

**Cohort Selection and Sampling Techniques to
Balance Time-Series Retrospective Studies**

by

Brian Bell Jr

S.B., Massachusetts Institute of Technology (2013)

Submitted to the Department of Electrical Engineering
and Computer Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2017

© Brian Bell Jr, MMXVII. All rights reserved.

The author hereby grants to MIT permission to reproduce and to
distribute publicly paper and electronic copies of this thesis document
in whole or in part in any medium now known or hereafter created.

Author

Department of Electrical Engineering
and Computer Science
January 29, 2017

Certified by

Una-May O'Reilly
Principal Research Scientist
Thesis Supervisor

Accepted by

Christopher Terman
Chairman, Masters of Engineering Thesis Committee

Cohort Selection and Sampling Techniques to Balance Time-Series Retrospective Studies

by Brian Bell Jr

Submitted to the Department of Electrical Engineering
and Computer Science on January 29, 2017, in partial fulfillment of the
requirements for the degree of Master of Engineering in Electrical Engineering and
Computer Science

Abstract

Comparing irregular and event-driven time series data points is beyond the capabilities of most statistical techniques. This limits the potential to run insightful retrospective studies on many cross-sectional time-series datasets. In order to unlock the value of these datasets, we need techniques to standardize observations with irregular events enough to compare them to each other, and ways to select and sample them so as to produce class balances for each strata at modeling time that lend themselves to statistically sound analysis.

In this study, we have developed two selection techniques and three sampling techniques for a characteristic cross-sectional time-series dataset. We found that using a Fluid-Balance Similarity-Based Dynamic Time Warp selection procedure with nearest neighbor parameter $k=1$ and using a Gamma distribution for sampling days produced consistently better class balance than all other methods when bootstrapped over 100 independent runs. We have written, documented and published open source MATLAB code for each selection and sampling technique, along with our bootstrap test.

To evaluate our results, we have developed the Class Imbalance Penalty, a new metric that gives the lowest scores to the selection and sampling runs that produce most comparable counts of treatment and non-treatment observations for all strata.

We validated our methods in the context of a study of diuretics treatment effects in ICU patients with Sepsis, drawn from the MIMIC II database. Starting from a group of 3,503 unique ICU stays from 2,341 study patients, with a Diuretics-treatment cohort of 349 unique ICU stays from 332 patients, we tested each selection and sampling technique, observing the trends across our different methods. We observed that sampling day was the stronger predictor of good class balance compared with selection technique, that the strongest similarity level ($k=1$) with the shortest history we considered produced the best results, and using a Gamma distribution for timepoint sampling most closely matched the distribution of actual administration days. Ultimately, we found strong evidence that our study lacked an important covariate, physician-id, to more fully account for seemingly unpredictable assignments to Diuretics-treatment in our dataset.

Thesis Supervisor: Una-May O'Reilly
Title: Principal Research Scientist

Acknowledgments

This thesis would not have been possible without my advisor, Una-May O'Reilly. Una-May is a driven, passionate master at her craft. Without her help I would not have been able to finish.

It also would not have been possible without the help and guidance of Kalyan Veeramachaneni, who supervised the early experimentation and software development phase of the project.

Thanks are due to Leo Celi M.D. (MIMIC Project Lead) and John Danziger M.D. for their assistance and medical advice in helping me formulate my characteristic problem and defining it.

I'd also like to thank my wife Carla, my grandmother, and my parents for all their support, love and tireless encouragement.

Contents

1	Introduction	15
1.1	Overview	15
1.2	Medical Background & Data Source	17
1.3	Research Questions & Contributions	18
1.4	Roadmap	20
2	Methods	21
2.1	Overview	21
2.2	Source Data	22
2.3	Patient-day Matrix: A Representation For Normalizing Patient Records	23
2.3.1	Description of Features	25
2.3.2	Identifier Variables	26
2.3.3	Non-temporal Variables	27
2.3.4	Temporal Variables	29
2.3.5	Elixhauser Variables	31
2.3.6	Variable Distributions	33
2.4	Experimental procedure: Propensity Scoring & Matching	37
2.4.1	Experimental Structure	38
2.5	Cohort Selection and Sampling Techniques	40
2.5.1	D- Patient Selection Method 1: Random Selection	40
2.5.2	D- Patient Selection Method 2: Fluid Balance Similarity-Based (FBSB) Selection	41
2.5.3	D- Patient Sampling Method 1: Median Day Sampling	42

2.5.4	D- Patient Sampling Method 2: Patient-Day Bucket Sampling	42
2.5.5	D- Patient Sampling Method 3: Gamma Sampling	43
2.6	Measuring our results: Class Imbalance Penalty	43
3	Results	47
3.1	Overview	47
3.2	Discussion	51
3.3	Experimental Results	54
3.3.1	Random Selection, Median Sampling	54
3.3.2	Random Selection, Patient-Day Sampling	56
3.3.3	DTW Sliding KNN Selection, Patient-Day Sampling $k = 1$. .	57
3.3.4	DTW Sliding KNN Selection, Patient-Day Sampling $k = 10$.	60
3.3.5	DTW Sliding KNN Selection, Patient-Day Sampling $k = 20$.	63
3.3.6	DTW Sliding KNN Selection, Patient-Day Sampling $k = 100$.	66
3.3.7	Random Selection, Gamma Distribution Sampling	69
3.3.8	DTW Sliding KNN Selection, Gamma Distribution Sampling k = 20	70
3.3.9	DTW Sliding KNN Selection, Gamma Distribution Sampling k = 1	73
4	Research Findings and Contributions	77
4.1	Research Findings and Contributions	77
4.2	Software	78
4.3	Future work	78
5	Glossary	81

List of Figures

2-1	Diagram of the MIMIC II Database and its component parts.	22
2-2	Diagram of two example patient timelines. Even in the case where patients have very similar event types, counts and timelines, small differences in timing can make analysis complex.	24
2-3	Matrix of days for each ICUSTAY. The Patient-Day column indicates the number of days since admission. The Post-Diuretics Length-Of-Stay or LOS column indicates the number of days the patient remained in the ICU after the first diuretics administration, including the first administration day.	24
2-4	Non-temporal variable distributions.	34
2-5	Temporal variable distributions.	35
2-6	Elixhauser variable distributions.	36
2-7	Building the Propensity Score model. Figure by Rammazotti [8] . . .	39
2-8	Class Imbalance Penalty equation	43
3-1	Plot of Imbalance Penalty scores for all experimental results. Parameter A refers to the number of consecutive lag days considered in FSBS DTW Selection. Parameter k refers to the number of closely matched patients considered to be tied at the end of the selection procedure we randomly select the final cohort of D- from these k observations. . . .	49
3-2	Fitting a Gamma curve (in red) to the histogram of treatment patient administration days.	50

List of Tables

2.1	Identifier variables and descriptions.	26
2.2	Non-temporal Variables and descriptions.	28
2.3	Temporal Variables and descriptions.	30
2.4	Elixhauser Variables and descriptions.	32
2.5	List of crowd-proposed, self-extracted covariates	38
3.1	Table of Imbalance Penalty values in order of increasing class balance	48
3.2	Quintiles - Random Selection, Median Sampling	54
3.3	LOS - Random Selection, Median Sampling	54
3.4	Mortality - Random Selection, Median Sampling	55
3.5	Quintiles - Random Selection, Patient Day Sampling	56
3.6	LOS - Random Selection, Patient Day Sampling	56
3.7	Mortality - Random Selection, Patient Day Sampling	56
3.8	Quintiles - DTW KNN (lag=5, k=1)	57
3.9	LOS - DTW KNN (lag=5, k=1)	57
3.10	Mortality - DTW KNN (lag=5, k=1)	57
3.11	Quintiles - DTW KNN (lag=4, k=1)	58
3.12	LOS - DTW KNN (lag=4, k=1)	58
3.13	Mortality - DTW KNN (lag=4, k=1)	58
3.14	Quintiles - DTW KNN (lag=3, k=1)	59
3.15	LOS - DTW KNN (lag=3, k=1)	59
3.16	Mortality - DTW KNN (lag=3, k=1)	59
3.17	Quintiles - DTW KNN (lag=5, k=10)	60

3.18	LOS - DTW KNN (lag=5, k=10)	60
3.19	Mortality - DTW KNN (lag=5, k=10)	60
3.20	Quintiles - DTW KNN (lag=4, k=10)	61
3.21	LOS - DTW KNN (lag=4, k=10)	61
3.22	Mortality - DTW KNN (lag=4, k=10)	61
3.23	Quintiles - DTW KNN (lag=3, k=10)	62
3.24	LOS - DTW KNN (lag=3, k=10)	62
3.25	Mortality - DTW KNN (lag=3, k=10)	62
3.26	Quintiles - DTW KNN (lag=5, k=20)	63
3.27	LOS - DTW KNN (lag=5, k=20)	63
3.28	Mortality - DTW KNN (lag=5, k=20)	63
3.29	Quintiles - DTW KNN (lag=4, k=20)	64
3.30	LOS - DTW KNN (lag=4, k=20)	64
3.31	Mortality - DTW KNN (lag=4, k=20)	64
3.32	Quintiles - DTW KNN (lag=3, k=20)	65
3.33	LOS - DTW KNN (lag=3, k=20)	65
3.34	Mortality - DTW KNN (lag=3, k=20)	65
3.35	Quintiles - DTW KNN (lag=5, k=100)	66
3.36	LOS - DTW KNN (lag=5, k=100)	66
3.37	Mortality - DTW KNN (lag=5, k=100)	66
3.38	Quintiles - DTW KNN (lag=4, k=100)	67
3.39	LOS - DTW KNN (lag=4, k=100)	67
3.40	Mortality - DTW KNN (lag=4, k=100)	67
3.41	Quintiles - DTW KNN (lag=3, k=100)	68
3.42	LOS - DTW KNN (lag=3, k=100)	68
3.43	Mortality - DTW KNN (lag=3, k=100)	68
3.44	Quintiles - Random Selection, Gamma Distribution Sampling	69
3.45	LOS - Random Selection, Gamma Distribution Sampling	69
3.46	Mortality - Random Selection, Gamma Distribution Sampling	69
3.47	Quintiles - DTW KNN Gamma (lag=5, k=20)	70

3.48	LOS - DTW KNN Gamma (lag=5, k=20)	70
3.49	Mortality - DTW KNN Gamma (lag=5, k=20)	70
3.50	Quintiles - DTW KNN Gamma (lag=4, k=20)	71
3.51	LOS - DTW KNN Gamma (lag=4, k=20)	71
3.52	Mortality - DTW KNN Gamma (lag=4, k=20)	71
3.53	Quintiles - DTW KNN Gamma (lag=3, k=20)	72
3.54	LOS - DTW KNN Gamma (lag=3, k=20)	72
3.55	Mortality - DTW KNN Gamma (lag=3, k=20)	72
3.56	Quintiles - DTW KNN Gamma (lag=5, k=1)	73
3.57	LOS - DTW KNN Gamma (lag=5, k=1)	73
3.58	Mortality - DTW KNN Gamma (lag=5, k=1)	73
3.59	Quintiles - DTW KNN Gamma (lag=4, k=1)	74
3.60	LOS - DTW KNN Gamma (lag=4, k=1)	74
3.61	Mortality - DTW KNN Gamma (lag=4, k=1)	74
3.62	Quintiles - DTW KNN Gamma (lag=3, k=1)	75
3.63	LOS - DTW KNN Gamma (lag=3, k=1)	75
3.64	Mortality - DTW KNN Gamma (lag=3, k=1)	75

Chapter 1

Introduction

1.1 Overview

Retrospective studies are very important for both pre-study analysis and for situations where running a full RCT may be too expensive, dangerous or unethical. How do we study things like the impact of putting a new police outpost in a given neighborhood, or the effect of AIDS therapy upon mortality? In either case, we cannot setup a normal RCT and randomly assign stations or therapy. Our best course of action would be to examine prior cases, then carefully model and account for each feature with respect to our response, and then we may be able to find some trends between the treatment-negative and treatment-positive groups that are significant and persistent.

Accounting for factors that are correlated with both the dependent variables and the independent variable, known as *confounders*, in a modeling problem poses difficult challenges to retrospective studies, and is a major focus of many retrospective study procedures. Improving upon existing techniques to better account for these confounders would increase the generalizability of subsequent retrospective analysis. Also, modifying these techniques to grapple with variables that vary over time, or time-series, is important for generalizing to a wide range of important problems. In recent years there has been an explosion in available time series datasets as collection mechanisms have improved and the expectations for publicly available data have risen. This enables many insightful and never-before-possible retrospective studies, if only

researchers can be equipped with practical techniques for analysis and comparison. Traditionally, time series data has seen its best applications in Univariate modeling settings, through the application of ARIMA and exponential smoothing techniques. This is not the way forward though for the vast majority of datasets — the other collected features need to be included in the modeling process if we are to build useful, realistic and generalizable models of the world, which can adequately deal with confounders. This kind of setup is called a cross-sectional time series problem.

This thesis centers around the two most significant data science hurdles for cross-sectional time series in a retrospective study. We address them as our two experimental questions. For event-based data, the timeline of results that we wish to consider as a single case or observation can vary widely. One observation may have 12 events spanning 12 hours, one per hour, while another observation may have 5 events in 1 hour and 1 event 36 hours later. Comparing these two observations is beyond the capabilities of many statistical techniques used in retrospective studies. This leads us to our first question:

- Question 1: How should we represent event-driven and irregular time-series data for use in a retrospective study?

Retrospective studies begin with a hypothesis about a given treatments effect on measured outcomes and a set of source data. Ultimately the goal of the study is to produce many comparable sets of treatment and non-treatment observations to use in evaluating the validity of the hypothesis. In a healthcare retrospective study, comparable usually means that each patient in the treatment group is matched with patients in the non-treatment group that have equivalent health status, and that we can achieve class balance across all levels of health status. Our second question concerns this class balance:

- Question 2: How do we select and sample cohorts from our study group to produce strata wherein observations have comparable features and likelihood of treatment at all levels (henceforth referred to as *class balanced quintiles*) despite the influence of confounding variables?

As a characteristic problem to explore and test new techniques for answering these questions, we have chosen to study a sepsis-patient study group from the MIMIC II database, extracting event records used to inform diuretics-related care decisions. First we tackle Question 1 — the representation problem inherent in cross-sectional time-series datasets — using a lag variable and binning approach, and examine the resulting matrix. Then we address Question 2, comparing and analyzing a number of cohort selection/sampling procedures for that matrix to test how to achieve excellent class balance for non-treatment observations in this data. The ultimate aims are to build and evaluate practical tools for addressing these fundamental questions and unlock the promise in many new and interesting datasets.

1.2 Medical Background & Data Source

Sepsis is a specific condition described as severe full body inflammation in response to a serious infection. Patients diagnosed with sepsis suffer from a long list of symptoms: very low blood pressure and heart racing, swelling, flushing, fever and hyperventilation. These patients are routinely treated through IV and thus their body fluid levels are elevated for a time. [3] Sometimes these patients' bodies resolve this issue themselves quickly and naturally, but when they do not doctors are left with a choice: wait until the patient's body is able to process the fluids and bring them under control (risky for the patient if this is too slow), or prescribe a diuretic drug to induce the patient's body to reduce fluid levels back to normal. [3] When the body has difficulty dealing with these conditions and these fluid levels remain high or unstable, drugs called diuretics are commonly used to reduce them by causing the patient to urinate out the excess water. These diuretics work by inducing the kidneys to expel sodium, in turn causing the body to release water from the bloodstream. Treatment of many conditions such as kidney failure, hypertension, heart failure, etc., regularly involve the use of diuretics.

Given how common and important this decision point is, investigating it through our ICU dataset is interesting and relevant — a researcher might want to use tech-

niques like the ones we propose to check if use of diuretics is beneficial or harmful for sepsis patients. Is it better to allow critical sepsis patients to resolve elevated fluid levels without drugs or to prescribe diuretics? Different doctors might judge the same situation very differently; given this grey area, lets see if our study results clearly identify any best practices regarding diuretics use.

This thesis will run a retrospective analysis into diuretics use as a *characteristic problem* and setting for building and testing new techniques. Our focus was on a set of 332 sepsis patients who received diuretics, as found in the MIMIC2 database compiled from Beth Israel Deaconess Medical Center in Boston, MA. From this we extracted several features that our medical partners identified as key indicators for diuretics treatment.

The next challenge was to assemble sepsis patients who did not receive diuretics who could be balanced with the ones who did. This can be done with a variety of selection and sampling techniques. One technique is better than another if it achieves better overall class balance across all 5 quintiles, that is, when the entire D+, D- group is stratified based on propensity, there are sufficient relative quantities of D+, D- in each strata of health status to make statistically sound comparisons. We used the Propensity Score modeling technique by Rosenbaum and Rubin [7] as our modeling procedure, allowing us to experiment with several selection and sampling techniques. Several of the proposed techniques appear competent at constructing *class balanced quintiles* of treatment and non-treatment patients with moderate amounts of data, and set the stage for making similar kinds of analysis straightforward with our software tools.

1.3 Research Questions & Contributions

This thesis asks the following questions for data science and statistical analysis in a cross-sectional time-series retrospective study:

- How do we prepare event-driven and irregular data for modeling?

- We smoothed our event-based data on days, taking mean and median for each variable, for each day. This makes previously incomparable observations comparable despite different timing and events.
- Which selection and sampling techniques produce the most balanced quintiles?
 - We developed two selection and three sampling techniques for preparing retrospective studies on irregular time-series data, and software for performing bootstrap iterations for combinations of techniques. We found that using a Fluid-Balance Similarity-Based Dynamic Time Warp selection procedure with nearest neighbor parameter $k=1$ and using a Gamma distribution for sampling days produced consistently better class balance than all other methods when bootstrapped over 100 independent runs.
- How can we quantitatively measure the balance of a set of quintiles?
 - We developed the Class Imbalance Penalty, a class balance metric for selecting quintiles for statistically sound comparisons, which enables relative ranking of stratified matching procedures. We have demonstrated that it gives low scores for the procedures that produce the most class-balanced quintile results for all strata and lend themselves to statistically sound future analysis.

For our retrospective study on Diuretics treatment effects for Sepsis patients, we ask the following questions for our study’s context:

- What differences in class balance do we observe for the tested selection and sampling techniques?
 - We found that sampling had a stronger effect on class balance than selection, and that using a Gamma distribution fit to pick sampling days produced more statistically comparable quintiles than all other methods, for any selection procedure.

- Is Fluid-Balance Similarity-Based selection effective as a health status similarity technique for sepsis patients?
 - We found that Fluid-Balance Similarity-Based DTW selection with $k=1$ and $\text{lag}=3$ in combination with Gamma distribution sampling is the best combination among the tested methods, but purely Random sampling outperforms it for many other values of k .
- What do our experimental procedure results mean for our study’s context?
 - We found strong evidence suggesting that physician and provider data are key missing covariates for predicting Diuretics administration in the ICU.

1.4 Roadmap

This thesis proceeds as follows:

- In **Chapter 2** we detail our data and the experimental procedures used in this thesis, including the Patient-Day Matrix, Dynamic-Time-Warp selection, Gamma Selection and the Imbalance Penalty
- In **Chapter 3** we present the results from our experiments and discuss our findings
- In **Chapter 4** we present our conclusions and describe opportunities for future work
- In **Chapter 5** we provide a glossary of terms used in this thesis
- Finally, we provide a bibliography of our sources

Chapter 2

Methods

2.1 Overview

This chapter outlines the methods and experimental procedures used in this thesis. To answer our 4 experimental questions, we need to both develop a test problem to use as a benchmark and construct an experimental procedure around the selection strategies we want to evaluate for optimal class balance. Each of the sections below will describe how a certain procedure was conducted or how a given step in the analysis was performed. It proceeds as follows:

- In Section 2.2 we describe the source data and reasons behind the selection of the dataset we use as our characteristic problem.
- In Section 2.3 we describe a transform to a new data representation called the Patient-Day Matrix
- In Section 2.4 we describe the experimental procedure, involving bootstrapping and Propensity Score matching on quintiles
- In Section 2.5 we describe the non-treatment selection and sampling procedures used within the experimental procedure
- In Section 2.6 we describe the metric used to evaluate class balance quantitatively

2.2 Source Data

The data used in our study was taken from the MIMIC II database (Multiparameter Intelligent Monitoring in Intensive Care), which has de-identified physiological data from thousands of patients who visited the ICU between 2001 and 2007 at Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts. [5] This database is a good starting point for this study for many reasons: it is freely available to referenced researchers in relevant fields, contains data from a wide range of inter-hospital perspectives (medical ICU, surgical ICU, cardiac care unit, cardiac surgery recovery unit), and contains high temporal resolution in many areas. Medical data is a very common setting for time-series retrospective studies, so this dataset is very realistic test problem for developing and evaluating new techniques.

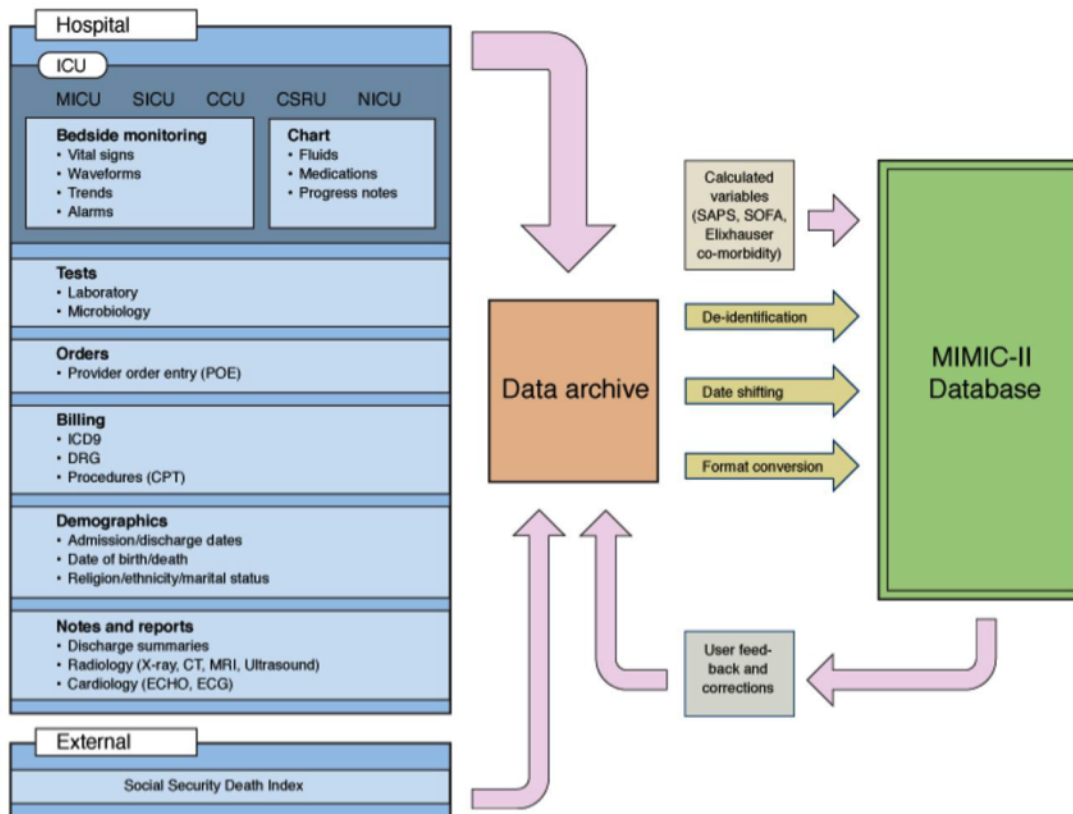


Figure 2-1: Diagram of the MIMIC II Database and its component parts.

The MIMIC II Clinical Database contains physiological values verified by nurses

(usually on an hourly basis), nurses notes, IV medications, fluid balances, demographics, physician orders, discharge summaries, ICD9 codes and more. These values were collected within the hospital using a Phillips CareVue Clinical Information System deployed in all the study ICUs. [5] Significant post-processing has been done by the team compiling the dataset at the hospital, to obtain integrated and unified records for each patient, and de-identify it in compliance with HIPAA (Health Insurance Portability and Accountability Act) standards.

Our study utilizes data gathered from the 32,000 patients represented in the *Clinical Database*. This database has records, which are recorded as *events* for example, a record for an administered medication would be identified by several features: Subject ID, Hospital Admission ID, ICUSTAY ID, the name of the drug, and a timestamp. We can think of these *events* as forming a time-series of data about the patient, describing the patients physiological changes and intervention events.

2.3 Patient-day Matrix: A Representation For Normalizing Patient Records

Event-driven time series data present a representation problem: how do we express and record changes in each time-bound feature in a consistent way across all patients in the dataset?

Figure 2-2 shows two hypothetical patient timelines with different intervals and sequences of events. Without any smoothing, an event from Patient 1 is tough to compare with a similar event for Patient 2 since they may occur at different times and have very different preceding and succeeding events. By smoothing over days, the contents of each day become comparable between the two patients.

We call the transformation applied in this study the Patient-Day-Matrix (PDM). For all patients in the study with recorded temporal events, we **generate a new row in the dataset for each 24-hour period they have been in the ICU**, and report the means and medians for each feature, for each day. Figure 2-3 shows an

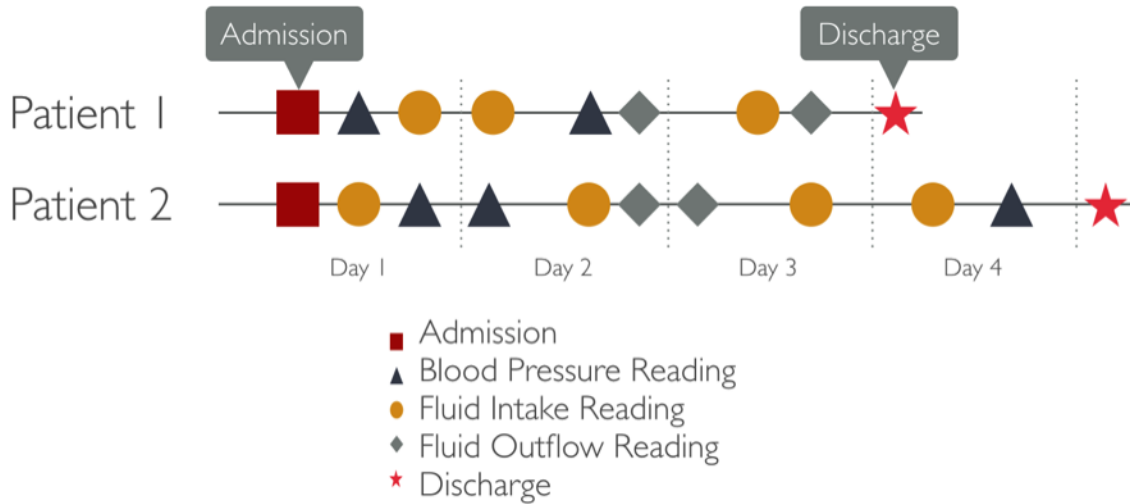


Figure 2-2: Diagram of two example patient timelines. Even in the case where patients have very similar event types, counts and timelines, small differences in timing can make analysis complex.

ICUSTAY	Patient-Day	Systolic BP Mean	Fluid Intake Mean	...	LOS	Diuretics on this day?
1234	1	110	1.3 L	...	3	0
1234	2	122	1.0 L	...	3	1
1234	3	115	2.5 L	...	3	0
1234	4	120	2.0 L	...	3	1

Figure 2-3: Matrix of days for each ICUSTAY. The Patient-Day column indicates the number of days since admission. The Post-Diuretics Length-Of-Stay or LOS column indicates the number of days the patient remained in the ICU after the first diuretics administration, including the first administration day.

example, with the incrementing Patient-Day column indicating the number of days since admission, beginning at 1. Two ICUSTAYs are considered to be independent, even if they record events for the same patient this indicates that the patient has been in the ICU multiple times.

2.3.1 Description of Features

After first exploring the data available in MIMIC II we wrote scripts to extract our new time-series interval features. Our aim with this extraction was to characterize general patterns in the data and smooth out the very noisy and irregular events from MIMIC II into a time-series of regular intervals, exposing seasonality, cyclicity and trending in the each patients record. The features fall into several categories:

- Identification features such as ICUstay and Patient-day are the primary keys for identifying a row of data. Each tuple of these is unique in the dataset and represents the interval for which the other features in the row apply.
- Non-temporal features consist of values describing descriptive characteristics of the patient for the entire ICUstay. They are recorded in each row such that the row represents a complete snapshot (to permit cross-sectional analysis) but remain constant for all the patient days in a given ICUstay.
- Temporal features consist of values that change over the interval and across intervals. For each temporal feature, the value recorded in the data is formed by querying for all the events (say, medications administered) falling between T and T-1 24-hour periods since admission to the ICU and computing the mean. If there are no events occurring in the interval, we record NaN to represent a null value.
- Exlihauser-related features consist of measurements recorded at hospital admission, screening for risks affecting patient outcomes. They are recorded as one overall score and several binaries variables for the presence of each condition. They are designed to record important comorbidities or conditions present on admission that are not related directly to the main reason for hospitalization, but that increase the intensity of resources used or increase the likelihood of a poor outcome. [4]

2.3.2 Identifier Variables

Table 2.1 summarizes the identifier variables in the extracted features. ICUstay and Patient day uniquely identify each row in the dataset.

	Name	Description
X1	ICUstay ID	Unique identifier for each stay in the ICU. A single patient may have several stays, and therefore have several IDs
X2	Subject ID	Unique identifier for each patient
X3	Day in the ICU (Patient day)	The days since the current ICUstay began

Table 2.1: Identifier variables and descriptions.

2.3.3 Non-temporal Variables

Table 2.2 summarizes the non-temporal variables in the extracted features. Non-temporal variables in the study refer to characteristic information about the patient and the ICUstay that dont vary across each day in the recorded ICU visit.

- X4, X5, X9 : These features are gathered from the patient demographic table in MIMIC II based upon subject ID.
- X6, X7 : These two features are compiled by looking for any events of the given type listed for that ICUstay ID. If there is at least one event on record, we record a 2 in each row for that ICUstay otherwise we record a 1.
- X8 : Mortality within 30 days is calculated using the timestamp for discharge from the ICU, and any available listing for date of death. For patients without any listing, we record 1 for No, and for those with a listing with a difference of less than 30 days from discharge time we record a 2.
- X10 : Post-Diuretics-Length-Of-Stay is an experimentally specific variable, expressing the number of patient-days the patient stays in the ICU following the first treatment of diuretics. It is a number strictly less than or equal to the number of total patient days for a given ICUSTAY. For a patient discharged on the same day as they received diuretics for the first time, this value is one. Note that for D- patients there is no administration and therefore no first administration day, so we record NaN for this feature. Later during preparation for modeling, we impute the administration days for D- patients using one of the tested sampling techniques, so both D+ and D- patients have values at modeling time.

	Name	Description
X4	Gender	Encoded as 1 for Male, and 2 for Female
X5	Ethnicity	Our MIMIC sample includes 15 specific racial groups, in addition to the following special categories: 'MULTI RACE ETHNICITY', 'OTHER', 'PATIENT DECLINED TO ANSWER', 'UNABLE TO OBTAIN', 'UNKNOWN/NOT SPECIFIED'
X6	Vasopressors given during ICU stay	Encoded as 1 for No or 2 for Yes
X7	Ventilation given during ICU stay	Encoded as 1 for No or 2 for Yes
X8	Mortality	Mortality within 30 days of the last day in ICU, with 1 for No and 2 for Yes
X9	Age	Age of each patient, recorded in MIMIC at time of admission
X10	Post Diuretics Length of Stay given during ICU stay	Length of Stay in the ICU after first administration of diuretics

Table 2.2: Non-temporal Variables and descriptions.

2.3.4 Temporal Variables

Table 2.3 summarizes the temporal variables in the extracted features. These are recorded as events in the MIMIC II source data, and are binned into features based upon the patients start day and day-length time intervals.

- X11 : If there are any diuretics medication events (from the medications table in MIMIC II) for this day, we record a 2 for yes otherwise we record 1 for an empty query.
- X12 : From the query for X11 if there are positive results we record the count of distinct administrations in the query for that day, greater than or equal to 1 administration. For patient days without any administrations, we record NaN.
- X13 : This feature is the simple arithmetic difference between Fluid Inputs and Fluid Outputs (Net Fluid Inputs or $(\text{mean}(X17) - \text{mean}(X16))$). This is some simple feature engineering since Fluids are important features when analyzing Diuretics administration and patients with Sepsis.
- X14, X15 : Systolic and Diastolic blood pressure averages are recorded separately. Each is a mean of all recorded values during the Patient-Day.
- X16, X17 : Fluid Inputs and Outputs are computed by averaging all the input and output events listed for a particular Patient-Day.
- X18 : Creatinine administered to the patient is averaged across the patient-day interval, or is listed as zero if there are no recorded administrations.
- X19 : SAPS score to asses the severity patients current condition this is a temporal variable since it is reassessed daily during the ICUstay. [9]
- X20 : SOFA score to asses the patients organ function this is a temporal variable that is reassessed for each Patient-Day. [11]

	Name	Description
X11	Diuretics given on this patient day (binary)	Encoded as 1 for No or 2 for Yes
X12	Number of times diuretics were given on this patient day	0 for none or 4 for four times for this patient, for this ICU stay, for this day
X13	Fluid balance Average	Net Fluid Inputs (Fluid Inputs - Fluid Outputs)
X14	Diastolic Blood Pressure (ABPmean)	Average of values recorded over the patient-day
X15	Systolic Blood Pressure Average (ABP)	Average of values recorded over the patient-day
X16	Fluid Outputs Average	Average of values recorded over the patient-day
X17	Fluid Inputs Average	Average of values recorded over the patient-day
X18	Creatinine Average	Average of values recorded over the patient-day
X19	Simplified Acute Physiology Score (SAPS) Score	Score between 0 and 163 expressing the severity of the patients condition
X20	Sequential Organ Failure Assessment (SOFA) Score	Score evaluating the level of the patients organ function

Table 2.3: Temporal Variables and descriptions.

2.3.5 Elixhauser Variables

Table 2.4 summarizes the Elixhauser-related variables in the extracted features. The Elixhauser Comorbidity Score is a rating expressing the presence of factors contributing to patient death, independent of the other aspects of the patients condition. [4] Each of these factors has been empirically demonstrated to correlate with substantial increases in length of stay, hospital charges, and mortality both for heterogeneous and homogeneous disease groups. [4] In addition to the overall score, several of the 30 individual components making up the Elixhauser Comorbidity Score that are suspected to have a relationship with Sepsis and its treatment have been included as binary variables.

- X21 : Score from 0 to 10 expressing concurrent presence of nonmalignant diseases.
- X22 through X30 : Binary variables indicating the concurrent presence of a specific disease along with the Sepsis we have filtered for in our sample.

	Name	Description
X21	Elixhauser Comorbidity Score	Integer from 0 to 10, recorded many times during the patient stay. If more than one observation exists for a patient-day then we average them
X22	Elixhauser Comorbidity Binaries or ECB (1 of 9)	-1 for no, 1 for yes, depending on whether or not the patient exhibits Congestive Heart Failure
X23	ECB (2 of 9)	Cardiac Arrhythmias
X24	ECB (3 of 9)	Valvular Disease
X25	ECB (4 of 9)	Hypertension
X26	ECB (5 of 9)	Diabetes (Uncomplicated)
X27	ECB (6 of 9)	Diabetes (Complicated)
X28	ECB (7 of 9)	Renal Failure
X29	ECB (8 of 9)	Liver Disease
X30	ECB (9 of 9)	Obesity

Table 2.4: Elixhauser Variables and descriptions.

The finalized Patient Data Matrix has these characteristics:

- 2,341 unique patients
- 3,503 unique ICU stays
- 2,807 unique hospital stays
- 332 unique patients prescribed Diuretics
- 349 unique ICU stays with Diuretics prescribed
- 32,678 unique Patient-Days
- 10,100 Patient-Days within 30 days of mortality

We wrote several robust scripts to extract various groups of features and then concatenated them into a single matrix. For each variable we then sampled and checked that the events in the patient record were being binned and summarized appropriately across the patient days, before finally saving the matrix as a MATLAB struct. This is a convenient format for later use in our experimental procedure. Figures 2-4, 2-5 and 2-6 show the distribution for each feature in the set of features (identifier variables such as ICUstay ID, and Subject ID are excluded). Please refer to the open source software on Github at the following link for more details:

- [github.mit.edu/ALFAGroup/Clinical_Time_Series_For_Diuretics_brian_bell](https://github.com/mit-alfagroup/Clinical_Time_Series_For_Diuretics_brian_bell)

2.3.6 Variable Distributions

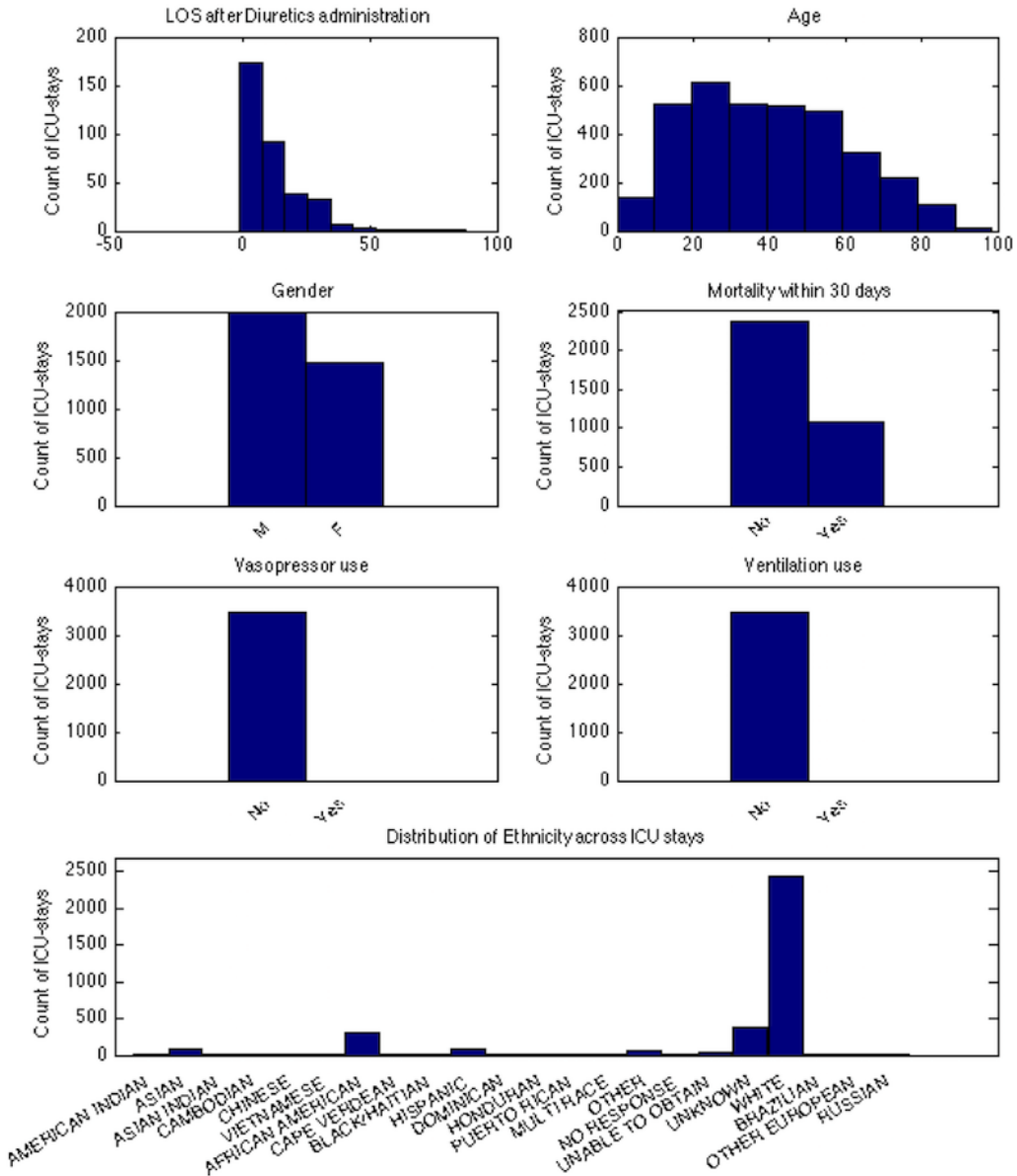


Figure 2-4: Non-temporal variable distributions.

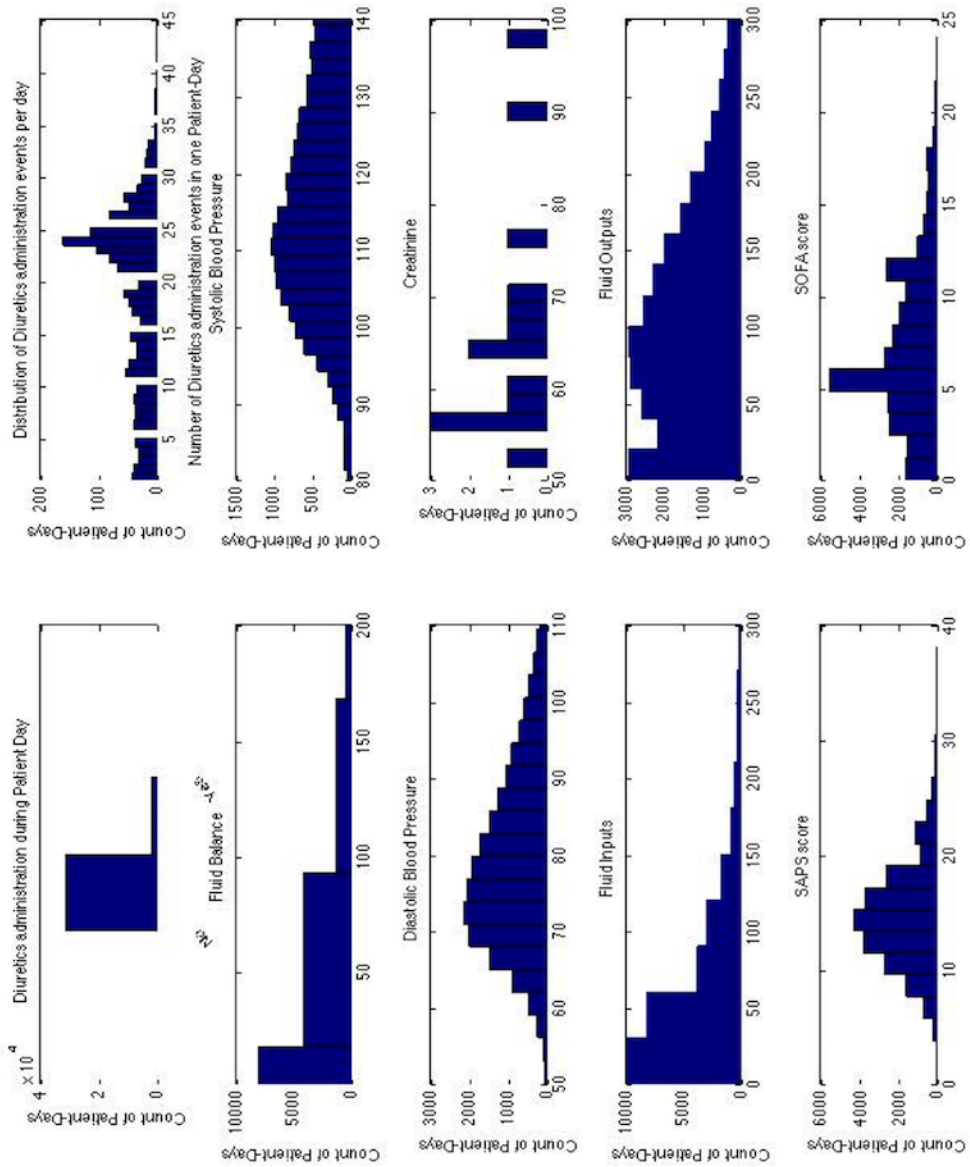


Figure 2-5: Temporal variable distributions.

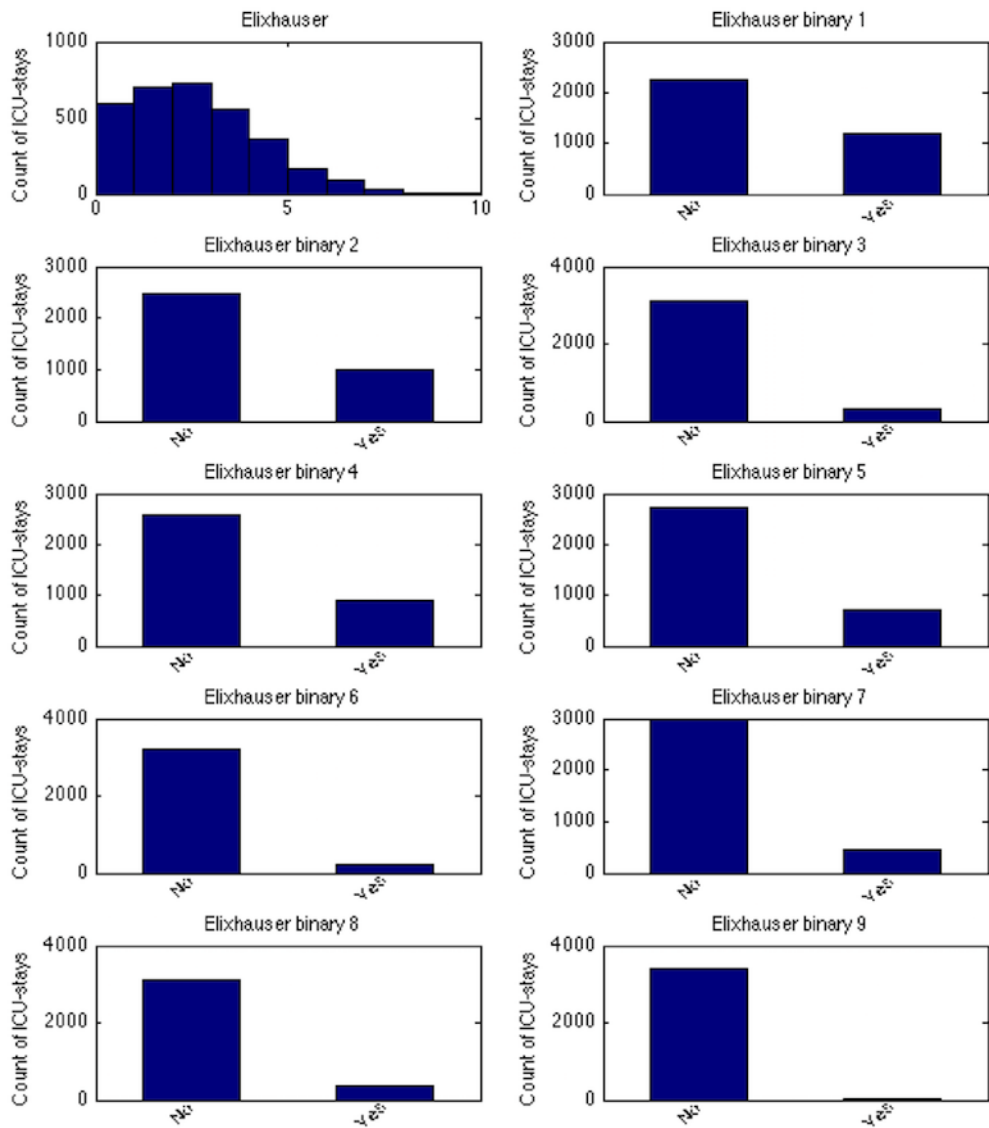


Figure 2-6: Elixhauser variable distributions.

2.4 Experimental procedure: Propensity Scoring & Matching

With our data matrix constructed, we move on to building a flexible and reusable experimental procedure to evaluate outcomes for observational study data. For our Observational Study, its key that we are able to compare and assert that the patients in the D+ treatment group and D- non-treatment group are similar in terms of likelihood to receive diuretics. We could achieve this by forming many sub-groups of patients and evaluating that all features are balanced among the groups. Unfortunately, for studies with more than a few features this balance becomes much more complex and it can also be difficult to achieve in cases where the data has a big imbalance in count of treatment versus non-treatment cases (like in our dataset). Additionally, our studys event-based data makes this balancing much more challenging because the difference in timing represents a big dimensional increase in the variation of the patient records, dramatically increasing the feature space to balance, to the point of being intractable.

Propensity Score Matching, first proposed by Rosenbaum and Rubin in 1984, has become a widely accepted technique for reducing bias when estimating the impact of treatment effects in an Observational Study. [1][7][8] It uses statistical modeling techniques to express the likelihood of a particular patient being assigned to a treatment, across all features. Then we are able to effectively work with observed data even if the assignment of patients to treatments is not random, and even if there are important differences in the patient characteristics between the treatment group and the control.

We will use Propensity Score Matching as our 'typical' observational study model, and test different non-treatment patient selection and sampling techniques to obtain class balance findings for each within the model. The next section outlines our procedure.

2.4.1 Experimental Structure

We have employed Rosenbaum and Rubins core approach in the heart of our study. Their method involves the 4 steps in **Bold** - the steps before and after are the additional evaluation, bootstrap and benchmarking steps we perform for this study:

Procedure
<p><i>Given an initial D+ patient cohort</i> <i>While num_runs < 100</i></p> <ul style="list-style-type: none"> • Step 1: Select D- patient cohort from Study Group • Step 2: Sample D- cohort to select administration days • Step 3: Build a propensity model on D+ U D- and select initial features • Step 4: Stratify and assess balance (Two-way ANOVA for F-Ratios) • Step 5: Refine the model (Add features and interactions) • Step 6: Decide whether the desired balance is achieved or go back to Step 3 • Step 7: Assign propensity value to each patient in the D+ U D-cohort. Rank patients by propensity and stratify into quintiles • Step 8: Save quintile values to a data structure <p>Step 9: For each quintile, compute its mean over 100 runs Step 10: Score the mean_quintile using the Class Imbalance Penalty to assess balance</p>

Table 2.5: List of crowd-proposed, self-extracted covariates

Inside Rosenbaum and Rubins procedure is an internal health status balance assessment that performs forward feature selection for a Logistic Regression model. [2][7] This is a separate goal, and necessary precondition for Class Balance. If the ANOVA F-Ratio for a feature not in the model is high, then it is added to the model, and if it remains high then its interactions are also added to the model. The ultimate health status balance achieved at the end of the procedure (as measured by the Imbalance Penalty) is a function of the input data and the procedure. By holding the dataset and procedure constant, we can modify Steps 1 and 2 to test their influence on the results.

Propensity score prediction is performed on the dataset using the binary column for diuretics administration as Y, and holding out the variable for number of administrations (which would add serious leakage if left in the model). With the model fit and scores for each row computed, the patients are stratified into equally sized quintiles based upon ordinal ranking of scores. The propensity scoring process results in quintiles that are balanced on features and balanced on the likelihood of diuretics administration. Stratifications without both Treatment and Non-treatment patients in each quintile are rejected and the entire process (beginning with D- patient selection) is re-run until there are both types of patients in each quintile when stratified by propensity scores. We expect from Rosenbaum and Rubin approximately 90% reduction in bias for each of the features when we stratify on the quintiles of the population propensity score, so we stratify on our estimated propensity score to achieve some of this reduction. [7][8] In any particular subclass that is relatively homogenous based upon propensity score, the distributions of the features are approximately the same between the treatment and non-treatment groups (approximately so because the propensity scores are binned rather than being exactly the same). Figure 2-7 illustrates the Rosenbaum and Rubin procedure:

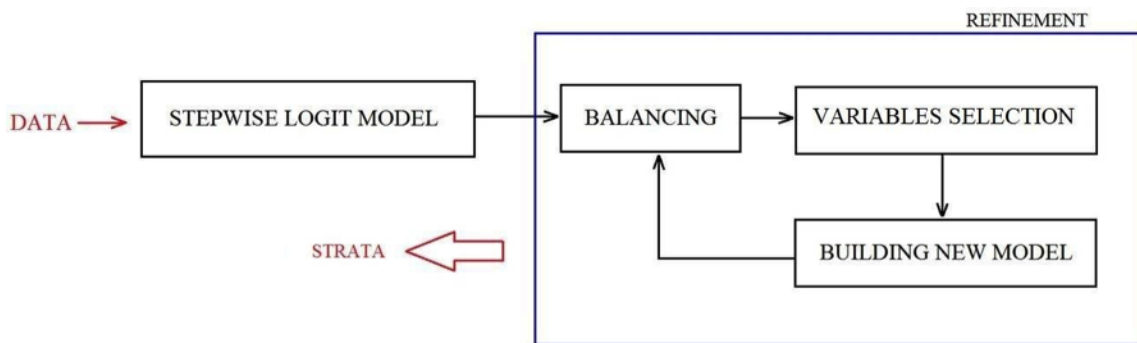


Figure 2-7: Building the Propensity Score model. Figure by Rammazotti [8]

To reduce the bias in our results, we have repeated the above process (with replacement) 100 times from initial sampling of patients through the collection of quintile data, and then aggregated the results. Since each run involved selecting D- patients and sampling, our aggregated results reduce the uncertainty of sampling and give us more stable imbalance penalty values to compare across techniques.

2.5 Cohort Selection and Sampling Techniques

To obtain our cohort, we need to select comparable non-treatment patients to include in our propensity scoring, and we need to sample the comparable slice of data for when each D- patient *would* have received diuretics. We need to somehow find the precise slices of D- patient data that we will take as representative cases. Even among the diuretics patients, there are a lot of rows we shouldn't include in the model. In this thesis we have considered 2 different methods of patient selection and 3 different methods for selecting non-treatment patient timepoints; each is described in detail below.

2.5.1 D- Patient Selection Method 1: Random Selection

The first Patient selection method we use is Random Selection. This is not a selection of any D- patient in the set. We must exclude patients who have NaNs on their administration days, and exclude patients with fewer Patient-Days than the sampling model we have chosen. For example, for median day sampling we are only randomly sampling from the subset of D- patients who have at least that many patient days, which is a much smaller set. In the Patient bucket sampling approach, for each D+ patient we will random select one D- patient from the subset who have at least as many patient days.

By using day-length slices exclusively, random selection treats each day in the ICUstay record like a fully independent observation, assuming that only the data for a single day is being used to select patients for treatment. If data across multiple days is what is actually relevant, then we aren't constructing our cohort with matches

that are comparable based upon the most important factors.

2.5.2 D- Patient Selection Method 2: Fluid Balance Similarity-Based (FBSB) Selection

Fluid Balance Similarity-Based (FBSB) selection based upon each D+ patients time-series is the more strict patient selection method we tested. We use Dynamic Time Warping to measure the nearness of each D- patient to a given D+ patient, based upon a series of consecutive days Fluid Balance feature (univariate time series approach). For the lag=A days leading up to first administration of diuretics, we collect the time-series of Fluid Balance values for each D+ patient, discarding those with missing Fluid Balance values for this range. Then for each remaining D+ patient, we loop through all the PDM rows corresponding to D- patients and use a sliding window to find the optimal match of lag=A sequential days. The DTW scores allow us to select the k number of patients which show the closest match for the lag=A consecutive days. Then we select a patient at random from the top k to sample for inclusion in the cohort.

Unlike random selection, matching based upon time-series has the potential to account for non-independent patient days. By matching each treatment patient with a non-treatment patient that shares not only the same pre-diuretics stay length (if combined with same-day sampling), but also similar statistical makeup in the days prior, we may be able to achieve a better matching. Selecting how many prior days to consider is a challenge matching too closely leads to overfitting, and could also cause us to include confounding data from other overlapping medical conditions (other than sepsis) that a patient may have during longer ICU stays. For this reason, we perform a parameter search over number of lag days from 3-5 (parameter A). We also search over k, the number of patients we randomly select from after recording DTW similarity scores.

2.5.3 D- Patient Sampling Method 1: Median Day Sampling

Collecting features based upon on the median administration day of the Treatment group is a straightforward way to select timepoints for Non-treatment patients. First, we collect all treatment patients and add their first diuretics administration to the propensity scoring dataset. Then we take the median of the patient day column to find the median day timepoint of administration. Given that the median day for all diuretics administrations was the 11th day, we sample all non-treatment patients in our cohort on the 11th day for the propensity scoring dataset. This method doesn't take into account how factors such as length of stay may affect the treatment decision. Sampling Method 2 is designed to take some of these concerns into account.

2.5.4 D- Patient Sampling Method 2: Patient-Day Bucket Sampling

The length of stay in the ICU might be an important variable when analyzing the effects of diuretics. A patient that has only been in the ICU for a short while might have vastly different signals at the time of administration compared with a patient that has been in the ICU for a long time. If a patient is rushed a diuretic, then that patient likely has sepsis issues as a primary condition, versus a patient who is given a diuretic after a few weeks; such a patient may have been admitted to the ICU with a different condition and slowly developed a sepsis condition later on. **By matching each first-administration day from the Treatment group with a Non-treatment patient day at the same timepoint (e.g. both sampled on the 12th day in the ICU),** we ensure that any effects of earlier or later administration are indeed represented in both the drug and control groups, improving on the Median day selection in Sampling Method 1. In addition, we also ensure that there are a wide variety of administration timings in our sample, while keeping the same distribution between treatment and non-treatment groups.

2.5.5 D- Patient Sampling Method 3: Gamma Sampling

For this method, we fit a Gamma distribution to the histogram of first-administration days for D+ patients. With a good fit on the first-administration days, we can take a sample of timepoints from the distribution, which we will use as the sampling days for our D- patients. Given N non-treatment patients that have been selected by one of the patient selection procedures, we select N timepoints from the Gamma distribution. Looping over the N non-treatment patients, we can sample the first patient on the day given by the first Gamma timepoint value, sampling the second patient on the second Gamma value, and so forth. In this way we will approximate the population distribution of diuretics administration timepoints from the sample of D+ in our study group, and then ensure that our non-treatment patients administration days match.

2.6 Measuring our results: Class Imbalance Penalty

To interpret our quintile results quantitatively, we propose the following class imbalance penalty to minimize:

$$\text{Imbalance} = \left(\prod_{Q=1}^5 i_Q * \max(\text{abs}(\text{num_D_neg} - \text{num_D_pos}), j_Q) \right)^{1/5}$$

for $1 \leq i, 1 \leq j$

Figure 2-8: Class Imbalance Penalty equation

This score has several properties that are desirable for this kind of measurement. First, the multiplicative sum will penalize an observation with multiple poorly class-balanced quintiles more heavily than an observation with a single poorly class-balanced quintile, even if that single quintiles balance is much worse in terms of magnitude. This is an important feature, since many poorly balanced quintiles leaves us with few good ones to extract insights from in a retrospective study. Put another way, it is better to have 4 excellent quintiles and 1 very poor quintile to exclude from analysis than to have 5 mediocre quintiles that are of questionable statistical quality. We intend to design a penalty that we can minimize to identify quintiles that are of excellent statistical quality for use in retrospective studies. Next, the penalty within the quintile takes the absolute value of the difference between the D+ and D- treatment, which removes directionality from our metric. Directionality in class-balance doesnt matter to us since both D- skewed and D+ skewed quintiles will be of low statistical quality for later analysis. We tested our class imbalance metric numerous times on varying quintile data, to verify that the relative ranking of excellent to poor class balance functioned as expected.

There are 2 tunable parameters in this metric, **i** and **J**. These two are used to modulate the penaltys impact for certain kinds of class imbalances. In our work **i** is set to 1, but could be increased to widen the distance between observations. **i** could also be varied across quintiles to penalize errors in some quintiles more than others, if a study is primarily concerned with errors on the high or low end of the propensity score range. An example would be a study concerned with whether the patients who are most likely to receive treatment are more likely to receive very aggressive and destructive therapies in comparison to patients with lower propensity scores. For such a study, we would intend to identify selection and propensity scoring techniques that produce good quintiles in the upper range, and rank them more highly than others.

J allows the researcher to set a threshold for how much imbalance they will tolerate in their study. In studies with very large N, an imbalance in the thousands may not have any significant effect on the results, so we can rank results like those equally. In

this study, we set \mathbf{J} to 1 since our N is small and bounded by the number of sepsis patients with clean data in the MIMIC II dataset. If a researcher has done a variance analysis and has a known level for statistically significant deviations in the balance of a particular quintile or of all the quintiles in their study, \mathbf{J} can be set to that level to cleanly differentiate values inside and outside the interval for statistical significance.

Chapter 3

Results

3.1 Overview

In Chapter 2 we outlined the selection and sampling techniques that we used within our experimental procedure. In this chapter we present results from 21 bootstrapped procedural runs, each with a different set of parameters or pair of included techniques. Each result consists of the counts of treatment and non-treatment patients that fall in the five quintiles, the value of the Imbalance Penalty for those counts, and the average Length of Stay and 30-day Mortality of the patients in each treatment-quintile subgroup (2 x 5). Figure 3.1 shows the Imbalance Penalties for all 21 runs plotted against each other, and Table 3.1 displays the raw values, ranked from top to bottom in order of increasing class balance (lower is better).

Selection type, Sampling type	Imbalance Penalty	Imbalance Penalty for lag A=3	Imbalance Penalty for lag A=4	Imbalance Penalty for lag A=5
Random, Me- dian	42.77			
FSBS DTW k=10, Patient- day Bucket		36.7333	40.0994	39.9688
FSBS DTW k=20, Patient- day Bucket		35.7359	38.7192	41.3027
FSBS DTW k=100, Patient- day Bucket		35.6175	37.4682	39.7104
FSBS DTW k=1, Patient- day Bucket		33.9095	33.7424	32.885
Random, Patient-day Bucket	32.5034			
FSBS DTW k = 20, Gamma		30.4246	30.2361	26.6094
Random, Gamma	27.2375			
FSBS DTW k = 1, Gamma		20.0551	21.7113	23.2082

Table 3.1: Table of Imbalance Penalty values in order of increasing class balance

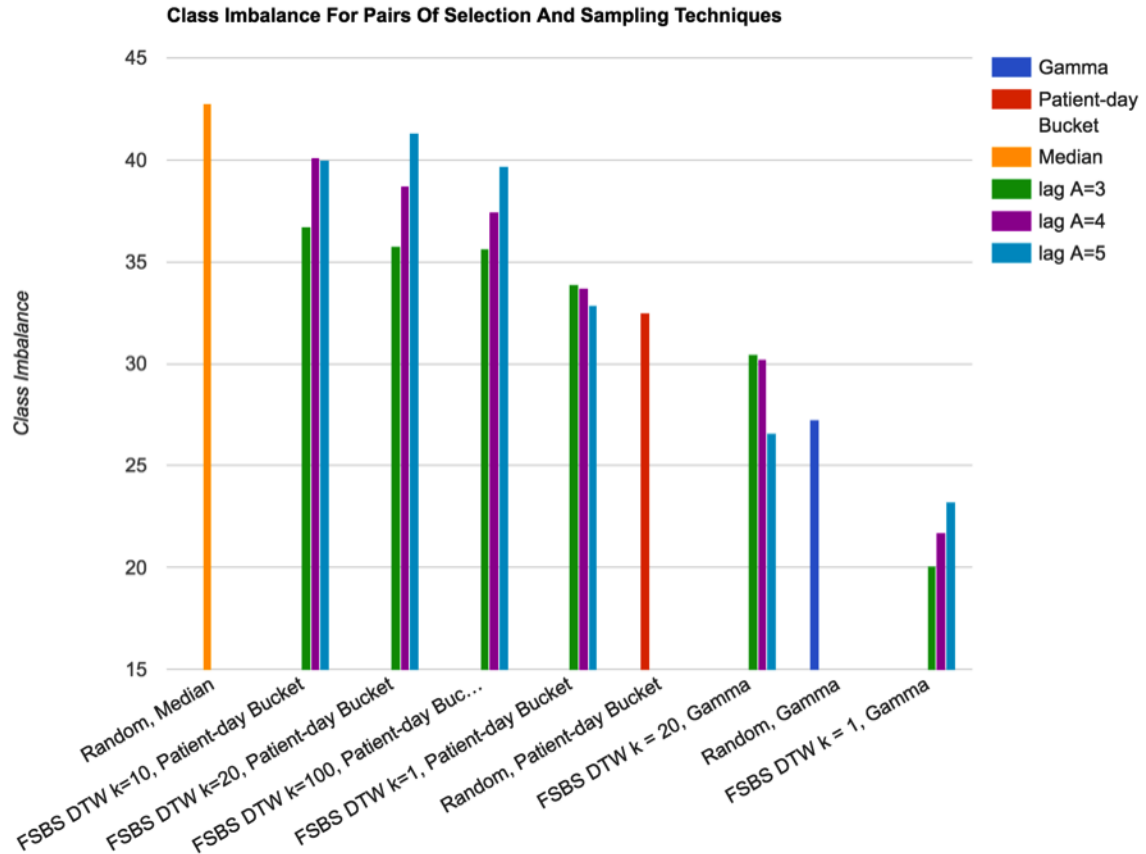


Figure 3-1: Plot of Imbalance Penalty scores for all experimental results. Parameter A refers to the number of consecutive lag days considered in FSBS DTW Selection. Parameter k refers to the number of closely matched patients considered to be tied at the end of the selection procedure we randomly select the final cohort of D- from these k observations.

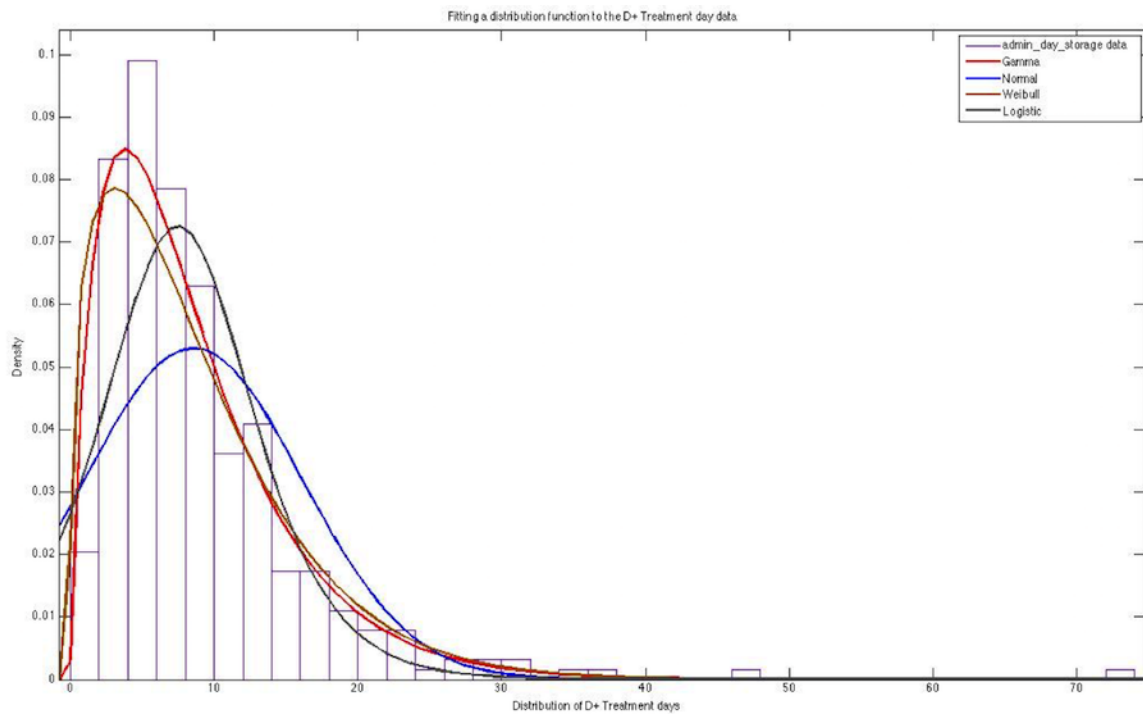


Figure 3-2: Fitting a Gamma curve (in red) to the histogram of treatment patient administration days.

3.2 Discussion

In our results, we see evidence that better balance is achieved when there is more randomness in the sampling scheme. If we think of each technique as a step up or step down in randomness in comparison to its sibling techniques, we can develop an intuitive understanding for why some schemes produce better balances than others. In addition we discuss some reasons why we may be seeing these results, interpreted for the medical context and the complexities of treatment assignment in the clinic in general.

Before we get to sampling lets first consider our selection techniques. For DTW $k=10,20$ and 100 , we see penalties trend downward with decreases in lag (we wish to minimize the penalty). With smaller lags, we are less strict with the fit that we require, allowing patients to be slightly less similar to each other. Being less strict with the patient criteria increases the balance, which can be interpreted as an improvement in selecting comparable patients across risk levels.

This selection pattern actually points to the fact that there may be some real differences picked up by the model between the D+ and D- patients, but unfortunately the covariates we would need are probably outside our dataset. If this is the case, then we might see the same pattern in the inability to achieve balanced quintiles because the D+ and D- patients are substantively different in a component which has a pervasive but unclear signal in the dataset i.e. we dont have these other missing covariates to include in the model to achieve the balance we want.

Also in the DTW results, we see that for $k=10, 20$ and 100 , increases in k generally lead to lower balance scores across all lag levels. This supports the prior interpretation that more randomness/less strict selection criteria leads to better balance because strict selection is finding some real differences that we arent able to account for in our model. With stricter selection, we find that our D- patient likelihoods of treatment are distributed too differently from D+ patient likelihoods to compare in a straightforward way.

Yet another result that supports this interpretation that we are missing a covari-

ate/randomness performs better is the fact that adding Gamma sampling to pick the administration day shifts the otherwise-predictable DTW procedures imbalance down dramatically. Both selection and sampling procedures are ordered according to this trend and improving randomness leads to consistently superior results compared to other strategies.

The one big exception in the results is for $k=1$. With $k=1$, we see a low balance score, with seems to shift less in response to changes in lag length. $K=1$ may perform differently from the other DTW settings because it is deterministic given a set of D+ patients — for each patient, the 1 D- patient in the set of sliding window samples that has the minimum error between the D+ and D- is chosen. We could suggest that choosing the best selection outperforms a random choice from a set of k good samples, but there is another possible explanation: given that randomness seems to give better balance than methods which more clearly differentiate the D+ from D- patients, and as seen in the other DTW methods, higher k (more randomness) gives better results, perhaps what we are seeing here is overfitting to the point of being very random. The fact that a given sample produces the closest match is just fitting to noise in the data — which gives us lower scores in similar vein with how the Gamma distribution sampling gives better balance results than the Patient-day bucket (we want to be random enough to fit the larger D+ population over a number of bootstrapped runs, not exactly fit the sample of D+ patients each and every time).

Comparing sampling techniques, we see that Random selection with Gamma sampling is vastly better than both Patient-day-bucket and Median sampling (and all non-Gamma DTW runs). In the randomness-is-better worldview, this makes sense since the Gamma distribution fit is a completely random selection of D- patients, which are then sampled on days determined by the distribution after it had been fit to the D+ administration days. It is only surpassed by another Gamma sampling procedure in our test (FSBS DTW with $k=1$). The patient-day bucket method is similar to Gamma, but may be overfitting to the exact sample of D- patients administration days since it enforces an exact match day-for-day, while the Gamma performs better because it attempts to model the larger population of D+ patients rather than exactly

match the sample. The same may be true of the Median sampling technique, which has the worst balance of any method attempted in this study. Perhaps the variance in the population sampling days is quite high, so then rigidly enforcing the median is a very poor way to mimic realistic sampling.

Our results suggest that the sampling day choice is very significant for achieving balance. The Median is the worst-fit for sampling day and shows the worst results overall across all methods. Meanwhile the Gamma fit ought to be among the best for sampling the administration day and we see that it shows the best overall, for both FSBS DTW and for Random selection. The DTW/Patient-day and Random/Patient-day bucket runs both fit to the sample of D+ patients observation-for-observation and show middling results, while the same DTW selection with Gamma sampling shows excellent performance for both k values tested. The ordering of the results follows the ordering of sampling techniques, and switching the selection technique gives a smaller magnitude change in performance compared with switching the sampling technique.

What do these results mean in a medical context? In the clinic, doctors assign treatments in psuedo-random ways. Evidence shows that simply varying the care provider may significantly influence treatment decisions, timing, and outcomes after controlling for confounding factors. [10] A previous study also concluded that physician's differing knowledge and opinions may have significant outcome effects, and that the physician should be modeled as part of the overall problem. [6] Given that two different doctors may see the same patient, but differ in their diagnosis or administration of diuretics, its not surprising that we find that we are able to model the results best when we select in more random ways. Said another way, the patient may be given or not given the treatment based upon the random assignment of doctor, so the physician ID is a crucial variable we are missing. In addition, the same doctor may change their opinions and decisions overtime as they see noteworthy cases, in some sense overfitting to their own sample, but only affecting the treatment of subsequent patients.

It is telling that the results seem to be ordered based upon how the patients administration days are sampled, and the greater effect of sampling compared to

selection on imbalance. The strong influence of sampling day is very likely related to the circumstances of our study setting: patients are admitted to the ICU for any number of reasons, and may have developed sepsis while already in the ICU. Many of the patients in our study have significant pre-conditions which caused them to be admitted to the ICU (sepsis may be a secondary diagnosis) and these conditions are both serious enough to put them into the ICU and only crudely approximated by covariates like the Elixhauser comorbidity score. More randomness in the sampling procedure is a better approximation for how diuretics are actually administered in the hospital: somewhat randomly, based upon preconditions, assignment to physicians, and inconsistent evaluation criteria.

3.3 Experimental Results

3.3.1 Random Selection, Median Sampling

Median Day selection for 100 loops

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	4.3400	84.600	92.1200	80.6700	56.2700	318
1	123.6600	42.4000	34.8800	46.3300	70.7300	318
Totals	128	127	127	127	127	636
Balance Score	42.770					

Table 3.2: Quintiles - Random Selection, Median Sampling

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	NaN	7.1388	8.1089	8.9593	9.4173
1	17.5891	11.2570	15.2871	13.9724	14.5147
Mean	NaN	9.1979	11.698	11.4659	11.966

Table 3.3: LOS - Random Selection, Median Sampling

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	NaN	0.2589	0.3066	0.3727	0.4197
1	0.3241	0.3380	0.3940	0.3810	0.3983
Mean	NaN	0.2985	0.3503	0.3769	0.409

Table 3.4: Mortality - Random Selection, Median Sampling

3.3.2 Random Selection, Patient-Day Sampling

Patient Day selection for 100 loops

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	19.6400	93.4500	83.2600	67.9200	53.7300	318
1	108.3600	33.5500	43.7400	59.0800	73.2700	318
Totals	128	127	127	127	127	636
Balance Score	32.5034					

Table 3.5: Quintiles - Random Selection, Patient Day Sampling

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	9.3591	7.5019	8.6239	9.0622	9.8048
1	16.3745	12.5455	14.7165	15.2783	15.2927
Mean	12.8668	10.0237	11.6702	12.1703	12.5488

Table 3.6: LOS - Random Selection, Patient Day Sampling

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4183	0.2509	0.2960	0.3474	0.3677
1	0.3664	0.2229	0.2802	0.3914	0.4287
Mean	0.3924	0.2369	0.2881	0.3694	0.3982

Table 3.7: Mortality - Random Selection, Patient Day Sampling

3.3.3 DTW Sliding KNN Selection, Patient-Day Sampling $k = 1$

DTW Sliding KNN Selection for 100 loops (lag=5, k=1)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	27.00	80.50	56.50	38.00	29.00	231
1	74.50	19.00	43.00	61.50	70.50	268.5
Totals	101.5	99.5	99.5	99.5	99.5	499.5
Balance Score	32.885					

Table 3.8: Quintiles - DTW KNN (lag=5, k=1)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	11.487	5.6297	11.3739	11.3684	12.6223
1	15.7675	11.5833	12.1405	15.7879	15.6023
Mean	13.6273	8.6065	11.7572	13.5782	14.1123

Table 3.9: LOS - DTW KNN (lag=5, k=1)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.2771	0.0372	0.1795	0.2500	0.3359
1	0.4549	0.2361	0.2771	0.3414	0.4231
Mean	0.366	0.1367	0.2283	0.2957	0.3795

Table 3.10: Mortality - DTW KNN (lag=5, k=1)

DTW Sliding KNN Selection for 100 loops (lag=4, k=1)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	19.5	78.00	62.50	42.50	33.00	235.5
1	82.00	21.50	37.00	57.00	66.50	264
Totals	101.5	99.5	99.5	99.5	99.5	499.5
Balance Score	33.7424					

Table 3.11: Quintiles - DTW KNN (lag=4, k=1)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	9.6395	9.0314	9.6414	11.0557	10.9081
1	14.4631	14.7412	14.7838	15.4259	15.0791
Mean	12.0513	11.8863	12.2126	13.2408	12.9936

Table 3.12: LOS - DTW KNN (lag=4, k=1)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.3842	0.1094	0.1758	0.4008	0.3943
1	0.4571	0.1414	0.3243	0.3296	0.4500
Mean	0.4206	0.1254	0.2501	0.3652	0.42215

Table 3.13: Mortality - DTW KNN (lag=4, k=1)

DTW Sliding KNN Selection for 100 loops (lag=3, k=1)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	27.50	78.50	57.50	32.50	34.50	230.5
1	74.50	20.00	41.00	66.00	64.00	265.5
Totals	102.0	98.5	98.5	98.5	98.5	496.0
Balance Score	33.9095					

Table 3.14: Quintiles - DTW KNN (lag=3, k=1)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	7.9033	8.3818	8.5537	12.3238	10.2786
1	14.5570	11.0750	15.0854	15.2785	15.2074
Mean	11.2302	9.7284	11.8196	13.8012	12.743

Table 3.15: LOS - DTW KNN (lag=3, k=1)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.2767	0.1337	0.2433	0.2762	0.3777
1	0.4568	0.3000	0.2439	0.3125	0.4847
Mean	0.36675	0.21685	0.2436	0.2944	0.4312

Table 3.16: Mortality - DTW KNN (lag=3, k=1)

3.3.4 DTW Sliding KNN Selection, Patient-Day Sampling k = 10

DTW Sliding KNN Selection for 100 loops (lag=5, k=10)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	25.2200	90.5300	77.9400	63.6300	40.0700	297.39
1	87.0100	19.7600	32.3500	46.6600	70.2200	256
Totals	112.123	110.29	110.29	110.29	110.29	553.39
Balance Score	39.9688					

Table 3.17: Quintiles - DTW KNN (lag=5, k=10)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	8.0482	6.6644	6.2519	7.1738	8.1898
1	15.7117	14.5676	12.8737	14.9575	14.8806
Mean	11.880	10.616	9.5628	11.0657	11.5352

Table 3.18: LOS - DTW KNN (lag=5, k=10)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4396	0.1180	0.1730	0.2543	0.3245
1	0.4631	0.2280	0.2613	0.4008	0.3587
Mean	0.4514	0.173	0.2171	0.3276	0.3416

Table 3.19: Mortality - DTW KNN (lag=5, k=10)

DTW Sliding KNN Selection for 100 loops (lag=4, k=10)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	26.1200	90.4900	77.4700	63.8500	39.4600	297.39
1	86.1900	19.7800	32.8000	46.4200	70.8100	256
Totals	112.31	110.27	110.27	110.27	110.27	553.39
Balance Score	40.0994					

Table 3.20: Quintiles - DTW KNN (lag=4, k=10)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	7.9097	6.7577	6.1272	7.1692	8.4025
1	15.6664	15.1871	12.7955	14.6762	14.9731
Mean	11.7881	10.9724	9.4614	10.9227	11.6878

Table 3.21: LOS - DTW KNN (lag=4, k=10)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4235	0.1162	0.1808	0.2508	0.3334
1	0.4666	0.2081	0.2658	0.3970	0.3606
Mean	0.4451	0.1622	0.2233	0.3239	0.347

Table 3.22: Mortality - DTW KNN (lag=4, k=10)

DTW Sliding KNN Selection for 100 loops (lag=3, k=10)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	27.2500	87.3700	75.8000	62.4900	40.8400	297.39
1	84.1400	22.2200	33.7900	47.1000	68.7500	256
Totals	111.39	109.59	109.59	109.59	109.59	549.75
Balance Score	36.7333					

Table 3.23: Quintiles - DTW KNN (lag=3, k=10)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	7.6368	6.5729	6.7216	7.9647	8.5307
1	16.0547	13.8960	13.1424	14.0822	15.1833
Mean	11.8458	10.2345	9.932	11.0235	11.857

Table 3.24: LOS - DTW KNN (lag=3, k=10)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4393	0.0918	0.1590	0.2564	0.3238
1	0.4861	0.1624	0.2633	0.3984	0.3610
Mean	0.4627	0.1271	0.2112	0.3274	0.3424

Table 3.25: Mortality - DTW KNN (lag=3, k=10)

3.3.5 DTW Sliding KNN Selection, Patient-Day Sampling k = 20

DTW Sliding KNN Selection for 100 loops (lag=5, k=20)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	25.44	87.43	78.63	66.37	40.65	298.52
1	87.00	23.09	31.89	44.15	69.87	256.0
Totals	112.44	110.52	110.52	110.52	110.52	554.52
Balance Score	41.3027					

Table 3.26: Quintiles - DTW KNN (lag=5, k=20)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	7.4793	7.5384	6.4885	7.3945	8.1202
1	15.2045	16.2707	12.5663	15.4497	14.7366
Mean	11.3419	11.9046	9.5274	11.4221	11.4284

Table 3.27: LOS - DTW KNN (lag=5, k=20)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4235	0.1316	0.1532	0.2038	0.3041
1	0.4905	0.2101	0.2865	0.3899	0.3320
Mean	0.4570	0.1709	0.2199	0.2969	0.3181

Table 3.28: Mortality - DTW KNN (lag=5, k=20)

DTW Sliding KNN Selection for 100 loops (lag=4, k=20)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	26.53	86.44	76.81	66.16	42.80	298.74
1	86.13	24.08	33.71	44.36	67.72	256.00
Totals	112.66	110.52	110.52	110.52	110.52	554.74
Balance Score	38.7192					

Table 3.29: Quintiles - DTW KNN (lag=4, k=20)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	7.2046	7.4306	7.0458	7.8608	8.3544
1	15.4672	14.8185	13.1088	15.5117	14.6441
Mean	11.3359	11.1246	10.0773	11.6863	11.4993

Table 3.30: LOS - DTW KNN (lag=4, k=20)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4149	0.1310	0.1635	0.2169	0.2850
1	0.4962	0.2063	0.3052	0.3872	0.3224
Mean	0.4556	0.16865	0.23435	0.30205	0.3037

Table 3.31: Mortality - DTW KNN (lag=4, k=20)

DTW Sliding KNN Selection for 100 loops (lag=3, k=20)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	28.50	85.57	76.91	63.22	42.85	297.05
1	83.67	24.65	33.31	47.00	67.37	256.0
Totals	112.17	110.22	110.22	110.22	110.22	553.05
Balance Score	35.7359					

Table 3.32: Quintiles - DTW KNN (lag=3, k=20)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	6.7148	7.3655	7.1231	7.9791	8.5633
1	15.2830	14.1629	13.5118	15.00	15.2797
Mean	10.9989	10.7642	10.3175	11.4896	11.9215

Table 3.33: LOS - DTW KNN (lag=3, k=20)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4270	0.1180	0.1542	0.2374	0.2912
1	0.4895	0.1827	0.2956	0.4025	0.3387
Mean	0.45825	0.1503	0.2249	0.31995	0.31495

Table 3.34: Mortality - DTW KNN (lag=3, k=20)

3.3.6 DTW Sliding KNN Selection, Patient-Day Sampling k = 100

DTW Sliding KNN Selection for 100 loops (lag=5, k=100)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	22.22	91.25	74.42	65.05	42.12	295.06
1	89.44	18.60	35.43	44.80	67.73	256.0
Totals	111.66	109.85	109.85	109.85	109.85	551.06
Balance Score	39.7104					

Table 3.35: Quintiles - DTW KNN (lag=5, k=100)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	7.1924	12.2795	9.6569	8.2640	8.0489
1	15.4171	18.7420	14.2674	12.7498	14.9097
Mean	11.30475	15.5108	11.96215	10.5069	11.4793

Table 3.36: LOS - DTW KNN (lag=5, k=100)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4137	0.1182	0.2128	0.2461	0.2904
1	0.4835	0.2016	0.2016	0.3632	0.3939
Mean	0.4486	0.1599	0.2072	0.30465	0.34215

Table 3.37: Mortality - DTW KNN (lag=5, k=100)

DTW Sliding KNN Selection for 100 loops (lag=4, k=100)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	24.67	89.10	73.20	64.17	42.32	293.46
1	86.83	20.39	36.29	45.32	67.17	256.0
Totals	111.5	109.49	109.49	109.49	109.49	549.46
Balance Score	37.4682					

Table 3.38: Quintiles - DTW KNN (lag=4, k=100)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	7.3111	12.3164	9.3260	8.2009	7.9107
1	15.3751	18.2712	14.2527	12.8592	14.9620
Mean	11.3431	15.2938	11.78935	10.5301	11.4364

Table 3.39: LOS - DTW KNN (lag=4, k=100)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4167	0.1476	0.2158	0.2507	0.2756
1	0.4856	0.2006	0.2004	0.3717	0.3964
Mean	0.45115	0.1741	0.2081	0.3112	0.3360

Table 3.40: Mortality - DTW KNN (lag=4, k=100)

DTW Sliding KNN Selection for 100 loops (lag=3, k=100)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	27.82	86.72	72.71	64.09	42.73	294.07
1	83.77	22.90	36.91	45.53	66.89	256.0
Totals	111.59	109.62	109.62	109.62	109.62	550.07
Balance Score	35.6175					

Table 3.41: Quintiles - DTW KNN (lag=3, k=100)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	7.4824	11.6767	8.9106	8.1565	7.8589
1	15.5291	16.2324	14.4464	12.8036	15.2996
Mean	11.5057	13.95455	11.6785	10.4800	11.57925

Table 3.42: LOS - DTW KNN (lag=3, k=100)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.4280	0.1355	0.2275	0.2506	0.3125
1	0.4883	0.1684	0.2360	0.3795	0.3939
Mean	0.45815	0.15195	0.23175	0.31505	0.3532

Table 3.43: Mortality - DTW KNN (lag=3, k=100)

3.3.7 Random Selection, Gamma Distribution Sampling

Gamma Distribution Sampling for 100 loops

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	69.72	88.14	73.37	52.26	33.79	317.28
1	58.2	38.7	53.47	74.58	93.05	318.0
Totals	127.92	126.84	126.84	126.84	126.84	635.28
Balance Score	27.2375					

Table 3.44: Quintiles - Random Selection, Gamma Distribution Sampling

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	9.5445	8.4369	9.4883	8.9087	8.5752
1	16.7992	13.8703	14.5682	15.0555	15.5426
Mean	13.1719	11.1536	12.02825	11.9821	12.0589

Table 3.45: LOS - Random Selection, Gamma Distribution Sampling

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.3370	0.2708	0.2973	0.3407	0.3797
1	0.2807	0.2409	0.3113	0.4120	0.4377
Mean	0.30885	0.25585	0.3043	0.3763	0.4087

Table 3.46: Mortality - Random Selection, Gamma Distribution Sampling

3.3.8 DTW Sliding KNN Selection, Gamma Distribution Sampling $k = 20$

DTW Sliding KNN Selection, Gamma Distribution Sampling for 100 loops (lag=5, k=20)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	30.74	81.39	70.34	55.94	42.09	280.50
1	76.24	23.66	34.71	49.11	62.96	246.68
Totals	106.98	105.05	105.05	105.05	105.05	527.18
Balance Score	26.6094					

Table 3.47: Quintiles - DTW KNN Gamma (lag=5, k=20)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	14.0633	11.0149	12.1528	15.4714	18.0658
1	15.0712	12.9009	15.0189	15.5030	15.4294
Mean	14.5673	11.9579	13.5859	15.4872	16.7476

Table 3.48: LOS - DTW KNN Gamma (lag=5, k=20)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.3400	0.1364	0.1668	0.2364	0.2671
1	0.4664	0.2237	0.2869	0.3206	0.3961
Mean	0.4032	0.1801	0.2269	0.2785	0.3316

Table 3.49: Mortality - DTW KNN Gamma (lag=5, k=20)

DTW Sliding KNN Selection, Gamma Distribution Sampling for 100 loops (lag=4, k=20)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	30.26	80.75	70.49	58.86	41.73	282.09
1	77.33	24.71	34.97	46.60	63.73	247.34
Totals	107.59	105.46	105.46	105.46	105.46	529.43
Balance Score	30.2361					

Table 3.50: Quintiles - DTW KNN Gamma (lag=4, k=20)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	13.4753	11.1939	12.5986	15.8599	16.7457
1	15.0578	12.3544	15.7088	14.9584	15.7092
Mean	14.2666	11.7742	14.1537	15.4092	16.2275

Table 3.51: LOS - DTW KNN Gamma (lag=4, k=20)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.3424	0.1392	0.1769	0.2273	0.2689
1	0.4702	0.2034	0.2867	0.3377	0.3841
Mean	0.4063	0.1713	0.2318	0.2825	0.3265

Table 3.52: Mortality - DTW KNN Gamma (lag=4, k=20)

DTW Sliding KNN Selection, Gamma Distribution Sampling for 100 loops (lag=3, k=20)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	29.63	81.23	69.82	58.81	41.42	280.91
1	77.99	24.16	35.57	46.58	63.97	248.27
Totals	107.62	105.39	105.39	105.39	105.39	529.18
Balance Score	30.4246					

Table 3.53: Quintiles - DTW KNN Gamma (lag=3, k=20)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	13.0138	10.9366	13.1595	16.6766	16.7075
1	15.2058	11.6706	15.5358	15.0552	15.6676
Mean	14.1098	11.3036	14.3477	15.8659	16.1876

Table 3.54: LOS - DTW KNN Gamma (lag=3, k=20)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.3715	0.1473	0.1733	0.2319	0.2736
1	0.4678	0.2084	0.2801	0.3467	0.3872
Mean	0.4197	0.1779	0.2267	0.2893	0.3304

Table 3.55: Mortality - DTW KNN Gamma (lag=3, k=20)

3.3.9 DTW Sliding KNN Selection, Gamma Distribution Sampling $k = 1$

DTW Sliding KNN Selection, Gamma Distribution Sampling for 100 loops (lag=5, k=1)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	14.68	56.42	44.49	32.69	23.73	172.01
1	62.37	18.76	30.69	42.49	51.45	205.76
Totals	77.05	75.18	75.18	75.18	75.18	377.77
Balance Score	23.2082					

Table 3.56: Quintiles - DTW KNN Gamma (lag=5, k=1)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	12.0145	9.1422	8.9191	10.0417	10.9814
1	15.4009	13.2308	14.6591	15.7964	15.7827
Mean	13.7077	11.1865	11.7891	12.9191	13.3821

Table 3.57: LOS - DTW KNN Gamma (lag=5, k=1)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.3595	0.1530	0.1933	0.2764	0.3122
1	0.4212	0.2290	0.2733	0.3279	0.3979
Mean	0.3903	0.1910	0.2333	0.3022	0.3550

Table 3.58: Mortality - DTW KNN Gamma (lag=5, k=1)

DTW Sliding KNN Selection, Gamma Distribution Sampling for 100 loops (lag=4, k=1)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	16.48	56.47	44.02	33.70	24.03	174.70
1	61.42	19.64	32.09	42.41	52.08	207.64
Totals	77.90	76.11	76.11	76.11	76.11	382.34
Balance Score	21.7113					

Table 3.59: Quintiles - DTW KNN Gamma (lag=4, k=1)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	11.6509	9.5520	8.6210	10.3951	10.5056
1	15.8760	13.7004	14.5218	15.4692	16.1794
Mean	14.2	11.7	14.1	15.4	16.2

Table 3.60: LOS - DTW KNN Gamma (lag=4, k=1)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.3439	0.1502	0.1916	0.2703	0.3233
1	0.4343	0.2095	0.2544	0.3237	0.3845
Mean	0.4	0.1	0.2	0.2	0.3

Table 3.61: Mortality - DTW KNN Gamma (lag=4, k=1)

DTW Sliding KNN Selection, Gamma Distribution Sampling for 100 loops (lag=3, k=1)

Treatment	Q1	Q2	Q3	Q4	Q5	Totals
0	15.83	54.58	42.65	33.61	24.12	170.79
1	60.86	20.05	31.98	41.02	50.51	204.42
Totals	76.69	74.63	74.63	74.63	74.63	375.21
Balance Score	20.0551					

Table 3.62: Quintiles - DTW KNN Gamma (lag=3, k=1)

Treatment	LOS Q1	LOS Q2	LOS Q3	LOS Q4	LOS Q5
0	11.7924	10.3545	8.9622	10.1732	10.8221
1	15.2749	13.6663	14.9963	15.4726	15.7023
Mean	13.5337	12.0104	11.9793	12.8229	13.2622

Table 3.63: LOS - DTW KNN Gamma (lag=3, k=1)

Treatment	MOR Q1	MOR Q2	MOR Q3	MOR Q4	MOR Q5
0	0.3627	0.1597	0.1978	0.2649	0.3136
1	0.4232	0.2132	0.2582	0.3150	0.3846
Mean	0.3930	0.1865	0.2280	0.2900	0.3491

Table 3.64: Mortality - DTW KNN Gamma (lag=3, k=1)

Chapter 4

Research Findings and Contributions

4.1 Research Findings and Contributions

This thesis demonstrates the following methods for clinical time-series data:

- An effective representation for standardizing event-driven and irregular data using smoothing on days, taking mean and median for each variable, for each day. This makes previously incomparable observations comparable despite different timing and events.
- Development of two selection and three sampling techniques for preparing retrospective studies on irregular time-series data, and software for performing bootstrap iterations for combinations of techniques. We found that using a Fluid-Balance Similarity-Based Dynamic Time Warp selection procedure with nearest neighbor parameter $k=1$ and using a Gamma distribution for sampling days produced consistently better class balance than all other methods when bootstrapped over 100 independent runs.
- Development of the Class Imbalance Penalty, a class balance metric for selecting quintiles for statistically sound comparisons, which enables relative ranking of

stratified matching procedures. We have demonstrated that it gives low scores for the procedures that produce the most class-balanced quintile results for all strata and lend themselves to statistically sound future analysis.

From our retrospective study on Diuretics treatment effects for Sepsis patients, we demonstrate the following findings, which hold for our study's context:

- Analysis and findings showing that sampling had a stronger effect on class balance than selection, and using a Gamma distribution fit to pick sampling days produced more statistically comparable quintiles than all other methods, for any selection procedure.
- Analysis and findings showing that Fluid-Balance Similarity-Based DTW selection with $k=1$ and $\text{lag}=3$ in combination with Gamma distribution sampling is the best combination among the tested methods, but purely Random sampling outperforms it for many values of k .
- Analysis and findings strongly suggesting that physician and provider data are key missing covariates for predicting Diuretics administration in the ICU

4.2 Software

- We have written, documented and published open source MATLAB code at [github.mit.edu/ALFAGroup/Clinical_Time_Series_For_Diuretics_brian_bell](https://github.com/mit-alfagroup/Clinical_Time_Series_For_Diuretics_brian_bell) for each selection and sampling technique, along with our bootstrap test.

4.3 Future work

- We strongly suspect that the addition of physician-related covariates would improve the balance seen for our characteristic problem. We should confirm that the relative performance of the techniques tested still holds.
- Re-implement the experimental procedure code in Python to better architect the system and improve readability and modularity.

- Try more flexible modeling approaches, including non-linear techniques such as Decision Trees, which may be better able to accurately score, and by extension produce better balance across quintiles. [12]

Chapter 5

Glossary

Balanced Quintiles

The desired end-goal state of Rosenbaum and Rubins Propensity Score Matching Procedure. During Propensity Score Matching, a statistical model is trained upon a set of covariates that may include confounding variables, and a treatment/non-treatment variable as the response. The model's prediction from 0-1 is an approximate measure of the likelihood of treatment.

Cohort

The cohort is a sample of larger Study Group, considered for a single experiment. It consists of approximately equal numbers of treatment and non-treatment patients.

Cohort Selection

We call the act of picking representative non-treatment observations from the Study Group to match a set of treatment observations Cohort Selection.

Diuretics

A class of Medical drugs given to treat Sepsis patients. Typically given via IV, they cause patients to release excess fluids, helping them reduce systemic inflammation.

Length of Stay

Given that a patient has been in the ICU for a least 24 hours (a criteria to be in the study group), Length of Stay expresses the number of days the patient has spent in the ICU, expressed as a ceiling. So a patient who has been in the ICU for 30 hours has a Length-of-Stay of 2 days.

Logistic Regression

A modeling technique proposed by Cox in 1958, used primarily as an alternative to early techniques such as linear regression or additive models when predicting a categorical response, especially a binary one. [2]Because the logistic function resembles a step function, it mainly produces values toward the endpoints of the range from 0.0-1.0 and so it makes for a natural fit for binary classification problems.

Propensity Score Matching

Statistical technique used to analyze a particular intervention by accounting for bias due to confounding variables. In particular, it helps address situations where the outcome may actually be linked to the characteristics that influenced the treatment decision, when in fact we want to isolate the effect of the treatment. By mimicking randomized assignment by creating cohorts that are comparable on all observed covariates, we can address concerns of systemic bias.

Sepsis

Medical condition described as severe full body inflammation in response to a serious infection. Patients diagnosed with sepsis suffer from a long list of symptoms: very low blood pressure and heart racing, swelling, flushing, fever and hyperventilation.

Study Group

The set of all Sepsis patients gathered from MIMIC with at least 24 hours stay in the ICU.

Thirty Day Mortality

Given a set of patients who have been in the ICU for at least 24 hours and subsequently released, 30-day mortality is the population mean for the proportion of the original set of patients who die within 30 days of their release from the hospital. It is a common quality-of-care metric, with better healthcare centers showing lower rates.

Time Series

A genre of modeling problems involving covariates that change upon one or more dimensions, including time. Examples of this kind of signal abound in the medical field, such as a log of a patients heart rate each minute, accumulated across multiple days. Modeling this data can be challenging because it must often incorporate data from different time scales (mixing daily data with hourly data for example), non time-bound data, and also the influence of previous days information in each time-varying signal. While other kinds of modeling problems can be tackled with a lenient eye toward concerns for the independence of each observation from the other observations, time-series data is often very strongly dependent on prior observations.

Bibliography

- [1] Peter C. Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27, 2008.
- [2] D. R. Cox. The analysis of binary data. 1970.
- [3] Greg S. Martin M.D. Marc Moss M.D. David M. Mannino M.D., Stephanie Eaton M.D. The epidemiology of sepsis in the united states from 1979 through 2000. *The New England Journal Of Medicine*, 2003.
- [4] Harris DR Coffey RM Elixhauser A, Steiner C. Comorbidity measures for use with administrative data. *Med Care*, 1998.
- [5] Daniel J. Scott Mauricio Villarroel, Gari D. Clifford. *User guide and documentation for the MIMIC II database*. 2012.
- [6] Erik Hemberg Miguel Paredes and Una-May O'Reilly. On the challenges of using propensity score matching to study intensive care unit patients. *Conference on Neural Information Processing Systems (NIPS) Machine Learning in Healthcare Workshop*, 2016.
- [7] Donald B. Rubin Paul R. Rosenbaum. Reducing bias in observational studies using sub-classification on the propensity score. *Journal of the American Statistical Association*, 1984.
- [8] D. Rammazotti. The effect of diuretics administration on outcomes of mortality and mean duration of i.c.u. stay (masters thesis). 2012.
- [9] Fabienne Saulnier M.D. Stanley Lemeshow PhD, Jean-Roger Le Gall M.D. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Journal of American Medical Association*, 1993.
- [10] Kerr E. A. Klamerus M. Zikmund-Fisher B. J. Holleman R. Hofer T. P. Subramanian, U. Treatment decisions for complex patients: differences between primary care physicians and mid-level providers. *The American Journal of Managed Care*, 15(6):373–380, 2009.
- [11] De Mendona A Bruining H Reinhart CK Suter PM Thijs LG Vincent JL Moreno R Takala J, Willatts S. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Med*, 22:707–710, 1996.

- [12] Lessler J. Funk M. J. Westreich, D. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833, 2010.