A Sequence Based Approach for Predicting Clinical Events

Anam Zahid

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Computer Science and Systems

University of Washington

2016

Reading Committee:
Martine De Cock
Ankur Teredesai
Shanu Sushmita

Program Authorized to Offer Degree to:
Institute of Technology - Tacoma

University of Washington

Abstract

A Sequence Based Approach for
Predicting Clinical Events

Anam Zahid

Chair of the Supervisory Committee:
Professor Martine De Cock
Institute of Technology

The data associated to each patient increases almost linearly as the patient flows through the continuum of care. Analysis of the data collected during a patient's admission to the hospital reveals that it grows vertically as well as horizontally as a variety of readings are taken for the patient. In general, machine learning techniques are designed and evaluated to predict clinical events at one particular time point during this process (on admission to the hospital, or on discharge). This highlights one of the key challenges of making predictive solutions applicable to the real world setting, as it limits the interventions that can be taken *while* the patient is at the hospital, to avoid undesirable clinical outcomes down the road. To address this challenge, we have proposed a novel framework of at-admit and sequence based models that predict clinical outcomes accurately at different time points of a patient's hospital stay and perform consistently better than a retrospectively designed solution.

Hospitalizations account for about half of all healthcare expenses, and it has been estimated that 13% of the inpatients in the United States use more than half of all hospital resources through repeated admissions [4]. Therefore, the clinical outcome chosen for this work is predicting thirty day readmissions for the "all cause" population. We compare our proposed approach to the state of the art readmission modeling approach of retrospective feature creation, and see an average improvement of 7% in the area under the curve as well as significant improvements in precision, accuracy and recall.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to extend my sincere gratitude to everyone who has helped me during this thesis and throughout the Masters program. I am extremely grateful for their invaluable support, guidance and feedback throughout this journey.

My deepest appreciation belongs to the members of my advisory committee and particularly Professor Martine De Cock from whom I have learnt a tremendous amount of things.

I want to thank my colleagues at KenSci without whom this work would not have been possible. For helping me make the best out of the program and giving me the opportunity to learn through such variety of valuable real world experiences.

I am extremely grateful to my family and friends for their continuous support, motivation and positivity.

Thank you,
Anam Zahid

Dedication

To my dearest Zoya, who has always been my happy place. Never stop being the curious, full of life and thoughtful person you are. You will always have a huggie monster waiting for you.

Chapter 1

Introduction

The clinical data available for a person increases almost linearly over the course of a person's lifespan. With the introduction of new IoT (internet of things) devices within and outside of hospitals, there is a rising opportunity as well as a challenge in the healthcare domain to capture the increasing data meaningfully [15]. The data grows not only vertically but also horizontally for example when the patient is in the hospital more readings for the same lab tests are taken (vertical expansion) as well new lab tests are taken (horizontal expansion). This increase of data requires predictive models to be more dynamic in feature space rather than static feature based individual models.

One key challenge in the healthcare domain is the gap between academic solutions developed everyday and their readiness to be used by clinicians in the real world setting. We have seen an incredible improvement in results as more and more sophisticated methods are developed [6]. The question is if those methods are designed and evaluated on how they will perform if used in a real clinical setting and whether the performance metrics can hold up as reported from retrospective evaluation.

The majority of the methods developed for clinical events predictions are designed to predict an event in the future based on a static set of accumulated features till a certain time point. In reality, the state of a patient evolves with every new clinical reading or lab test result, which does not get reflected in such static models. This growing sequence of data for each patient can be captured using sequence based features as used in sequence prediction techniques [18] as inputs to event based predictive models. In this thesis we propose a framework of sequence based models that capture the transitions of a patient's health throughout a hospital stay, as well as an "at-admit" model to predict a clinical outcome based on a fixed history of data available for the patient.

1.1 Use Case: Thirty Day Readmissions

Readmissions to hospitals are a huge problem hurting the quality of care and significantly increasing the costs [1, 12, 10]. Significant work has been done in building solutions to predict which patients are at a higher risk for hospital readmission [9, 11, 2]. This risk assessment could be used to help target the delivery of resource-intensive interventions to the patients at greatest risk [6]. To this end, patients need to be risk stratified while they are admitted in the hospital, to provide clinically relevant interventions in time well before the discharge. Hospital readmissions also have an impact on cost as the Centers for Medicare Medicaid Services (CMS) began using readmission rate as a publicly reported metric, with plans to lower reimbursement to hospitals with excess risk-standardized readmission rates.

Under the Affordable Care Act, the Readmission Reduction Program (HRRP) requires CMS to reduce payments to IPPS hospitals with excess readmissions and one of the ways it is calculated is based on thirty day readmissions [5, 10]. In this work we predict thirty day readmissions as the clinical use case of the proposed approach. This means that it is a binary classification task where we predict whether or not the patient will be readmitted to the hospital within 30 days after discharge. The population is "all cause" meaning that the patient cohort is a "mixed bag" of various conditions. We propose a novel at-admit and sequence based approach to tackle some of the challenges mentioned above and build more dynamic and prospective models to predict clinical outcomes.

1.2 Thesis Overview

We will first discuss in detail the dataset used in the thesis and all the preprocessing steps performed on the data in Chapter 2. We are using Electronic Health Records (EHR) data from 6 facilities of a major U.S. healthcare provider. The data components are explained in Table 2.1. There are a total of approximately 92k unique encounters and 65k unique patients in the raw data. We perform data cleaning on each data component as explained in Section 2.2. After cleaning the data we do feature engineering and select the features to be created which are mentioned in Table 2.2. The data aggregations and transformations are different for each modeling setup i.e at-admit model and sequence based models which are discussed in Section 2.4.

After the data processing, we propose our approach to predict hospital readmissions at different times during the encounter starting with the time of admission prediction model in Chapter 3. This is the at-admit model designed similarly to the state of the art readmission models but it only uses data available for a patient at the time of admission (first admission or readmission). We compare it with the baseline LACE index score and show that it performs significantly better in all metrics of evaluation mentioned in Section 3.4.2.

Next we propose a sequence based modeling approach to predict clinical events like readmissions during the hospital encounter in Chapter 4. We com-

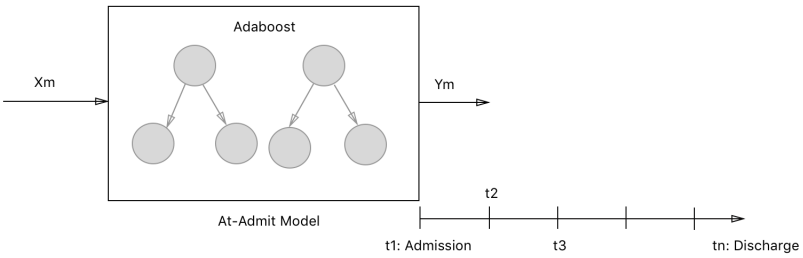

Figure 1.1: At-Admit Modeling Technique

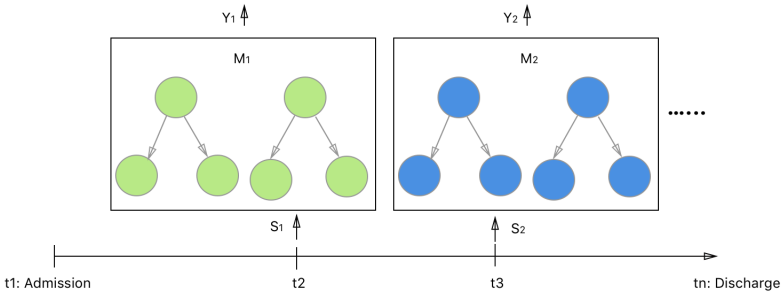

Figure 1.2: Sequence Based Modeling Technique

pare it to a retrospective readmission model baseline and predict using both approaches at each day of a patient's encounter at the hospital. The results show that our sequence features based models outperform the baseline in all metrics.

Finally we show results from the combination of at-admit and sequence based models and offer our observations and ideas for future work in Chapter 5.

Chapter 2

Data Pre-processing

2.1 Data Introduction

In this research we predict a patient's risk of readmission within thirty days of discharge for the all cause population. This work is based on a major healthcare provider's electronic medical records (EMR) from its 6 facilities across the U.S. It contains two years of data from January, 2014 to January, 2016. Just like any other clinical datasets it consists of anonymized patient records divided into various groups of data like diagnoses, lab results, demographics and more. Real world data is generally not in a shape that can be used for modeling and experimentation as is. This chapter highlights all the steps taken to transform and clean the raw data into various input types that are used for training and testing in the experiments explained in Chapters 3, 4 and 5.

2.1.1 Raw Data

In total the data contains approximately 92k unique encounters and 65k unique patients to the hospital. An encounter is defined as an admission to the hospital in the context of this research. The raw data consists of five files representing the components collected in EMR data. Table 2.1 provides the size of the raw files as well as the features created from each.

Data Component	Size	Description
Population	92000 rows	Age, Gender, Marital Status etc
Labs	38.8M rows	17 Clinical Readings (Hemoglobin level, Creatinine level etc.)
Diagnoses	12.5M rows	37 diagnoses (Arrhythmia, Peptic Ulcer, Fluid Disorder etc.)

Table 2.1: Data components and summary of extracted relevant attributes

We took various steps to capture data in the manner that is needed for our

two modeling scenarios:

- At-Admit Modeling (Chapter 3)
- Sequence Modeling (Chapter 4)

These steps include tasks like data cleaning, feature extraction and selection, and data transformations that are described in detail below.

2.2 Data Cleaning

Real world datasets and especially healthcare datasets usually have noise that can make the experimental results unreliable. Across the three data components we use in this research, we take different steps of data cleaning to address this issue.

2.2.1 Population

The population data contains general demographic information for patients as well as the admission and discharge information for each encounter. For cleaning the noise in this file, we remove the data that fulfills any one of the following conditions:

- Rows with bad data in ID fields (facility ID, patient ID, encounter ID)
- Rows with bad data in encounter admission and discharge date fields
- Rows with encounters for which thirty day readmission ground truth cannot be determined i.e. the encounters where the discharge date is later than thirty days before the maximum data extraction date
- Columns with unreliable data which were identified by the data provider administrator (demographics like race and more)

2.2.2 Diagnoses

The diagnoses data contains the ICD-9 and ICD-10 codes associated to a patient's appointments in the facilities. The International Classification of Diseases (ICD) is a system used by physicians and other healthcare providers to classify and code all diagnoses, symptoms and procedures recorded in conjunction with hospital care in the United States. This includes diagnoses that happened during an in-patient as well as out-patient visit. The association between diagnoses and population is made using encounter dates and patient IDs. We remove the data that contains any of the following issues:

- Rows with bad data in ID fields
- Rows with bad data in appointment date fields
- Rows with invalid data in ICD code fields

2.2.3 Labs

This file contains the various lab tests and their results for patients. The association between lab orders and population data occurs in a similar fashion as diagnosis. The ID fields along with the mapping of lab order date and encounter dates is used to extract lab tests taken during an encounter. We use a two step process to clean the lab results data. First the data containing any of the following conditions is removed:

- Rows with bad data in ID fields
- Rows with bad data in lab order field
- Rows with non-numeric data in lab test result field

A second level of cleaning is done to make the lab test names uniform for each lab test that is used in the features. A total of 17 lab tests were selected to be of most importance to the risk of readmission problem, a detail of which is given in Section 2.3 on feature engineering of this chapter. In the data there is a lot of variation in the names used for recording each of these 17 lab tests. For example the lab test 'Glucose' is named in various different ways like 'GLUCOSE,COBAS (19MAR13; upd ur 14OCT15 abs)','GLUCOSE POCT (WRB I-STAT)12SEP2014','GLUCOSE (WRB-COBAS) 05SEPT2014' and many other variations. This noise is due to lack of standardization at the data entry level. This makes it difficult to extract the maximum amount of data for each of the 17 lab tests while not including incorrect lab tests.

We have tried two methods to deal with this variation noise in the lab test names. In the first method the following two steps are used to deduplicate lab test names:

Algorithm

lab tests = $[L_1, L_2, \ldots..L_{17}]$
lab test standards = possible min, max values for each Li
for Li in lab tests **do**
  variations[Li] = Search Li sub-string matching in all lab tests names
  for vi in variations **do**
    if min or max values in vi are not within lab test standards for Li **then**
      Delete vi from variations
    end if
  end for
end for

This method of grouping same lab tests together has two major problems. One is that it requires to know a possible minimum and maximum value for each lab test. The literature and domain expertise allows to infer the normal range of values for each lab test but to be able to come up with a definite possible minimum and maximum for every lab test is not trivial. The reason is that although the values for any lab test may surround a certain range in general,

there still might be some patients that are outliers to that range i.e. a very sick patient. The second problem with this approach is that even if there is a known minimum and maximum value for a lab test, there might be more lab tests with a similar name that also have values within those minimum and maximum values. Due to these identified issues, we use another method to tackle de-duplication of lab test names more robustly.

Algorithm

lab tests = [$L_1, L_2, \ldots..L_{17}$]
for Li in lab tests **do**
  variations[Li] = Search Li sub-string in all lab test names
  standards[Li] = top 3 vi in variations by number of rows and verified by domain expert
  for vi in variations-standards **do**
    variations-ks = Kolmogorov–Smirnov test (vi, standards[Li])
    if p-value of variation-ks is less than 0.5 **then**
      Delete vi from variations
    end if
  end for
end for

In this method the first filter is to get a golden set of lab test names for each of the 17 lab tests by picking the ones with most records (usually these have 70-80% more rows than the rest of the variations). These extracted standard lab tests are also verified by a domain expert to be the same tests as the corresponding lab test needed. After getting these standard validated lab tests, the similarity between their probability distribution and that of other variations is measured using a Kolmogorov–Smirnov test [20]. Any variation below a threshold of p value 0.5 is discarded. Using this method, a valid comparison is possible of the data with the standard drawn out of the same data and thus a good representative of the same population. This also doesn't bound the values to be in a particular range but instead only checks the similarity of distributions between variations and the validated correct labs. This technique gives considerably better results and the variations extracted are renamed in the data to their corresponding standard lab test name.

2.3 Feature Engineering

Feature engineering is the process of transforming raw attributes to meaningful features for the purpose of improving the predictive accuracy of machine learning models built on top of them. These features are created using domain expertise and literature reviews of previous work [14][16]. The features created from each data component from Table 2.1 are described in Table 2.2.

The response variable is a binary flag set for thirty day readmission. For each encounter to the hospital, if the patient got readmitted to the hospital within thirty days of discharge then the thirty day readmission flag is set to

Features	Type	Usage
Age	Integer	At-Admit Modeling
Gender	Boolean	At-Admit Modeling
Prior admission count	Integer	At-Admit Modeling
Emergency visits in last 6 months	Integer	At-Admit Modeling
Comorbidity ever diagnosed for comorbidities in Table 2.4	Boolean	At-Admit Modeling
The last time comorbidity was diagnosed for comorbidities in Table 2.4	Integer	At-Admit Modeling
The number of times comorbidity was diagnosed for comorbidities in Table 2.4	Integer	At-Admit Modeling
Average lab results for labs in Table 2.3	Float	Sequence Modeling
Thirty day Readmission flag	Boolean	At-Admit Modeling, Sequence Modeling

Table 2.2: Summary of features constructed for the machine learning models

1 else it is set to 0. Prior admission count for each encounter of a patient is the number of encounters the patient had at the hospital before the current encounter. Similarly the emergency visits count for each encounter is the number of emergency visits for the patient in the last 6 months before the current encounter.

Each of these attributes is used in the techniques mentioned alongside. These techniques are discussed in detail in the Chapter 3 and 4.

2.3.1 Feature Selection

It is important to decrease the feature space to reduce noise for the predictive model. Some modeling techniques are more sensitive to feature noise than others. As explained in Section 2.2.3, for the selection of relevant labs we incorporated information from a literature review [14] and a domain expert's input to select a set of 17 lab tests. Similarly, for the selection of relevant diagnoses, the most significant chronic conditions/comorbidities are considered for the prediction of readmission. The three features created for each diagnosis are based on a domain expert's feedback on how to best capture the diagnosis history of a patient.

2.4 Data Transformation

After feature creation, we shape the data to bring it into the right input format for each of the modeling techniques from Chapter 3 and 4. The granularity

No.	Lab tests	Top 2 Name Variations
1.	Glucose	1) GLUCOSE POCT (WRB I-STAT)12SEP2014 2) GLUCOSE (WRB-COBAS) 05SEPT2014
2.	Hematocrit	1) HCT,DXH (27MAR14) 2) HCT (WRB/QMAIN) 7DEC2014
3.	Hemoglobin	1) HGB,DXH (27MAR14) 2) HGBMAMC6AUG2007
4.	White Blood Cells	1) WBC,DXH (27MAR14) 2) WBC COUNT (WRB/QMAIN) 29JULY2013
5.	Blood Urea Nitrogen	1) UREA NITROGEN,COBAS (4APR13;upd 14OCT15 abs) 2) UREA NITROGEN (C8000)
6.	Sodium	1) SODIUM,COBAS (4APR13; upd ur 14OCT15) 2) SODIUM (WRB-COBAS) 16AUG2014
7.	Creatinine	1) CREATININE IDMS,COBAS 2) CREATININE (WRB-COBAS) 19JUNE2012
8.	Calcium	1) CALCIUM,COBAS (4APR13;upd ur5AUG15;s/p14OCT15 abs) 2) CALCIUM,COBAS (25FEB16)
9.	Protein	1) PROTEIN TOTAL,COBAS (13MAR13; upd 14OCT15 abs) 2) COBAS PROTEIN TOTAL
10.	Potassium	1) POTASSIUM,COBAS (19APR13) 2) POTASSIUM,COBAS (11MAR16)
11.	Platelets	1) PLT,DXH (27MAR14) 2) PLT
12.	INR	1) STA INR MAMC(12/10) 2) INR (STA-COMPACT 03/09)
13.	Red blood cells	1) RBC,DXH (27MAR14) 2) RBC
14.	Albumin	1) ALBUMIN,COBAS (11MAR13;upd AMR 14OCT15) 2) ALBUMIN (C8000)
15.	Alanine Aminotransferase	1) ALANINE AMINOTRANS-FERASE,COBAS (15MAR13;upd14OCT15' 2) COBAS ALT
16.	Aspartate Aminotransferase	1) AST,COBAS (11MAR13;upd 14OCT15 abs) 2) AST (S/P 5600-1)2010MAR01-MJM
17.	Bilirubin	1) BILIRUBIN TOTAL,COBAS (27MAR13;upd 14OCT15 abs) 2) BILIRUBIN TOTAL(S/P 5600-2)

Table 2.3: 17 lab tests extracted for feature creation and their top 2 variations in names

No.	Comorbidities
1.	Anemia
2.	Arrhythmia
3.	Asthma
4.	Atherosclerosis
5.	Cancer
6.	Cardio Respiratory Failure Shock
7.	CHF
8.	Chronic Kidney Disease
9.	Cerebrovascular Disease
10.	Complicated Diabetes
11.	Connect Tissue
12.	COPD
13.	Acute Coronary Syndrome
14.	Dementia
15.	Depression
16.	Alcohol Abuse
17.	Fluid Disorders
18.	Gastrointestinal Disorder
19.	HIV
20.	Other Liver Disease
21.	Lung Disorders
22.	Malnutrition
23.	Myocardial
24.	Nephritis
25.	Other Heart Disease
26.	Paralysis
27.	Peptic Ulcer
28.	Peripheral Vascular Disease
29.	Pneumonia
30.	Other Psychiatric Disorder
31.	Renal Failure
32.	Rheumatic
33.	Sepsis
34.	Solid Tumor
35.	Ulcer
36.	Uncomplicated Diabetes
37.	Urinary Tract Disorder

Table 2.4: 37 comorbidities extracted for feature creation

and shape of features varies for each technique. At the time of admission of patients, only their previous clinical history is available. Retrospective models are trained on data from the whole encounter which makes them unfit to predict at the time of admission with no data available from the encounter yet. To resolve this problem, models can be trained to predict risk of readmission based on only the clinical history for a patient (see Chapter 3).

During the hospital stay more data comes in, in the form of labs, vitals and diagnoses for the patient. For each day in an in-patient stay, models can be built on similar patients who have stayed that long in the hospital to predict the risk of readmission (see Chapter 4). The details of data transformations to prepare the data in the format required by each of the modeling techniques are described in the sections below. Note that these sections correspond to the methods of Chapter 3 and 4 respectively.

2.4.1 At-Admit Modeling

The goal of this model is to predict the risk of thirty day readmission for patients at the time of admission. In this scenario there is no knowledge yet of the future events within the current encounter. Using just the clinical history of the patient, we can predict which patients are at high risk. The input feature vector is $X_i = x_{i1}, x_{i2}, \ldots x_{in}$ for an encounter i that contains the features mentioned in Table 2.2. Each row is designed to represent an encounter at the hospital for every patient. The goal is to predict Y (1/0) if the patient will be readmitted within thirty days of discharge.

Model	Input Vector Xi	Output Yi (1/0)
M	X_1 : $x_{11}, x_{12}, \ldots x_{1n}$	Y_1
	X_2 : $x_{21}, x_{22}, \ldots x_{2n}$	Y_2
	.....	...
	X_m : $x_{m1}, x_{m2}, \ldots x_{mn}$	Y_m

Table 2.5: Data format for at-admit modeling where m = number of encounters in the data and n = number of features from Population and Diagnoses as described in Table 2.2

2.4.2 Sequence Modeling

Sequence prediction is used to predict on outcome when the input data is in the form of sequential events [13]. A hospital encounter can be thought of a sequence of events happening to the patient, making it an interesting scenario to use for sequence prediction. Details of our sequence modeling method are provided in Chapter 4. For this technique we create an input vector which consists of a sequence of daily averages of lab test features. The length of the input vector changes for each length of stay. For the scope of this thesis, the experiments have been run on lengths of stay between 1 to 10 days. The distribution of the

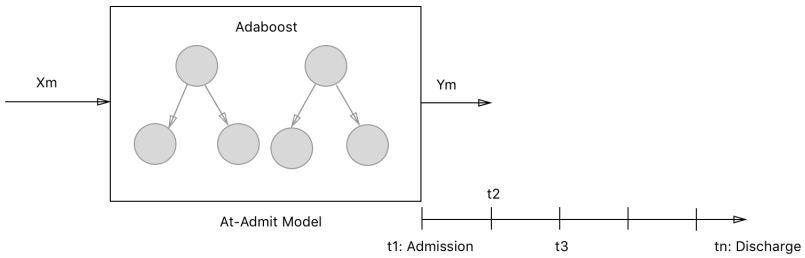

Figure 2.1: At-Admit Modeling Technique

length of stays in the data is shown in Figure x. For each of S sequence in Table 2.6, the goal is to predict Y (1/0) if the patient will be readmitted within thirty days of discharge. The dataset is divided into sets where all encounters within each set have the same length of stay. Models are trained and tested within each set.

LOS	Models	Input Sequence S	Output Y (1/0)
1	M_1 i=88699	$S_1 : x_{11}, x_{12}, \ldots x_{1n}$	Y_1
		$S_2 : x_{21}, x_{22}, \ldots x_{2n}$	Y_2
		....	...
		$S_i : x_{i1}, x_{i2}, \ldots x_{in}$	Y_i
2	M_2 j=79899	$S_1 : x_{11}, x_{12}, \ldots x_{1m}$	Y_1
		$S_2 : x_{21}, x_{22}, \ldots x_{2m}$	Y_2
		....	...
		$S_j : x_{j1}, x_{j2}, \ldots x_{jm}$	Y_j
..	....	.....	.....
10	M_{10} k=6193	$S_1 : x_{11}, x_{12}, \ldots x_{1o}$	Y_1
		$S_2 : x_{21}, x_{22}, \ldots x_{2o}$	Y_2
		....	...
		$S_k : x_{k1}, x_{k2}, \ldots x_{ko}$	Y_k

Table 2.6: Data format for sequence modeling where n = 17 lab test features, m = 2n, ... , o = kn

After performing the above mentioned steps, the data is ready to be consumed for predictive modeling. The next chapters discuss each kind of the modeling approaches and the experimental results gathered.

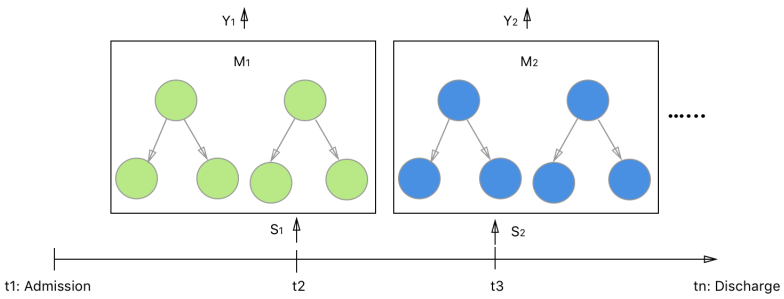

Figure 2.2: Sequence Based Modeling Technique

Chapter 3

Admission Model

3.1 Readmission Modeling

Most of the work done previously on predicting risk of readmission is based on retrospective modeling. For example the well known LACE index score [17] [21], often used as a baseline technique for predicting risk of thirty day readmission, is a retrospective calculation which requires events from the encounter like length of stay as well as the diagnoses associated to the encounter. These models are designed to make predictions at/after discharge of patients from the hospital to predict their risk of readmission. This limits the interventions that can be taken to avoid the readmission, because the hospital stay has already come to an end by the time the prediction is made.

In our solution we combine the retrospective readmission modeling techniques with daily predictive models to provide a framework that can predict risk of readmission throughout the encounter of a patient at the hospital. In this chapter we discuss the first component of the framework which is the at-admit model.

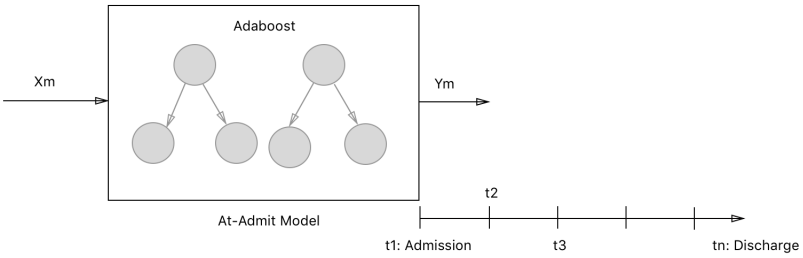

Figure 3.1: At-Admit Modeling Technique

3.2 Baseline LACE Index Score

The LACE index identifies patients that are at risk for readmission or death within thirty days of discharge. It incorporates four parameters.

- "L" stands for the length of stay of the index admission.
- "A" stands for the acuity of the admission. Specifically, if the patient is through the Emergency Department vs. an elective admission.
- "C" stands for comorbidities, incorporating the Charlson Comorbidity Index.
- "E" stands for the number of Emergency Department visits within the last 6 months.

LACE scores range from 1-19 and as mentioned above, predicts the rate of readmission or death within thirty days of discharge. A score of $0 - 4$ = Low; $5 - 9$ = Moderate; and a score of 10 = High risk of readmission. LACE has been used extensively as a baseline measure to predict the risk of thirty day readmission. Multiple studies have evaluated the predictive value of the LACE index [17][3]. Essentially, of the various tools available, LACE has been studied most extensively. It has been shown to have a moderate to high predictive value in identifying those patients at risk for readmission.

3.3 At-Admit Modeling Approach

The at-admit modeling setup is designed similar to retrospective readmission modeling. The model is trained to predict the risk of thirty day readmission of a patient at the time of admission using only the clinical history of the patient.

Each row in the data represents an encounter to the hospital. All the features are generated from previous encounters for the patient. A risk score is predicted which is the risk of that patient getting readmitted within thirty days of being discharged from the current encounter.

3.3.1 Features

The features used for training the models are listed in Table 3.1. The lab test features are not used in the at-admit model because the lab tests are taken during the encounter and are not available at the time of admission. The Sequence based and Till-Day cumulative techniques discussed in Chapter 4 use the lab test features. Finally the predictions from at-admit model and Till-Day cumulative model are combined together to arrive at the overall model.

3.3.2 Model

Another very successful technique for building ensembles is boosting [7]. In boosting, one maintains a vector of weights for the training instances, which

Group	Features	Type
Demographics	Age	Integer
	Gender	Boolean
Medical Diagnosis	For each of 37 comorbidities in Table 2.4	
	1) If the comorbidity was ever diagnosed	Boolean
	2) The last time it was diagnosed	Integer
	3) The number of times it was diagnosed	Integer
Others	Prior admission count	Integer
	Emergency visits in last 6 months	Integer
Response Variable	Thirty Readmission flag	Boolean

Table 3.1: Features summary for At-Admit Model

is initialized with uniform weights. Next, one iteratively trains a model on the weighted examples, and increases the weights of the examples that were misclassified by this new model. This iterative process continues until the new model no longer does better than random guessing. Then all trained models are collected in an ensemble. When classifying new instances, the outputs of the models are combined with weighted voting. The term boosted decision trees refers to an ensemble of decision trees constructed through boosted.

We have used stochastic gradient boosting with 50 iterations where at each iteration of the algorithm, a base learner fits on a sub-sample of the training set drawn at random without replacement [8]. The sub-sample size is some constant fraction f of the size of the training set. When f = 1, the algorithm is deterministic and identical to the one described above. Smaller values of f introduce randomness into the algorithm and help prevent over-fitting, acting as a kind of regularization. We have used f = 0.5 meaning that one half of the training set is used to build each base learner. In our experiments we have used the R package 'ada' for the implementation of boosted decision trees.

3.4 Experimental Results

3.4.1 Dataset

The data used for training and testing is explained in detail in Chapter 2. Data from Population and Diagnosis files is used to extract features for the model mentioned in Table 3.1. After cleaning of the data, there are a total of 88699 encounters in the data with 63342 unique patients. The thirty day readmission rate is 12.3% in the data.

3.4.2 Evaluation Measures

We have used K-Fold cross validation to calculate the results across the dataset where k = 10. We measure the performance of the algorithms using sensitivity

(recall), precision, area under the curve (AUC) and accuracy for classification. For our binary classification task of readmission, the outcome can be positive or negative. We use the predicted probability and a threshold to assign the binary label. As readmission is an unbalanced problem, we use a threshold closer to the mean of the predicted risk scores. All test instances with risk score above the threshold are labelled as readmission class 'yes' (1) and the others as readmission class 'no' (0). The prediction results for each test instance may or may not match the actual class. Therefore, we get the following results:

- True positive (TP): correctly classified as positive.
- False positive (FP): incorrectly classified as positive.
- True negative (TN): correctly classified as negative.
- False negative (FN): incorrectly classified as negative.

Based on the counts of these four measures, we compute the following evaluation metrics:

$$Sensitivity/Recall = \frac{TP}{TP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.3}$$

Along with accuracy values we also use area under the curve [Figure **??**]. Some classiers output a probability that an instance belongs to the positive class, and require the selection of a threshold to turn the probability into a class label. For example, we might choose to label encounters to have readmission if and only if the classier says they are going to have readmission with probability at least .75. Depending on how we choose the threshold (e.g. 0.2, 0.4, 0.6, 0.75, etc.), the FPR and the TPR will vary. We can test our classier for various choices of the threshold, and plot the corresponding (FPR,TPR) points. The resulting curve is called a ROC curve

3.4.3 Results

The results are calculated by using 10-cross validation. We have implemented the LACE score index as well as the at-admit model based on the data available at the time of admission . The thresholds for both techniques are set to be the mean of the predicted values. Every test instance with a risk score greater or equal than the threshold is labeled 1 (thirty day readmission yes) and every test instance with risk score less than the threshold is labelled as 0 (thirty day readmission no). The results are summarized in Table 3.2:

Analyzing the results it is evident that the at-admit model outperforms the LACE index in catching more high risk patients at admission. The AUC for

Metric	LACE	At-Admit Model
Recall	0.4366 (threshold = 1.14)	0.6058 (threshold = 0.12)
Precision	0.2036 (threshold = 1.14)	0.2103 (threshold = 0.12)
Accuracy	0.7200 (threshold = 1.14)	0.6712 (threshold = 0.12)
AUC	0.5982	0.6431

Table 3.2: Metrics Summary for At-Admit Model

	1	0
1	4774	6161
0	18671	59093

Table 3.3: Confusion Matrix for LACE at admission

	1	0
1	7283	4740
0	27344	58202

Table 3.4: Confusion Matrix for At-Admit Model

LACE at admission is lower than the AUC for the at-admit model. The LACE index is also less precise in identifying the patients that should be prioritized by the hospital which will result in additional costs of resources allocated.

Chapter 4

Sequence Based Modeling

Sequences are a significant type of data which occur frequently in several scientific, medical, security, and other domains [19]. Sequence data mining provides helpful tools for exploring useful knowledge hidden in large sequence datasets. Many prediction problems in clinical care settings are inherently sequence prediction problems: given a sequence of events like vital readings or lab tests taken during an encounter, we wish to predict a future outcome related to it.

Despite the sequence based nature of clinical events, most of the clinical studies use a retrospective approach in their feature creation and predictive modeling setup [9]. In Chapter 3 we have presented a modified alternative to retrospective readmission models to predict at the time of admission. In this chapter we propose a new sequence based technique to predict a clinical outcome, like thirty day readmissions, using clinical events of an encounter as sequence inputs.

4.1 Retrospective Model Baseline

To compare our sequence based model we use the conventional approach of retrospective modeling for clinical scenarios as the baseline. Since our goal is to provide a framework solution which is designed for production scenario, this baseline helps in better validation of the sequence based approach as opposed to using a less sophisticated approach like a majority baseline or random guessing. In this baseline approach we use the same attributes from data that we will use in the sequence model as well as the same machine learning model. The difference is the design setup of these two approaches.

4.1.1 Features

Among the data components described in Table 2.1, lab data changes during an encounter of a patient. This is why we use the lab features as elements of our

Models	Input Vector X	Output Y (1/0)
M i = 88699 (number of instances)	X_1 : $x_{11}, x_{12}, \ldots x_{1n}$ X_2 : $x_{21}, x_{22}, \ldots x_{2n}$ .... X_i : $x_{i1}, x_{i2}, \ldots x_{in}$	Y_1 Y_2 ... Y_i

Table 4.1: Data format for retrospective baseline training with n = 17 lab test features averaged across the encounter

sequences. Similarly for the baseline we use the same 17 lab tests mentioned in Table 2.3 averaged across the whole encounter as input features.

4.1.2 Modeling Setup

We use boosted decision trees as the predictive model which are described in detail in Section 3.3.2. Each input row to train the model is a patient encounter in the hospital with average lab tests as features.

For comparing the baseline to the sequence based approach, we want to analyse how it performs when deployed in a clinical setting. When a patient is in the hospital, the value of the averaged lab tests will be the average value till the time of prediction. So although the retrospectively trained models are trained over features calculated over the entire encounter, the prediction is made based on feature calculation on the data available for the encounter thus far. This impacts the performance of the retrospective models on predictions made during the encounter.

For testing the model trained on the setup as described in Table 4.1, the average of the features is the average until the day of prediction from day 1 to day 10. The results are compared with the sequence based approach in Section 4.4.

4.2 Sequence Based Modeling

When the patient is admitted to a hospital, the data associated to the patient expands vertically and horizontally. We have designed a sequence modeling approach to make predictions on this expanding sequence of events. This approach helps in addressing the following two challenges of retrospective modeling:

- Inability to capture trends of changing patient state using the growing data points during the encounter in the features.
- Unsuitability of the model design for prediction during the encounter of the patient and before the discharge from the hospital.

4.2.1 Features

We have used the 17 lab tests mentioned in Table 2.3 as features for the sequence models. For each day of hospital stay a sequence is created of daily average lab test values for all the 17 lab tests. For example for the Day 3 sequence model, the input consists of a sequence of average values of the 17 lab tests for day 1, 2 and 3. This sequence is then used to train and test the sequence model for Day 3. The data format for the sequence modeling approach is shown in Table 4.2.

LOS	Models	Input Sequence S	Output Y (1/0)
1	M_1 i=88699	$S_1 : x_{11}, x_{12}, \ldots x_{1n}$	Y_1
		$S_2 : x_{21}, x_{22}, \ldots x_{2n}$	Y_2
		....	...
		$S_i : x_{i1}, x_{i2}, \ldots x_{in}$	Y_i
2	M_2 j=79899	$S_1 : x_{11}, x_{12}, \ldots x_{1m}$	Y_1
		$S_2 : x_{21}, x_{22}, \ldots x_{2m}$	Y_2
		....	...
		$S_j : x_{j1}, x_{j2}, \ldots x_{jm}$	Y_j
..	....	.....	.....
10	M_{10} k=6193	$S_1 : x_{11}, x_{12}, \ldots x_{1o}$	Y_1
		$S_2 : x_{21}, x_{22}, \ldots x_{2o}$	Y_2
		....	...
		$S_k : x_{k1}, x_{k2}, \ldots x_{ko}$	Y_k

Table 4.2: Data format for sequence modeling where n = 17 lab test features averaged across each day of encounter, m = 2n, ... , o = kn

4.2.2 Model

To make a valid comparison between techniques, we have used boosted decision trees in the sequence based approach as well. In total there are ten boosted decision trees predicting for each length of stay between 1-10 using the sequence of daily lab readings as input. The sequences enable the predictive model to encapsulate the growing data for a patient and their changing condition over the encounter till the time of prediction in the boosting iterations. The design of this approach is shown in Figure 4.1.

4.3 Experimental Results

The evaluation measures mentioned in Section 3.4.2 are used here as well to compare the sequence based modeling technique to the retrospective baseline.

K-fold cross validation is done with k=10 to calculate metrics for the 10 models. In each fold of cross validation, the same encounters are used for training and testing the sequence based as well as the baseline model. The number of encounters available for each of the 10 models are shown in Figure 4.2. This

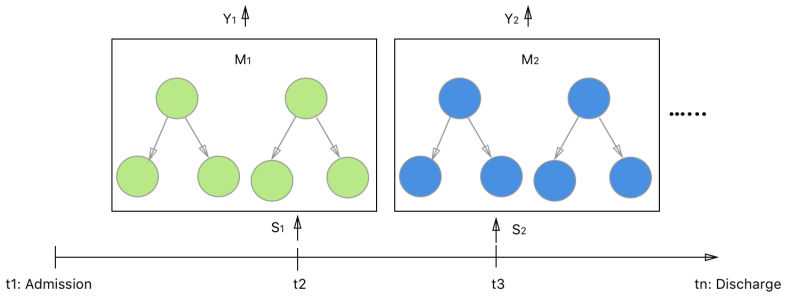

Figure 4.1: Sequence Based Modeling Technique

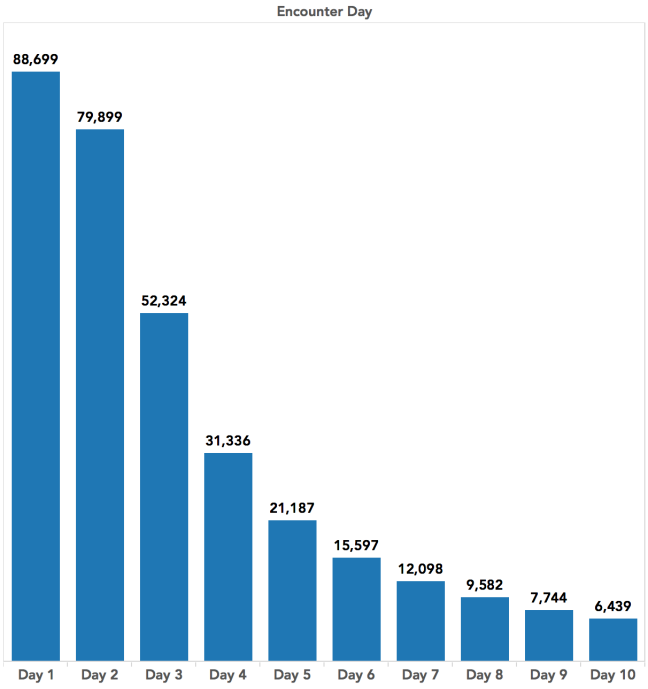

Figure 4.2: Number of input encounters for each model from day 1 to day 10.

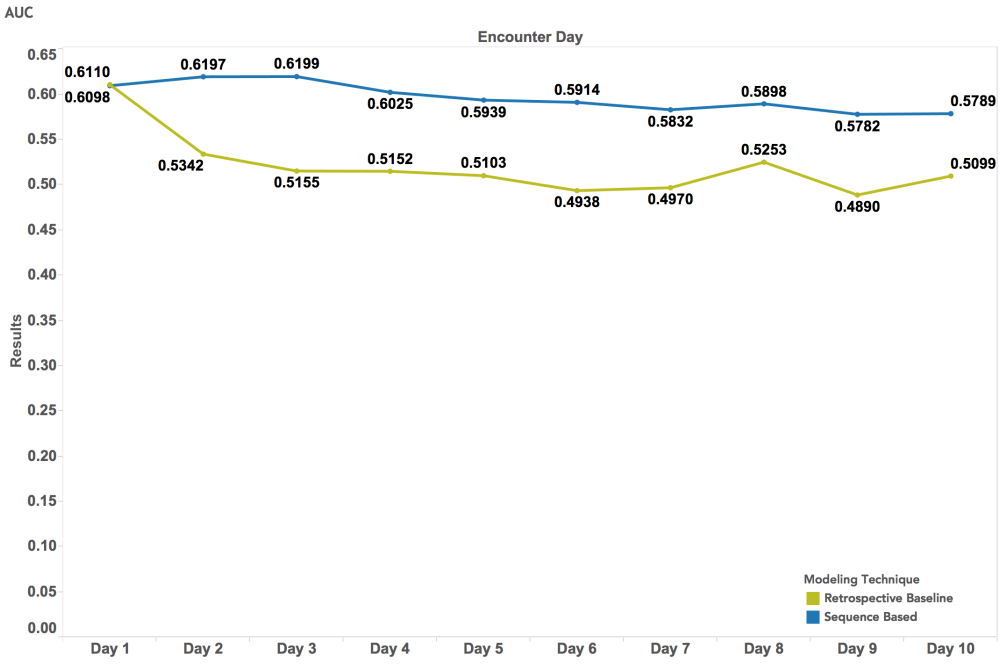

Figure 4.3: AUC comparison between Sequence based models and Retrospective baseline models where the positive class represents patients who will readmit within thirty days of discharge

shows the distribution of length of stays of the patients at the hospital in the dataset.

For Figure 4.4, 4.5 and 4.6 the thresholds to assign the label are set using the mean value of the predicted risk scores. Analyzing the results we can see that the sequence based approach outperforms the baseline approach in all metrics. The AUC comparison shows that the baseline approach performs almost as good as random guessing while predicting for patients who have length of stays greater than two days.

Another observation from the accuracy and AUC results of both techniques is that as the length of stay increases it becomes difficult to accurately differentiate between patients who are going to be readmitted and the patients who are not going to be readmitted. This is because the patients staying longer fall under the sicker population and the differentiation within that population is difficult. This is also because the number of instances to learn from is lesser as the length of stay increases as we have seen in the length of stay distribution in Figure 4.2.

The power of sequence based versus accumulated features is reflected in the trend lines of all metrics as we go from day 1 to day 10. The first day results are highly similar because the sequence input has the same features as the baseline technique, however as the sequence grows, the metrics show significantly better

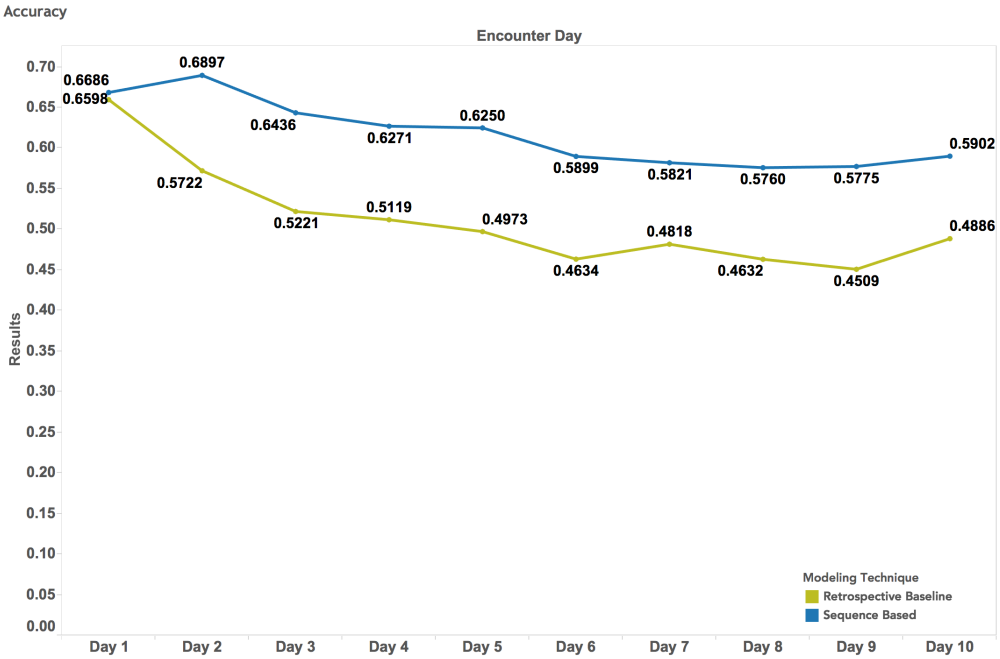

Figure 4.4: Accuracy comparison between Sequence based models and Retrospective baseline models

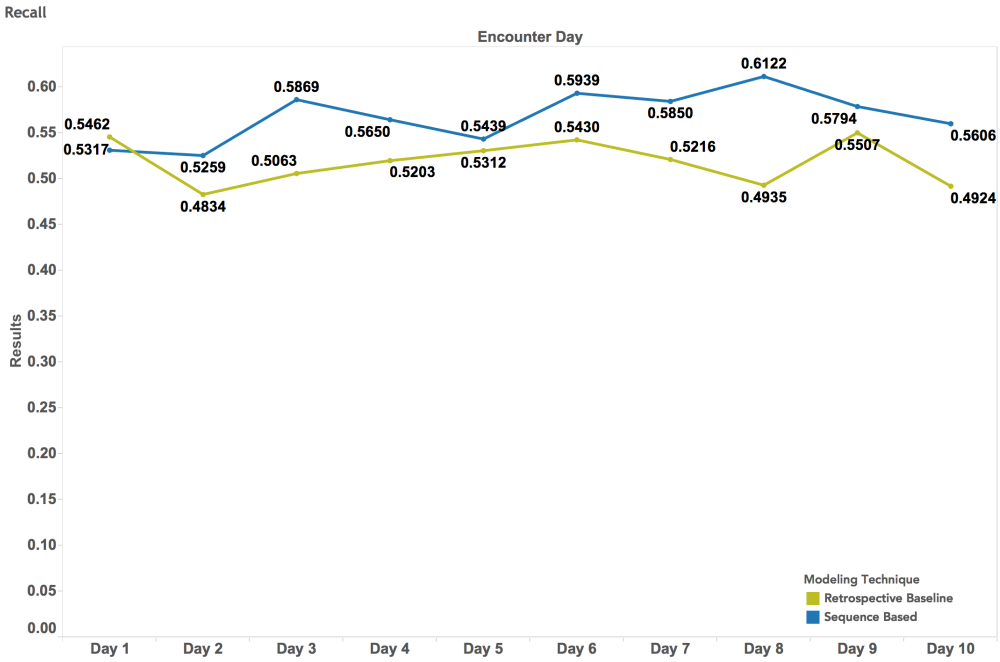

Figure 4.5: Recall comparison between Sequence based models and Retrospective baseline models

results for the sequence based approach.

Through these results we have validated that a sequenced based approach of handling clinical events like lab readings to predict a future event like thirty day readmission performs better than cumulative retrospective modeling techniques. This makes it more suitable for deployment in the hospital system. Using the Sequence based models in combination with the at-admit model, we have constructed a framework that predicts thirty day readmission as a clinical outcome while encapsulating sensitivity to different times of prediction in a hospital stay as well as the trends of clinical events.

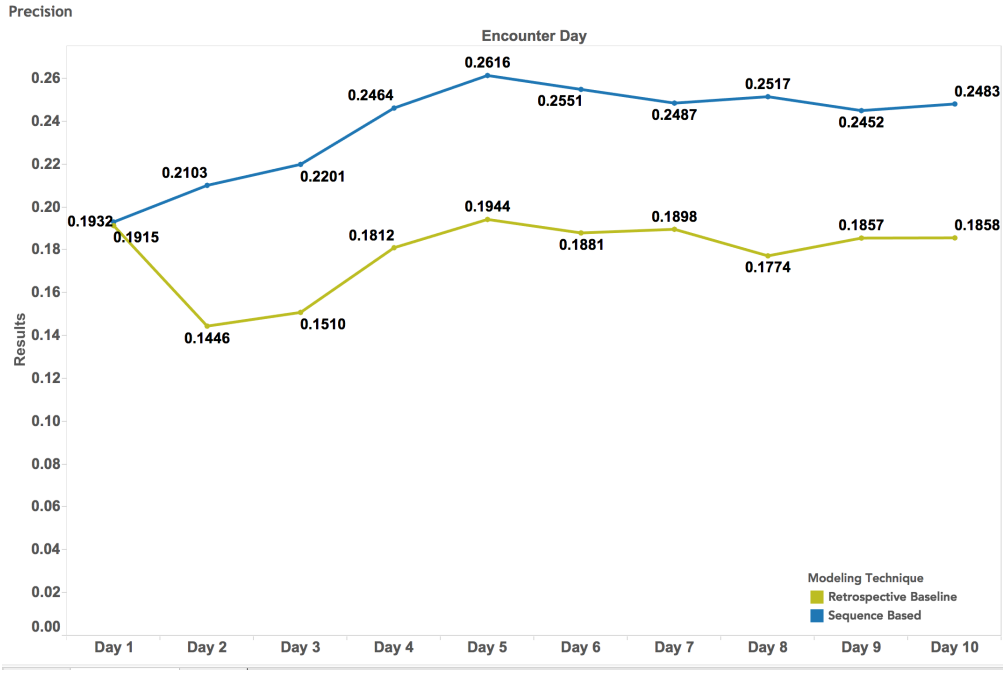

Figure 4.6: Precision comparison between Sequence based models and Retrospective baseline models

Chapter 5

Conclusion

In this thesis we have developed a technique for making clinical predictions, designed for the growing data in the real world setting. In each of the respective sections and chapters we have reviewed and mentioned the related work found in literature with an introduction in Chapter 1. The goal of our work was to propose a solution that effectively utilizes the clinical data that grows for a patient within and across hospital admissions.

We have used an EHR dataset to create two modeling setups for predicting thirty day all cause readmissions. The at-admit model predicts significantly better than the LACE index score for patients at the time of admission to the hospital (see Chapter 3). The sequence feature based models proposed in Chapter 4 predict better than a retrospective model baseline when compared on how they might be used in an operational clinical setting where predictions are needed everyday to prioritize and target interventions. We use the metrics of area under the curve (AUC), recall, precision and accuracy to evaluate the models. The sequence based models perform better in all metrics on all days (1-10) than the baseline. As the length of stay increases, we see a downward trend for the accuracy and precision of our models because not only the number of instances to learn from decreases but also it becomes more difficult to correctly label outcome in a set of sicker patients.

5.1 Results from At-Admit + Sequence Based Models

We also ran experiments to combine the predictions from at-admit models (see Chapter 3) with the sequence based models (see Chapter 4). We take the predicted score for the same patient from the at-admit model and the sequence based model and compute an average risk score using these two scores. We assign the output label based on the threshold in the similar fashion as described in Section 4.3 and Section 3.4.3 using the mean of the predicted values. The results are evaluated using AUC, accuracy, recall and precision which are shown

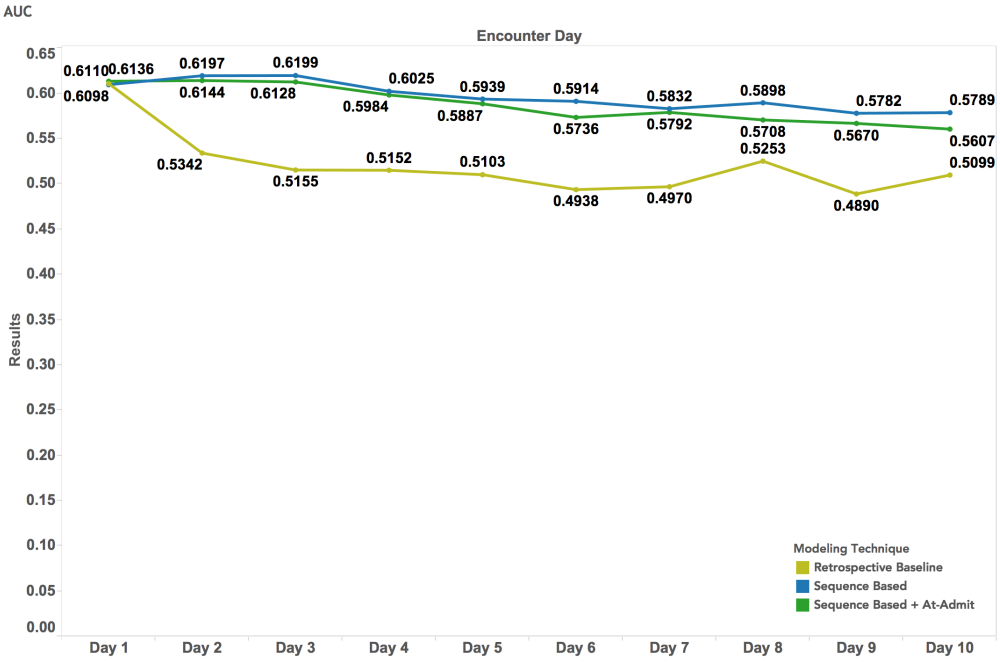

Figure 5.1: AUC comparison between combination of Sequence based models + At-Admit models, Sequence based models and Retrospective baseline models

in Tables 5.1,5.2,5.3,5.4 respectively.

Looking at the results from the combination of these two models, we can see that the sequence based models outperform the combination models (sequence based + at-admit) in AUC, accuracy and precision. However the combination models perform better in recall than the sequence based models. It is also interesting to note that the area under the curve is also very close for these two. Analyzing this we can say that using at-admit model to predict patient's risk at the time of admission and using the sequence based models independently while the patient is admitted to the hospital will give more accurate and precise predictions. However, if the goal is to optimize for recall and catch as many high risk patients as we can, than the combination of these two techniques can be used to predict risk scores when the patient is admitted to the hospital.

5.2 Summary and Future Work

Using the sequence based models and the at-admit model proposed in this thesis, we can predict thirty day readmission at various time points of a patient's encounter at the hospital. This solution is sensitive to the time of prediction during the encounter and is designed and evaluated from the perspective of the

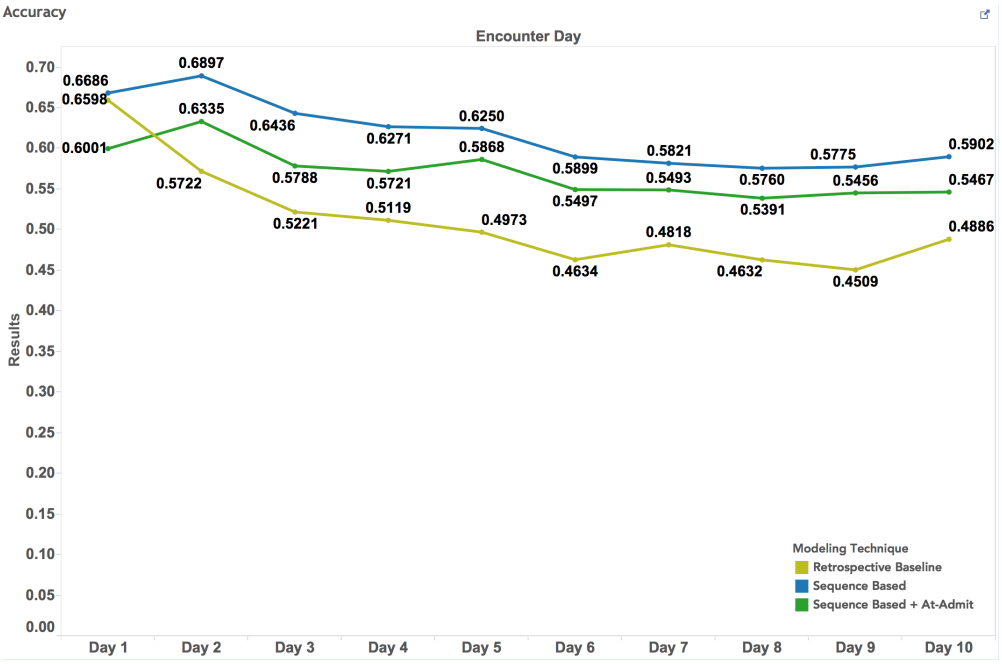

Figure 5.2: Accuracy comparison between combination of Sequence based models + At-Admit models, Sequence based models and Retrospective baseline models

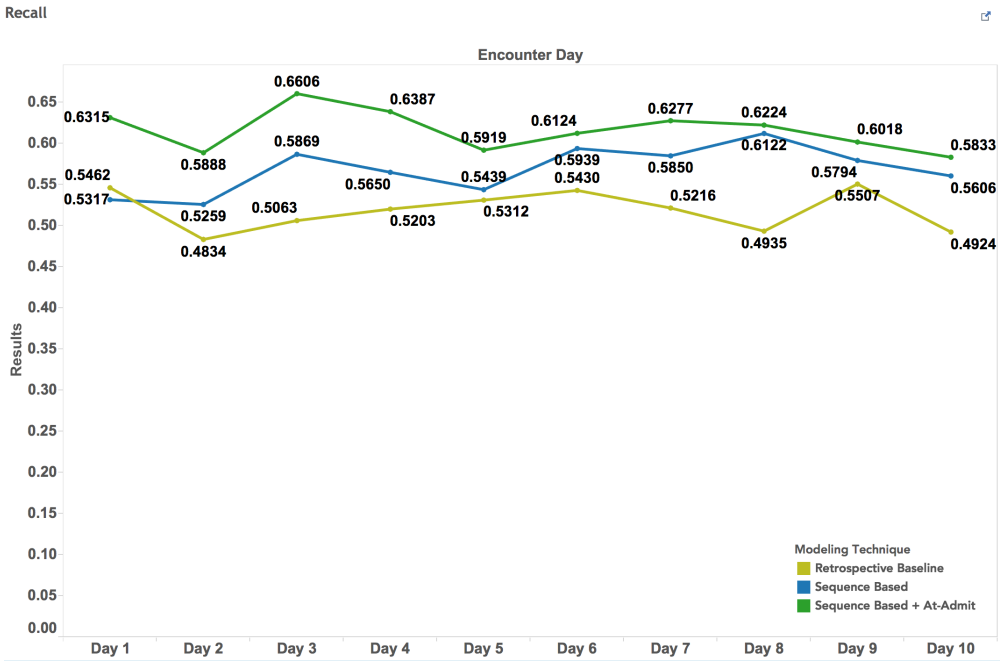

Figure 5.3: Recall comparison between combination of Sequence based models + At-Admit models, Sequence based models and Retrospective baseline models

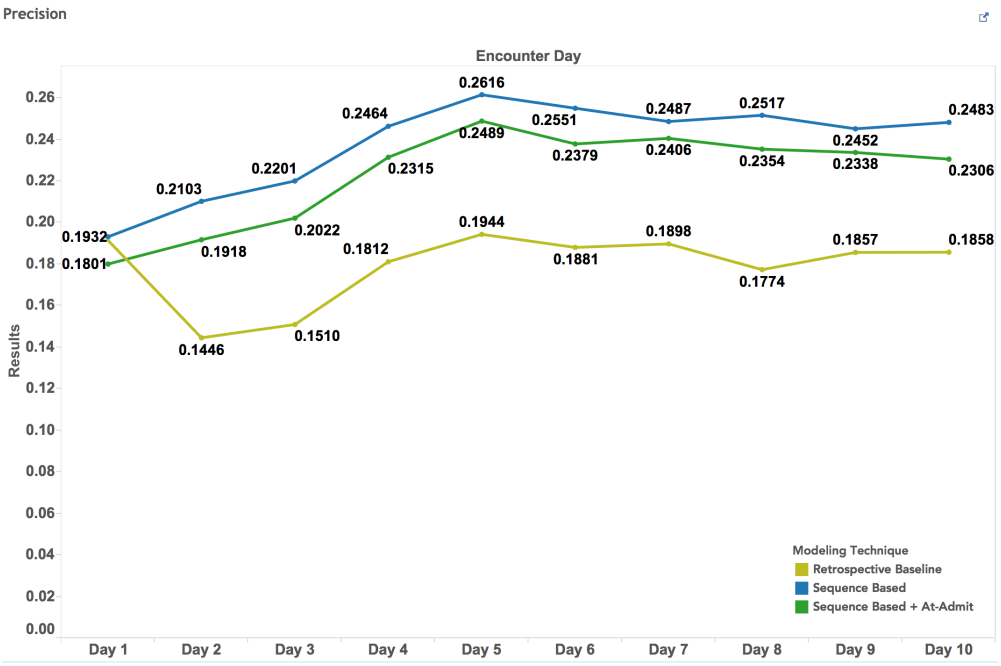

Figure 5.4: Precision comparison between combination of Sequence based models + At-Admit models, Sequence based models and Retrospective baseline models

real world setup. Its applicability is by far not limited to the problem of predicting whether a patient will be readmitted within thirty days after discharge from the hospital. Indeed, there are many similar problems to be solved in the healthcare domain that can be addressed using a framework for sequence based prediction. Some other useful scenarios to which the methods proposed in this thesis are applicable include:

- Predicting length of stay of the patient for an admission. Knowledge about the expected length of stay is valuable for better planning of resources, among other things.
- Identifying patients who are at high risk for developing a surgical site infection.
- Predicting risk of mortality for patients in the next 6 months of their hospital admission.

Finally, there are multiple dimensions along which our work can be extended to improve the solution:

- **Feature Space Expansion** In our solution we have used lab features for sequence based models and diagnoses and demographic features for the at-admit model. These can be expanded to include more clinical features like vital readings, prescriptions data and socio-demographic features. Inclusion of these features will aid in capturing the patient state better and improvement of results.
- **Model Expansion** We implemented boosted decision trees, support vector machines, logistic regression and naive bayes for retrospective event based modeling to predict thirty day readmissions. Boosted decision trees showed the best results and therefore we have used them in all our techniques to have a valid comparison. However the comparison can be expanded using all of these techniques to show that the sequence based technique performs better overall.
- **Dataset Expansion** We have used hospital admissions data from 6 facilities across U.S for this work. These experiments can be run with more hospital admissions datasets from more facilities to validate the results across a variety of samples.

Bibliography

[1] Carol M Ashton and Nelda P Wray. A conceptual framework for the study of early readmission as an indicator of quality of care. Social science & medicine, 43(11):1533–1541, 1996.

[2] Senjuti Basu Roy, Ankur Teredesai, Kiyana Zolfaghar, Rui Liu, David Hazel, Stacey Newman, and Albert Marinez. Dynamic hierarchical classification for patient risk-of-readmission. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pages 1691–1700. ACM, 2015.

[3] E. Ben-Chetrit, C. Chen-Shuali, E. Zimran, G. Munter, and G. Nesher. A simplified scoring tool for prediction of readmission in elderly patients hospitalized in internal medicine departments. Isr Med Assoc J, 14, 2012.

[4] Jochanan Benbassat and Mark Taragin. Hospital readmissions as a measure of quality of health care: advantages and limitations. Archives of internal medicine, 160(8):1074–1081, 2000.

[5] CMS. Readmissions reduction program (hrrp), 2016.

[6] Kansagara D, Englander H, Salanitro A, and et al. Risk prediction models for hospital readmission: A systematic review. JAMA, 306(15):1688–1698, 2011.

[7] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In European conference on computational learning theory, pages 23–37. Springer, 1995.

[8] Jerome H Friedman. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4):367–378, 2002.

[9] Brian W Jack, Veerappa K Chetty, David Anthony, Jeffrey L Greenwald, Gail M Sanchez, Anna E Johnson, Shaula R Forsythe, Julie K O'Donnell, Michael K Paasche-Orlow, Christopher Manasseh, et al. A reengineered hospital discharge program to decrease rehospitalization: a randomized trial. Annals of internal medicine, 150(3):178–187, 2009.

[10] Colleen K McIlvennan, Zubin J Eapen, and Larry A Allen. Hospital readmissions reduction program. Circulation, 131(20):1796–1803, 2015.

[11] Naylor MD, Brooten D, Campbell R, and et al. Comprehensive discharge planning and home follow-up of hospitalized elders: A randomized clinical trial. JAMA, 281(7):613–620, 1999.

[12] Mark V. Williams Stephen F. Jencks and Eric A. Coleman. Rehospitalizations among patients in the medicare fee-for-service program. New England Journal of Medicine, 360(14):1418–1428, 2009.

[13] Ron Sun and C Lee Giles. Sequence learning: from recognition and prediction to sequential decision making. IEEE Intelligent Systems, 16(4):67–70, 2001.

[14] Shanu Sushmita, Garima Khulbe, Aftab Hasan, Stacey Newman, Padmashree Ravindra, Senjuti Basu Roy, Martine De Cock, and Ankur Teredesai. Predicting 30-day risk and cost of" all-cause" hospital readmissions. In Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[15] Hariharasudhan Viswanathan, Baozhi Chen, and Dario Pompili. Research challenges in computation, communication, and context awareness for ubiquitous healthcare. IEEE Communications Magazine, 50(5):92–99, 2012.

[16] Douglas S Wakefield and David R Mehr. Risk factors for all-cause hospital readmission within 30 days of hospital discharge. JCOM, 20(5), 2013.

[17] Hao Wang, Richard D. Robinson, Carlos Johnson, Nestor R. Zenarosa, Rani D. Jayswal, Joshua Keithley, and Kathleen A. Delaney. Using the lace index to predict hospital readmissions in congestive heart failure patients. BMC Cardiovascular Disorders, 14(1):97, 2014.

[18] Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. Finding progression stages in time-evolving event sequences. In Proceedings of the 23rd international conference on World wide web, pages 783–794. ACM, 2014.

[19] Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. Finding progression stages in time-evolving event sequences. In Proceedings of the 23rd International Conference on World Wide Web, WWW '14, pages 783–794, New York, NY, USA, 2014. ACM.

[20] Ian T Young. Proof without prejudice: use of the kolmogorov-smirnov test for the analysis of histograms from flow systems and other sources. Journal of Histochemistry & Cytochemistry, 25(7):935–941, 1977.

[21] Bichen Zheng, Jinghe Zhang, Sang Won Yoon, Sarah S Lam, Mohammad Khasawneh, and Srikanth Poranki. Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Systems with Applications, 42(20):7110–7120, 2015.