more data and thus obtain better results. For some of the variables that could not be used because of missing data, such as laboratory results, binning the measurements into normal and abnormal categories, along with another category for patients that are missing these lab results, similar to the approach taken for GCS, could be used to add these other covariates into the model. Also, varying some of the parameters used during extraction could also be helpful. For example, t_b was set to be 180 (3 hours), but perhaps increasing or decreasing this number could allow for more patients to be included and matched. Also, the time windows used for the rolling statistics (median, minimum, and maximum) were set at $t_{r,min} = t_{r,max} = t_{r,median} = 60$ (1 hour). However, we could vary these windows so that they are more appropriate. For example, we may want to keep $t_{r,median}$ small (at 60) since this was useful for filtering noise, but perhaps changing $t_{r,min}$ and $t_{r,max}$

to a larger value, such as 1440 (1 day), could help us capture the worst patient state during the stay and may be more useful for predicting the outcome of this patient.

Another area of future work was in the Cox PH model. Some of the covariates that were used to fit the model need to be investigated because they did not seem appropriate from the output of the model. For example, temperature was not statistically significant in any of the models. However, this may be due to its inclusion in a linear factor form, and a patient's discharge from the hospital does not seem like it should depend linearly on the temperature. Instead, the highest risk of healthy discharge should probably be around 37°C, with temperatures either above or below this temperature lowering the risk of discharge. Also, the confidence intervals of some the covariates was calculated to be from 0 to infinity, which could signal some problem with the model, such as collinearity. Currently, we used the same set of covariates for all the models, which could contribute to this problem, but using a different set of covariates that are more relevant to different admission diagnosis groups could be tested.

For the patient matching, only a limited number of parameters were tested. The distance metrics (Euclidean, Manhattan, and maximum) that were used were the basic distance metrics provided by R. However, a patient trajectory in the ICU could be thought of as having two components, the mean predictor throughout their stay and the trajectory of the differences from the mean for each timepoint of their stay. The distance metrics used do not really capture both of these components well, and using a more sophisticated metric could help significantly with the match. Similarly, the patients in the similar cohort had an equal weight in calculating the prediction for the index patient, whether the similar patient was the most similar or the 500^{th} most similar to the index patient. Instead of using a basic k-nearest neighbor prediction in this way, we could use a variation of a k-nearest neighbor, for example weighing the outcomes of the

similar patients differently, to achieve better results. Also, only matches of up to 500 were used. However, many of models had the best AUROC at a cohort size of 500. Testing with different cohort sizes, for example by increasing the maximum number of matches, could also be useful in the future.

Overall, this type of method may be useful, but presents many challenges requiring both methodological and engineering research. However, even though the model did not perform very well, the similar cohort could be displayed clearly, from which the physician could either choose to accept or reject the matches and potentially make a more informed recommendation based on the larger sample size of the patients. By addressing the challenges, this type of method would become more useful by presenting to the physician more closely resembling cohort of similar patients.

Appendix A

The table that contains the coefficient of each covariate in the final model described in chapter 3 is shown in Table A-1. The first column is the name of the covariate, second is the value of the coefficient of the covariates (β in the Cox PH equation), third is the standard error, fourth is the Z value (the coefficient divided by the standard error of the coefficient), and fifth is the p-value of the coefficient. For categorical covariates (i.e. ethnicity, unitadmitsource, and gcs_group), the coefficients were calculated compared to the first category alphabetically, which were: African American for ethnicity, Emergency Room for unitadmitsource, and ALL_HIGH for gcs_group.

Table A-1. Table showing the statistically significant covariates ($p < 0.001$) of each Cox Proportional Hazards model fit on the specific diagnosis groups.

APACHE admission diagnosis group: Cardiovascular				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
age	-4.314E-03	7.359E-04	-5.862	4.58E-09
ethnicity: Caucasian	1.446E-01	3.249E-02	4.449	8.64E-06
ethnicity: Hispanic	4.475E-01	5.453E-02	8.207	2.27E-16

unitadmitsource: Operating Room	-1.477E-01	2.666E-02	-5.54	3.02E-08
cardiovascular	-2.324E-01	5.572E-02	-4.17	3.04E-05
endocrine	1.639E-01	2.675E-02	6.125	9.07E-10
hematology	-1.572E-01	4.627E-02	-3.397	0.000682
neurologic	2.311E-01	3.253E-02	7.104	1.21E-12
pulmonary	-3.178E-01	2.849E-02	-11.155	< 2e-16
renal	-1.978E-01	2.977E-02	-6.644	3.05E-11
sepsis	-5.206E-01	1.017E-01	-5.118	3.09E-07
gcs_group: ALL_LOW	-3.724E+00	3.54E-01	-10.518	< 2e-16
gcs_group: INTERMED	-2.383E+00	1.698E-01	-14.035	< 2e-16
gcs_group: MISSING	-1.597E-01	2.08E-02	-7.678	1.61E-14
gcs_group: SOME_HIGH	-9.658E-01	5.119E-02	-18.865	< 2e-16
bp_med: first spline term	-5.266E+02	1.266E+02	-4.161	3.17E-05
bp_med: second spline term	-1.104E+03	2.58E+02	-4.278	1.88E-05
bp_max	-3.733E-02	2.143E-03	-17.421	< 2e-16
bp_min	3.937E-02	2.111E-03	18.649	< 2e-16
heartrate	1.544E-02	2.755E-03	5.604	2.10E-08
heartrate_med: first spline term	-6.782E+01	7.343E+00	-9.236	< 2e-16

heartrate_med: second spline term	-1.453E+02	1.531E+01	-9.492	< 2e-16
heartrate_max	-2.147E-02	2.095E-03	-10.249	< 2e-16
respiration	1.877E-02	3.884E-03	4.834	1.34E-06
respiration_med: first spline term	-1.736E+01	3.248E+00	-5.344	9.09E-08
respiration_med: second spline term	-3.724E+01	6.682E+00	-5.573	2.51E-08
respiration_max	-3.212E-02	2.697E-03	-11.908	< 2e-16
respiration_min	5.106E-02	3.351E-03	15.237	< 2e-16
sao2	-1.719E-02	1.764E-03	-9.746	< 2e-16
APACHE admission diagnosis group: Gastrointestinal				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
age	-6.689E-03	1.248E-03	-5.362	8.24E-08
male	-1.225E-01	3.633E-02	-3.371	0.000748
ethnicity: Hispanic	4.548E-01	9.259E-02	4.911	9.04E-07
unitadmitsource: Other	-2.557E-01	5.072E-02	-5.041	4.62E-07
pulmonary	-1.922E-01	5.586E-02	-3.441	0.000581
sepsis	-3.742E-01	1.037E-01	-3.607	0.000310
gcs_group: ALL_LOW	-1.993E+00	5.786E-01	-3.445	0.000571
gcs_group: INTERMED	-2.173E+00	2.379E-01	-9.134	< 2e-16
gcs_group: MISSING	-1.559E-01	3.991E-02	-3.906	9.39E-05

gcs_group: SOME_HIGH	-1.017E+00	7.334E-02	-13.868	< 2e-16
bp_max	-4.197E-02	4.247E-03	-9.882	< 2e-16
bp_min	4.137E-02	4.04E-03	10.24	< 2e-16
heartrate	3.672E-02	5.161E-03	7.115	1.12E-12
heartrate_med: first spline term	-1.157E+01	2.978E+00	-3.885	0.000102
heartrate_med: second spline term	-2.945E+01	5.027E+00	-5.858	4.69E-09
heartrate_max	-4.261E-02	4.187E-03	-10.177	< 2e-16
respiration_max	-2.992E-02	5.193E-03	-5.762	8.29E-09
respiration_min	4.116E-02	6.194E-03	6.646	3.01E-11
sao2	-1.589E-02	4.031E-03	-3.943	8.04E-05
APACHE admission diagnosis group: Hematological				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
gcs_group: SOME_HIGH	-1.099E+00	3.263E-01	-3.369	0.000755
bp_max	-5.808E-02	1.762E-02	-3.297	0.000977
APACHE admission diagnosis group: Metabolic				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
age	-7.752E-03	1.07E-03	-7.247	4.25E-13
unitadmitsource: Floor	-2.582E-01	7.347E-02	-3.514	0.000442
unitadmitsource: Other	-4.485E-01	6.237E-02	-7.191	6.45E-13

pulmonary	-1.998E-01	5.288E-02	-3.778	0.000158
toxicology	2.614E-01	4.641E-02	5.632	1.78E-08
gcs_group: ALL_LOW	-2.584E+00	7.079E-01	-3.65	0.000262
gcs_group: INTERMED	-1.556E+00	1.356E-01	-11.476	< 2e-16
gcs_group: SOME_HIGH	-7.936E-01	6.030E-02	-13.16	< 2e-16
bp	1.414E-02	4.107E-03	3.444	0.000573
bp_max	-4.679E-02	3.937E-03	-11.885	< 2e-16
bp_min	4.822E-02	3.862E-03	12.485	< 2e-16
heartrate	1.781E-02	4.12E-03	4.323	1.54E-05
heartrate_med: first spline term	-1.575E+01	4.383E+00	-3.594	0.000326
heartrate_med: second spline term	-5.134E+01	9.081E+00	-5.653	1.58E-08
heartrate_max	-3.29E-02	3.287E-03	-10.01	< 2e-16
respiration	2.425E-02	6.442E-03	3.764	0.000167
respiration_max	-2.668E-02	4.507E-03	-5.92	3.21E-09
respiration_min	3.752E-02	5.636E-03	6.657	2.79E-11
sao2	-1.477E-02	3.327E-03	-4.44	8.98E-06
APACHE admission diagnosis group: Musculoskeletal				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
male	-3.963E-01	9.974E-02	-3.973	7.08E-05

gcs_group: INTERMED	-1.602E+00	4.593E-01	-3.488	0.000486
gcs_group: SOME_HIGH	-8.182E-01	1.816E-01	-4.506	6.60E-06
sao2	-3.369E-02	9.287E-03	-3.628	0.000286
APACHE admission diagnosis group: Neurological				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
age	-6.63E-03	9.697E-04	-6.838	8.05E-12
ethnicity: Other/Unknown	3.462E-01	7.34E-02	4.716	2.40E-06
unitadmitsource: Operating Room	2.466E-01	4.913E-02	5.019	5.19E-07
pulmonary	-4.36E-01	5.149E-02	-8.468	< 2e-16
gcs_group: ALL_LOW	-3.299E+00	5.781E-01	-5.707	1.15E-08
gcs_group: INTERMED	-2.553E+00	1.548E-01	-16.494	< 2e-16
gcs_group: MISSING	-1.331E-01	3.335E-02	-3.991	6.58E-05
gcs_group: SOME_HIGH	-8.822E-01	4.687E-02	-18.822	< 2e-16
bp_max	-4.502E-02	3.387E-03	-13.292	< 2e-16
bp_min	3.767E-02	3.35E-03	11.245	< 2e-16
heartrate	2.593E-02	4.164E-03	6.228	4.73E-10
heartrate_med: second spline term	-6.728E+00	1.422E+00	-4.732	2.23E-06

heartrate_max	-4.491E-02	3.405E-03	-13.187	< 2e-16
heartrate_min	1.698E-02	3.782E-03	4.49	7.11E-06
respiration_med: first spline term	-9.079E+00	2.613E+00	-3.474	0.000512
respiration_med: second spline term	-2.031E+01	5.395E+00	-3.765	0.000167
respiration_max	-2.06E-02	4.246E-03	-4.852	1.22E-06
respiration_min	5.384E-02	5.888E-03	9.144	< 2e-16
sao2	-2.451E-02	3.704E-03	-6.618	3.64E-11
APACHE admission diagnosis group: Renal				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
surgery	4.85E-01	1.456E-01	3.332	0.000863
gcs_group: INTERMED	-2.733E+00	7.107E-01	-3.846	0.000120
gcs_group: SOME_HIGH	-7.285E-01	1.382E-01	-5.271	1.35E-07
bp_med: first spline term	-3.778E+01	9.826E+00	-3.845	0.000121
bp_med: second spline term	-6.783E+01	1.977E+01	-3.432	0.000600
bp_max	-2.575E-02	7.408E-03	-3.476	0.000510
bp_min	5.55E-02	8.647E-03	6.419	1.38E-10
heartrate_max	-5.344E-02	9.578E-03	-5.58	2.40E-08
respiration_min	5.7E-02	1.293E-02	4.409	1.04E-05
sao2	-2.981E-02	4.572E-03	-6.519	7.07E-11

APACHE admission diagnosis group: Respiratory				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
age	-8.228E-03	1.189E-03	-6.918	4.60E-12
ethnicity: Asian	5.606E-01	1.43E-01	3.919	8.89E-05
ethnicity: Hispanic	4.209E-01	8.609E-02	4.888	1.02E-06
unitadmitsource: Floor	-3.974E-01	5.047E-02	-7.874	3.45E-15
unitadmitsource: Other	-4.236E-01	5.116E-02	-8.279	< 2e-16
pneumonia	-2.227E-01	4.241E-02	-5.252	1.51E-07
gcs_group: ALL_LOW	-3.371E+00	7.078E-01	-4.762	1.92E-06
gcs_group: INTERMED	-2.567E+00	1.909E-01	-13.452	< 2e-16
gcs_group: MISSING	-3.288E-01	3.829E-02	-8.587	< 2e-16
gcs_group: SOME_HIGH	-1.221E+00	6.599E-02	-18.505	< 2e-16
bp_max	-3.236E-02	3.686E-03	-8.779	< 2e-16
bp_min	4.638E-02	3.896E-03	11.903	< 2e-16
heartrate	4.292E-02	4.639E-03	9.252	< 2e-16
heartrate_med: first spline term	-1.231E+02	1.856E+01	-6.632	3.32E-11
heartrate_med: second spline term	-2.645E+02	1.39E+02	-6.747	1.51E-11
heartrate_max	-4.535E-02	4.024E-03	-11.27	< 2e-16

respiration_max	-2.19E-02	4.503E-03	-4.862	1.16E-06
respiration_min	3.651E-02	5.737E-03	6.364	1.97E-10
sao2	-1.97E-02	2.997E-03	-6.572	4.98E-11
APACHE admission diagnosis group: Sepsis				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
ethnicity: Hispanic	3.547E-01	1.054E-01	3.365	0.000765
unitadmitsource: Other	-5.034E-01	6.818E-02	-7.384	1.54E-13
endocrine	2.125E-01	4.853E-02	4.378	1.20E-05
`infectious diseases`	1.82E-01	4.478E-02	4.064	4.83E-05
pulmonary	-4.502E-01	5.528E-02	-8.144	3.82E-16
sepsis	4.354E-01	6.858E-02	6.349	2.17E-10
gcs_group: ALL_LOW	-2.667E+00	5.787E-01	-4.608	4.06E-06
gcs_group: INTERMED	-2.053E+00	1.636E-01	-12.554	< 2e-16
gcs_group: MISSING	-4.417E-01	4.578E-02	-9.649	< 2e-16
gcs_group: SOME_HIGH	-7.089E-01	6.118E-02	-11.588	< 2e-16
bp_med: first spline term	-1.552E+02	2.847E+01	-5.45	5.03E-08
bp_med: second spline term	-3.321E+02	6E+01	-5.535	3.12E-08
bp_max	-3.532E-02	4.258E-03	-8.295	< 2e-16
bp_min	6.096E-02	4.584E-03	13.299	< 2e-16

heartrate_med: first spline term	-3.346E+01	7.943E+00	-4.212	2.53E-05
heartrate_med: second spline term	-8.106E+01	1.599E+01	-5.07	3.99E-07
heartrate_max	-3.011E-02	4.479E-03	-6.722	1.79E-11
respiration_min	4.046E-02	6.87E-03	5.889	3.89E-09
sao2	-2.004E-02	3.757E-03	-5.334	9.58E-08
APACHE admission diagnosis group: Trauma				
Covariate	**coef**	**se(coef)**	**z**	**Pr(>|z|)**
age	-5.331E-03	1.426E-03	-3.738	0.000185
male	-1.914E-01	5.744E-02	-3.333	0.000860
pulmonary	-2.857E-01	8.556E-02	-3.34	0.000839
toxicology	5.77E-01	1.363E-01	4.233	2.31E-05
gcs_group: ALL_LOW	-3.64E+00	1.002E+00	-3.634	0.000279
gcs_group: INTERMED	-2.194E+00	2.398E-01	-9.153	< 2e-16
gcs_group: MISSING	-2.118E-01	5.794E-02	-3.655	0.000257
gcs_group: SOME_HIGH	-1.17E+00	9.990E-02	-11.709	< 2e-16
bp_max	-4.446E-02	6.545E-03	-6.793	1.10E-11
bp_min	3.812E-02	6.152E-03	6.196	5.79E-10
heartrate	2.551E-02	7.163E-03	3.561	0.000369
heartrate_max	-4.503E-02	5.868E-03	-7.674	1.67E-14

respiration_med: second spline term	-3.883E+01	1.166E+01	-3.331	0.000864
respiration_max	-2.866E-02	8.204E-03	-3.493	0.000477
respiration_min	3.895E-02	1.045E-02	3.727	0.000194
sao2	-3.133E-02	5.396E-03	-5.806	6.40E-09

THIS PAGE IS INTENTIONALLY LEFT BLANK

Bibliography

[1] *ANZICS-APD-Data-Dictionary.Pdf*. https://www.anzics.com.au/wp-content/uploads/2018/08/ANZICS-APD-Data-Dictionary.pdf. Accessed 27 May 2019.

[2] Hoffrage, Ulrich, et al. "Natural Frequencies Improve Bayesian Reasoning in Simple and Complex Inference Tasks." *Frontiers in Psychology*, vol. 6, 2015. *Frontiers*, doi:10.3389/fpsyg.2015.01473.

[3] Johnson, Alistair E. W., et al. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data*, vol. 3, May 2016, p. 160035. *www.nature.com*, doi:10.1038/sdata.2016.35.

[4] Kaplan, E. L., and Paul Meier. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association*, vol. 53, no. 282, June 1958, pp. 457–81. *Taylor and Francis+NEJM*, doi:10.1080/01621459.1958.10501452.

[5] Kleinbaum, David G., and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. 3rd ed, Springer, 2012.

[6] Pollard, Tom J., et al. "The EICU Collaborative Research Database, a Freely Available Multi-Center Database for Critical Care Research." *Scientific Data*, vol. 5, Sept. 2018, p. 180178. *www.nature.com*, doi:10.1038/sdata.2018.178.

[7] Rajkomar, Alvin, et al. "Scalable and Accurate Deep Learning for Electronic Health Records." *Npj Digital Medicine*, vol. 1, no. 1, Dec. 2018. *arXiv.org*, doi:10.1038/s41746-018-0029-1.

[8] Sharafoddini, Anis, et al. "Patient Similarity in Prediction Models Based on Health Data: A Scoping Review." *JMIR Medical Informatics*, vol. 5, no. 1, 2017, p. e7. *medinform.jmir.org*, doi:10.2196/medinform.6730.

[9] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A Survey of Collaborative Filtering Techniques." *Advances in Artificial Intelligence*, 2009, doi:10.1155/2009/421425.

[10] Teasdale, Graham, and Bryan Jennett. "ASSESSMENT OF COMA AND IMPAIRED CONSCIOUSNESS: A Practical Scale." *The Lancet*, vol. 304, no. 7872, July 1974, pp. 81–84. *ScienceDirect*, doi:10.1016/S0140-6736(74)91639-0.

[11] "Tmerge Function | R Documentation." *RDocumentation*, https://www.rdocumentation.org/packages/survival/versions/2.43-3/topics/tmerge. Accessed 27 May 2019.

[12] Topol, Eric J. “High-Performance Medicine: The Convergence of Human and Artificial Intelligence.” *Nature Medicine*, vol. 25, no. 1, Jan. 2019, p. 44. *www.nature.com*, doi:10.1038/s41591-018-0300-7.

[13] Zimmerman, Jack E., et al. “Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital Mortality Assessment for Today’s Critically Ill Patients*.” *Critical Care Medicine*, vol. 34, no. 5, May 2006, p. 1297. *journals.lww.com*, doi:10.1097/01.CCM.0000215112.84523.F0.