Learning to Prevent Healthcare-Associated Infections: Leveraging Data Across Time and Space to Improve Local Predictions

by

Jenna Wiens

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2014

(C) Massachusetts Institute of Technology 2014. All rights reserved.

Author ........................................
Department of Electrical Engineering and Computer Science
May 21st, 2014

Certified by ..................
John Guttag
Dugald C. Jackson Professor
Thesis Supervisor

Accepted by ......................
Léslie A. Kolodziejski
Chair of the Committee on Graduate Students
Learning to Prevent Healthcare-Associated Infections:
Leveraging Data Across Time and Space to Improve Local Predictions

by

Jenna Wiens

Submitted to the Department of Electrical Engineering and Computer Science
on May 21st, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract
The proliferation of electronic medical records holds out the promise of using machine learning and data mining to build models that will help healthcare providers improve patient outcomes. However, building useful models from these datasets presents many technical problems. Among the challenges are the large number of factors (both intrinsic and extrinsic) influencing a patient’s risk of an adverse outcome, the inherent evolution of that risk over time, and the relative rarity of adverse outcomes, institutional differences and the lack of ground truth.

In this thesis we tackle these challenges in the context of predicting healthcare-associated infections (HAIs). HAIs are a serious problem in US acute care hospitals, affecting approximately 4% of all inpatients on any given day. Despite best efforts to reduce incidence, HAIs remain stubbornly prevalent. We hypothesize that one of the reasons why is lack of an effective clinical tool for accurately measuring patient risk.

Therefore, we develop accurate models for predicting which patients are at risk of acquiring an infection with Clostridium difficile (a common HAI). In contrast to previous work, we take a novel data-centric approach, leveraging the contents of EMRs from over 100,000 hospital admissions. We show how, by adapting techniques from time-series classification, transfer learning and multitask learning, we can learn more accurate models for patient risk stratification.

Our model, based on thousands of variables both time-varying and time-invariant, does not remain static but changes over the course of a patient admission. Applied to a held-out validation set of 25,000 patient admissions, our model achieved an area under the receiver operating characteristic curve of 0.81 (95% CI 0.78-0.84). The model has been successfully integrated into the health record system at a large hospital in the US, and is being used to produce daily risk estimates for each inpatient.

While more complex than traditional risk stratification methods, the widespread development and use of such data-driven models could ultimately enable cost-effective, targeted prevention strategies that reduce the incidence of HAIs.

Thesis Supervisor: John Guttag
Title: Dugald C. Jackson Professor
Acknowledgments

This work would not have been possible without the guidance and support of many. While this dissertation is all about the research, my PhD was shaped mostly by the people I met along the way.

John Guttag has been a phenomenal advisor throughout my graduate career. I am inspired by John’s ability to always see the big picture, while also grasping the important technical details of a research project (while juggling 10 or more different projects). John has made my graduate school experience not only intellectually stimulating, but also enjoyable.

Many of the initial ideas for this work came during an internship at MSR, under the supervision of Eric Horvitz. Eric’s enthusiasm for research is contagious. I have lost count of the number of times a discussion with him has led to a new insight or direction. He has been fundamental to this work, as have others in particular Paul Koch and Hank Rappaport.

This work has focused on improving patient care, and to that end our clinical collaborators have been critical. In particular, Wayne Campbell, Ella Franklin and Mark Smith at MedStar have been wonderful collaborators. Wayne’s expertise in infectious disease has proven invaluable, and Ella’s thoughtful insight into the clinical application of the tool has helped us realize the project beyond a simple proof of concept.

Many people have contributed to this work indirectly, through discussions (both formal and informal). Polina Golland, a member of my thesis committee, has been both a sounding board and a supportive mentor. Zeeshan Syed helped shape my early years as a graduate student, along with my senior officemates Ali Shoeb and Eugene Shih. Collin Stultz’s constructive criticism led to improvements. My current labmates Anima, Gartheeban, Jen, Yun, Joel, Guha, Orly, and Amy have been wonderful colleagues. I feel truly lucky to have gotten the opportunity to work with such a fantastic group of researchers.

Whether they knew it or not, the friends I have made along the way (many listed above) but also Lisa Burton, Christy and Jamie Teherani, Ellan Spero, Molly Roberts and David Fenning have played a crucial role in my PhD. Their friendship, along with the friendship of many others (the list is too long to include here) gave me courage during periods of self-doubt. Finally, I acknowledge my family: my mother, my father, my siblings and my fiancé, who have tirelessly supported and cheered me along since day one. Their love gives me the confidence to advance in the direction of my dreams.
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>Background</td>
<td>25</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>25</td>
</tr>
<tr>
<td>2.2</td>
<td>The Data</td>
<td>26</td>
</tr>
<tr>
<td>2.3</td>
<td>Identifying Positive Cases for <em>C. difficile</em></td>
<td>31</td>
</tr>
<tr>
<td>2.4</td>
<td>Colonization Pressure</td>
<td>32</td>
</tr>
<tr>
<td>2.5</td>
<td>Preprocessing the Data</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>Leveraging the Richness of the Electronic Medical Record</td>
<td>37</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>37</td>
</tr>
<tr>
<td>3.2</td>
<td>The Data</td>
<td>39</td>
</tr>
<tr>
<td>3.3</td>
<td>Materials &amp; Methods</td>
<td>41</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Feature Extraction</td>
<td>41</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Learning to Predict Risk</td>
<td>44</td>
</tr>
<tr>
<td>3.4</td>
<td>Experiments and Results</td>
<td>45</td>
</tr>
<tr>
<td>3.5</td>
<td>Summary &amp; Conclusion</td>
<td>51</td>
</tr>
<tr>
<td>4</td>
<td>Patient Risk Stratification as a Time-Series Classification Task</td>
<td>53</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>4.2</td>
<td>The Data</td>
<td>55</td>
</tr>
<tr>
<td>4.3</td>
<td>Methods</td>
<td>56</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Feature Extraction</td>
<td>56</td>
</tr>
</tbody>
</table>
5 Risk Stratification with Time-Varying Coefficients: A Multitask Approach

5.1 Introduction .......................................................... 71
5.2 The Data ................................................................. 73
5.3 Methods ................................................................. 73
  5.3.1 Feature Extraction ................................................. 73
  5.3.2 Learning to Predict Daily Risk .................................. 74
5.4 Experiments and Results ............................................ 77
  5.4.1 Evaluating the Model .............................................. 80
5.5 Conclusion ............................................................. 86

6 Transferring Knowledge Across Hospitals ........................................... 87

6.1 Introduction ............................................................. 87
  6.1.1 Background on Transfer Learning ................................. 89
6.2 Methods ................................................................. 90
  6.2.1 Data and Preprocessing ............................................ 90
  6.2.2 Risk Stratification .................................................. 94
6.3 Experiments and Results ............................................. 96
  6.3.1 Including Source Data Helps ..................................... 97
  6.3.2 Target-Specific Features Are Important ......................... 100
  6.3.3 More Data is Not Always Better ................................ 102
  6.3.4 Not All Transfer is Equal ......................................... 103
6.4 Discussion & Conclusion ............................................. 104
7 Translation to Clinical Practice

7.1 Introduction .......................................................... 107
7.2 Review of Current Clinical Guidelines .......................... 108
  7.2.1 Antimicrobial Stewardship .................................. 109
  7.2.2 PPI Cessation ..................................................... 111
  7.2.3 Environmental Cleaning ....................................... 112
  7.2.4 Isolation/Contact Precautions/Hand hygiene .............. 114
  7.2.5 Probiotics ......................................................... 115
7.3 A Cost-Benefit Analysis ........................................... 117
7.4 Summary & Conclusion ............................................. 120

8 Summary and Conclusion ............................................. 123

A Database Details ....................................................... 129

B Testing Protocol for C. difficile .................................... 135
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Patient Laboratory Results</td>
<td>28</td>
</tr>
<tr>
<td>2-2</td>
<td>ICD-9 Codes</td>
<td>29</td>
</tr>
<tr>
<td>2-3</td>
<td>Patient Medications</td>
<td>29</td>
</tr>
<tr>
<td>2-4</td>
<td>Colonization Pressure</td>
<td>34</td>
</tr>
<tr>
<td>3-1</td>
<td>Study Population</td>
<td>41</td>
</tr>
<tr>
<td>3-2</td>
<td>Risk Period</td>
<td>42</td>
</tr>
<tr>
<td>3-3</td>
<td>ROC Curves</td>
<td>47</td>
</tr>
<tr>
<td>3-4</td>
<td>Calibration Plots</td>
<td>49</td>
</tr>
<tr>
<td>3-5</td>
<td>Net Reclassification Improvement</td>
<td>50</td>
</tr>
<tr>
<td>3-6</td>
<td>Confusion Matrix</td>
<td>50</td>
</tr>
<tr>
<td>3-7</td>
<td>Positive Predictive Value</td>
<td>52</td>
</tr>
<tr>
<td>4-1</td>
<td>Distribution of SVM Output</td>
<td>59</td>
</tr>
<tr>
<td>4-2</td>
<td>Patient Risk Profiles</td>
<td>60</td>
</tr>
<tr>
<td>4-3</td>
<td>A Two-Step Approach to Risk Stratification</td>
<td>62</td>
</tr>
<tr>
<td>4-4</td>
<td>Smoothing the Risk Profiles</td>
<td>65</td>
</tr>
<tr>
<td>4-5</td>
<td>ROC Curve</td>
<td>67</td>
</tr>
<tr>
<td>4-6</td>
<td>Confusion Matrix</td>
<td>69</td>
</tr>
<tr>
<td>4-7</td>
<td>The Variation of Risk Across Quintiles</td>
<td>69</td>
</tr>
<tr>
<td>4-8</td>
<td>Feature Weights</td>
<td>70</td>
</tr>
<tr>
<td>5-1</td>
<td>Dividing the Problem into Multiple Tasks</td>
<td>76</td>
</tr>
<tr>
<td>5-2</td>
<td>The Variation in the Relative Importance of Features Over Time</td>
<td>79</td>
</tr>
</tbody>
</table>
# List of Tables

1.1 Thesis Outline ...................................................... 23

2.1 Patient Locations .................................................. 31

3.1 Study Population Characteristics ................................. 40
3.2 Curated and EMR Variables ....................................... 43
3.3 Results - EMR vs. Curated ........................................ 46

4.1 Time-Varying and Time-Invariant Variables .................... 57
4.2 Temporal Features .................................................. 61
4.3 Results .............................................................. 67

5.1 Time-Varying and Time-Invariant Variables .................... 74
5.2 The Variation in the Relative Importance of Features Over Time ... 80
5.3 Risk Factors ......................................................... 81

6.1 Available Training Data ............................................ 91
6.2 Comparing Populations Across Hospitals ....................... 92
6.3 Differences in Feature Spaces ..................................... 93
6.4 Outline of Experiments ............................................. 97
6.5 Results .............................................................. 99

7.1 Net Financial Gain .................................................. 117

A.1 Relevant Tables in the Database ................................. 129
A.2 Patient Locations .................................................. 134
Chapter 1

Introduction

Over recent years, there has been an enormous growth in 1) our capacity to gather clinically relevant data and 2) the availability of such datasets. The collection of these data provides healthcare workers with a more complete picture of a patient’s health. In particular, the proliferation of electronic medical records (EMRs) holds out the promise of using machine learning and data mining to build models that will help improve patient outcomes. However, transforming patient data into actionable knowledge presents a barrage of pragmatic and technical challenges. But if we are successful in addressing these challenges, the knowledge embedded in these data has the potential to revolutionize clinical medicine.

One way in which these data can have an impact is through the development of accurate data-driven models for predicting avoidable bad patient outcomes. The hypothesis is that these data contain generalizable information that can help accurately identify a patient’s future pathological states. If pathologies are predicted far enough in advance, then it may be possible for healthcare workers to intervene. Such targeted interventions could, in turn, lead to better patient outcomes.

In recent years, there has been a significant amount of research effort devoted to using clinical data to predict patient outcomes [1–9]. In this dissertation, we focus on the specific task of predicting which patients in a hospital will acquire an infection with Clostridium difficile (C. difficile), a potentially avoidable bad outcome. C. difficile is a type of bacteria that takes over a patient’s gut when normal flora get
wiped out (often from receipt of antimicrobials). Infection with *C. difficile* can lead
to severe diarrhea and intestinal diseases (e.g., colitis), or even death. The infection
is often treated with specific antimicrobials: oral vancomycin and metronidazole (and
less frequently, fidaxomicin). However, it is estimated that approximately 20% of
cases relapse within 60 days [10]. The incidence of infection with *C. difficile* in the
US is estimated at 200,000 cases per year [11]; this is on par with the number of new
cases of invasive breast cancer discovered each year in the US [12].

Infection with *C. difficile* is a type of healthcare-associated infection (HAI). HAI s
are a serious problem in healthcare facilities in the US. It is estimated that, on
any given day, HAI s affect approximately 1 in every 25 inpatients in US acute care
hospitals [13]. In addition to *C. difficile*, other common HAI s include ventilator-
associated pneumonia, surgical site infection, and infections with methicillin-resistant
*Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus* (VRE).

Though many risk factors are well-known (e.g., healthcare-associated exposure,
age, underlying disease, etc.), HAI s continue to be a significant problem throughout
the world [14]. In recent years there have been numerous articles citing our inability
to prevent HAI s [15–17]. We hypothesize that one of the reasons why is because we
lack an effective clinical tool for accurately measuring patient risk.

In this work, we chose to focus on infections with *C. difficile* since it is one of
the most prevalent HAI s, and one of the most difficult to cure [15]. In our data,
infection with *C. difficile* is diagnosed based on the results of a stool test recorded in
the patient’s laboratory results.

We take a data-centric approach to the problem of developing a predictive model.
We leverage the contents of EMRs from over 100,000 patient admissions to three
different hospitals, “Hospital A”, “Hospital B” and “Hospital C” (ordered by size),
within the same hospital network. These clinical data contain information regarding:
medications, procedures, locations, healthcare staff, lab results, measurements of vi-
tals, demographics, patient history and admission details. We seek a mapping from
this information describing a patient to a value estimating the patient’s probability
of acquiring an infection. Such a risk-stratification model aims to order patients from
lowest to highest risk based on the contents of the EMRs.

Clinical guidelines have been proposed for identifying when a patient is likely to test positive for *C. difficile* during their hospitalization [18]. Such guidelines are based largely on symptoms associated with already having an infection, and thus are not useful for predicting in a proactive manner when a patient will become infected. In contrast, risk-stratification models aim to identify patients at high risk of acquiring an infection in the future.

Automated patient risk stratification, based on the contents of the patient’s EMR, can serve several purposes. Firstly, risk-stratification models can help clinicians match high-risk patients with the appropriate interventions or therapies. Secondly, data-driven models can help generate hypotheses regarding potential risk factors, in turn improving our understanding of the disease. Thirdly, such models could aid in designing more efficient clinical trials by identifying a study population at higher risk for disease, increasing the fraction of patients expected to test positive in the trial. This could significantly reduce the cost of a clinical trial without compromising the statistical power of the study.

Learning accurate risk-stratification models from EMR data presents a number of technical challenges:

- **Real Data:** The data are real clinical data; these data do not come from a curated dataset, and therefore are plagued with inconsistencies, incorrect entries, and missing values.

- **High dimensionality:** There can be thousands of binary variables representing each day of patient admission. It is likely that many of these variables affect a patient’s risk of acquiring *C. difficile*.

- **Temporal aspect:** The data are both time-invariant (e.g., a patient’s gender) and time-varying (e.g., the location of the patient within the hospital and the treatments given). These time-varying data suggest that as a patient spends time in the hospital, his/her actual risk of acquiring *C. difficile* will vary. It
is not obvious how to best incorporate these temporal dependencies into the model.

- **Paucity of cases:** Because bad outcomes are relatively rare events, there is often a paucity of positive cases to learn from.

- **Institutional differences** There are institutional differences to contend with. Institutional differences often imply that a one-size-fits-all model is not the best solution.

- **Lack of ground truth:** Once a patient is discharged from the hospital, if that patient does not return to a hospital within the network, we cannot know what happened to that patient and are forced to make assumptions about that patient’s outcome. Moreover, we are trying to estimate something that is unmeasurable (i.e., daily patient risk).

Several recent efforts have focused on building models for identifying patients at high risk of acquiring an infection with *C. difficile* [19–22]. In previous work, risk-stratification models for *C. difficile* have been developed using a small number of risk factors selected by clinical experts. Many of the risk factors considered have pertained to time-invariant features (e.g., patient history). Dubberke *et al.* were the first to consider a risk prediction model based on both variables collected at the time of admission and throughout the admission [20]. In their work, patient risk can be calculated online during the hospital stay. However, they ignore any trend in patient risk, and consider only time-invariant models. Moreover, they evaluate their model at only a single point in time.

We take a novel data-driven approach to patient risk stratification for *C. difficile*. Our final hospital-specific model produces a daily risk estimate that incorporates previous estimates of patient risk. This estimate is based on thousands of binary variables, both time-varying and time-invariant. In addition, our final model does not remain static but changes over the course of a patient admission, incorporating the changing relative importance of risk factors over time. In a final, comprehensive
set of evaluations we show how our model performs applied to held-out data. Unlike previous work, we evaluate how our model performs when applied to each day of a patient’s admission rather than a single point in time. This evaluation scheme is more representative of how the model will be applied in practice.

We have integrated our algorithm into the health record system at a large hospital in the US. Each day, the system calculates a probability for every adult inpatient, estimating his/her risk of acquiring an infection with *C. difficile*. Through a cost-benefit analysis, we illustrate how the selective targeting of high-risk patients (identified by our model) with specific interventions could lead to changes in clinical practice and ultimately a reduction in the incidence of *C. difficile*.

The contributions of this dissertation are of two kinds. In addition to developing a clinically useful tool for identifying patients at risk of developing *C. difficile*, we have made several contributions to the fields of machine learning and data mining. Much of the innovation has come in the form of feature engineering and problem formulation. Oftentimes when working in applications, these aspects are more important than the underlying classification method. The contributions of this dissertation are as follows:

- **Novel representations of EMR data incorporating both time-varying and time-invariant data.** We represent each day of a patient admission as a high-dimensional feature vector, composed of both variables collected once at the time of admission, and those that continue to be measured over the course of the hospitalization. In addition, we incorporate many hospital-specific features into our analysis (e.g., locations within the hospital and healthcare staff) that have been previously ignored. Finally, we develop a novel method for estimating the colonization pressure (i.e., patient exposure to the disease).

- **The demonstration of the benefit of taking a data-driven approach to building predictive models for *C. difficile*:** We compare classifiers learned on high-dimensional feature vectors automatically extracted from the EMR, with ones learned on a small number of known risk factors. Applied to the same held-out validation set, our high-dimensional representation of the data yields a signif-
significant improvement in the area under the receiver operating curve (AUROC). More concretely, for the same sensitivity (TPR=0.5), when evaluated on a year's worth of held-out data our approach leads to over 3,000 fewer misclassified patients compared to an approach based on a small number of variables identified by clinical experts.

- The reformulation of the problem as a time-series classification problem. We present an approach for patient risk stratification in which we incorporate the temporal changes in patient risk. We further distinguish high-risk patients from low-risk patients based on patterns in patient risk. Compared to the current approach employed in the clinical literature, our methods lead to a significant improvement in patient classification.

- The development of a novel approach for incorporating the changing relative importance of risk factors over time. We further extend our risk-stratification model to incorporate time-varying coefficients, by adapting techniques from multitask learning. We split the problem into multiple separate tasks, and learn these tasks jointly. The result is a shared component for each task from which the model is allowed to deviate, depending on the task-specific data. Compared to the our previous static model, this approach leads to consistently better performance, especially on patients with longer visits.

- The development of a novel framework for effectively incorporating data from different but overlapping feature spaces. We illustrate our approach in terms of the transfer of knowledge across three different hospitals within the same hospital network. We illustrate how current global approaches that ignore target-specific features can fail, while standard hospital-specific approaches often suffer from a lack of learning examples. To address these issues, we propose an approach that considers data from multiple hospitals while seeking a solution in the target feature space.

- The development and use of a more accurate evaluation scheme. We develop
an evaluation scheme based on the clinical use case in which patient risk is measured daily. Instead of evaluating the model only a constant number of days before an index event, as is typically done, we consider predictions made throughout the visit when measuring classification performance. In addition, we evaluate our models on data held-out temporally, i.e., patient admissions in the test set occur after all admissions in the training set. This approach to validation yields a predicted performance that is representative of how we expect the model to perform in practice.

The contributions of this dissertation are in the specific context of building risk-stratification models for predicting infections with \textit{C. difficile}. However, we believe the contributions extend to other problems in which: the data are high dimensional, or the data are time series, or the model must incorporate a temporal dependence, or where there are multiple tasks with different but overlapping feature spaces.

The remainder of this dissertation is organized as follows. In Chapter 2, we present a brief background on the EMR data we used in our analysis and the precise learning task. The next four chapters present separate but related pieces of research, each tackling specific challenges with respect to building risk-stratification models for \textit{C. difficile}. We do not present this work in an order identical to how the research was actually done, but in an order that builds on work from the previous chapter. Chapters 3 through 5 build in this way, leading to what we refer to as our "final" model in Chapter 5.

Each chapter investigates a different aspect of the problem. Therefore, the precise prediction task in each chapter changes to fit the specific research question. In addition, because this work took place over the course of a three year period the study populations vary. We updated our study populations over time to reflect changes in data collection at the hospitals. We summarize these differences in Table 1.1. Not only do the study populations vary but so do our evaluation techniques. Over time, we developed what we believe to be an evaluation scheme that best represents how we expect a model to perform in practice.

In Chapter 3, we explore the benefit of our proposed data-driven approach over
the expert-driven approach for feature extraction. In Chapter 4, we reformulate the problem as a time-series classification task, with the goal of identifying patterns of risk that are more likely to lead to worse outcomes. In Chapter 5, we present our multitask approach to building dynamic risk models, and compare the result with static risk models. In Chapter 6, we examine the transferability of data across different hospitals and present our learning approach to dealing with different but overlapping feature spaces. In Chapter 7, we discuss the integration of our final model into the health record system at the largest hospital, and examines how different interventions could be applied, guided by the risk tool, to reduce patient suffering. Finally, Chapter 8 summarizes the contributions of this dissertation, and presents ideas for future work in the area of data-driven medicine.
Chapter 3
Research Goal: We explore the benefit of building predictive models using the entire structured contents of the EMR.
Task: We make a single prediction for each patient 24 hours after admission.
Study Population: Patients admitted to Hospital C between April 2011 and April 2013, with a visit of at least 24 hours.

Chapter 4
Research Goal: We measure the evolution of predicted risk over the course of a hospitalization, and investigate how trends in daily risk could be used to improve predictions.
Task: We make a prediction for each day of a patient's admission and then use these predictions to update the current prediction. We evaluate the discriminative power of each model 2 days before the index event.
Study Population: Patients admitted to Hospital C between Jan 2010 and Dec 2010, with a visit of at least 7 days.

Chapter 5
Research Goal: We extend the model presented in Chapter 4 to include time-varying coefficients.
Task: We make multiple predictions for each patient, one for each day of a patient's admission. We evaluate the model throughout the entire visit (up to a positive test result, or discharge).
Study Population: Patients admitted to Hospital C between April 2011 and April 2013, with a visit of at least 3 days.

Chapter 6
Research Goal: We investigate the transferability of predictive models across different hospitals.
Task: We make a single prediction for each patient at the time of admission.
Study Population: Patients admitted to Hospital A, Hospital B, or Hospital C between April 2011 and April 2013.

Table 1.1: In Chapters 3-6 of this dissertation, we explore a different aspect of building models for predicting infections with C. difficile. We define the risk-stratification task and the study population based on the precise research goal.
Chapter 2

Background

2.1 Introduction

Healthcare is currently undergoing a change in record keeping, as hospitals transition from paper-based records to electronic medical records (EMRs). A basic EMR\textsuperscript{1} system includes the electronic capture of patient demographics, clinical notes, laboratory results, radiology reports, and medication orders. Transitioning from paper-based records to electronic records promises higher quality and more efficient care [23].

In 2008, the adoption rate for EMRs among acute care general medical and surgical hospitals in the US was approximately 9% [24]. Since then, the number of hospitals using EMR systems has increased dramatically. In 2012, a comprehensive survey concluded that 44% of acute care hospitals have at least a basic EMR system, but only 5% met the criteria of a comprehensive system [24].

These low adoption rates may be due to the fact that EMR systems are a long-term investment. Some physicians argue that cost savings can only be expected 10 years after the change [25]. This narrow perspective focuses on the savings that come with decreased storing and sorting of paper charts and ignores the potential of ‘Big Data’ in medicine. Many critics of electronic record systems fail to recognize the critical role EMR data can play in improving patient outcomes.

\textsuperscript{1}In the literature electronic health record (EHR) and electronic medical record (EMR) are often used interchangeably [23], here we use only EMR.
In addition to the high adoption cost, hospitals are wary of the lack of standards and structured data definitions in electronic health record systems [23]. At this point there are several widely deployed health information system solutions on the market (e.g., Epic, Amalga, and Cerner). The wide selection makes it difficult for hospitals to choose the right solution. It is not uncommon for a hospital to have an EMR system composed of software from more than one vendor. Unfortunately, this introduces a challenge for researchers when trying to work with EMR data at even a single hospital. When working with data across hospitals these challenges become compounded.

In this dissertation we have encountered these challenges firsthand. While our work considers only a single hospital network the variability in the EMR across hospitals and time was significant. The hospital network employs both Amalga and Cerner (MedConnect). Patient data are transferred across the two systems (e.g., Amalga receives Cerner laboratory data). For our work, we focused on EMR data stored in the Amalga system. During the time period for which we conducted our research, Amalga was updated significantly. Not only did the names of certain tables change, but so did the contents. We incorporated these changes into our work as they arose. Going forward, it will be important that algorithms and models built using EMR data are designed with this sort of flexibility in mind, since health information systems will continue to evolve.

2.2 The Data

As mentioned above, in this dissertation we use EMR data from a single hospital network. The Institutional Review Board of the Office of Research Integrity of the hospital network approved the statistical analysis of retrospective medical records. The organization operates ten hospitals in the Baltimore-Washington metropolitan area of the US. In our work we consider three different hospitals within this network: Hospital A, Hospital B and Hospital C. The majority of our work focuses on building predictive models for Hospital C, the largest of the three hospitals. Hospital C is a large teaching and research hospital; with 926 beds, it is among the 50 largest hospitals.
in the US. In Chapter 6, we provide more details about the other two hospitals, and how the three hospitals compare in terms of patient populations.

We consider inpatient visits between 2010 and 2013. Each chapter in this dissertation focuses on a slightly different problem, and therefore the study population changes. Additionally, overtime we updated the study population to reflect changes in the data collection and storage at the hospital. In each chapter we give more detail regarding the precise study population and the exclusion criteria applied.

In the EMR we have data pertaining to admission details, patient demographics, laboratory results, diagnoses, medications, patient locations, vitals, and procedures. These data are stored across several different tables in several different databases, described in detail in Appendix A.

In the EMR, each patient visit (or admission) is represented by a unique identifier an “EID”, and each patient is associated with a unique identifier an “OID”. These identifiers allows us to retrieve information across hospital databases for each admission. In addition, using these identifiers we can retrieve information about all previous visits within the network for a patient.

In the paragraphs that follow we briefly describe how relevant patient data are represented in the EMR. In Section 2.5 we describe in more detail how we map this information to a feature vector representing each day of a patient admission.

**Admission details:** Details of the admission (e.g., date and time of admission, date and time of discharge, and type of visit) are stored in a single table. This table also contains other information pertaining to the admission such as the financial class code, the source of the admission, the hospital service, and the attending doctor. We extract all this information for each visit included in our analysis.

**Patient demographics:** Information pertaining to patient demographics such as date of birth, gender, race, marital status, and city of residence are stored in a single table. Aside from the date of birth, all data in this table are categorical. Patient age at the time of admission is calculated using the admission date and the date of birth.
Laboratory results: Results pertaining to all ordered/observed laboratory tests are recorded in the EMR. Each laboratory test is associated with a unique identifier. Figure 2-1 gives an example of a row in the laboratory table. Each row in the table is associated with a patient admission, an observation identifier, an observation value, and an observation time. Each laboratory test is also associated with a reference range (e.g., 120-200 for cholesterol). If the observed value lies outside the normal range for that measurement, an abnormal flag is entered. Abnormal flags are either H=high, L=low, C=critical, or empty=normal. These flags are coded based on the reference ranges (defined by experts) in the health record system. When extracting laboratory results for patients we extract the observation identifier and the flag associated with the observation, as shown in Figure 2-1.

<table>
<thead>
<tr>
<th>ObservationIdentifier</th>
<th>ObservationName</th>
<th>ObservationValue</th>
<th>ReferenceRange</th>
<th>AbnormalFlags</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREAT</td>
<td>Creatinine</td>
<td>2.03</td>
<td>0.52-1.04</td>
<td>H</td>
</tr>
</tbody>
</table>

Figure 2-1: We represent each laboratory result by a concatenation of the observation identifier and the flag associated with the result.

Diagnoses: Patient diagnoses are encoded using ICD-9 codes. ICD-9 is a coding system developed by the International Statistical Classification of Diseases and Related Health Problems that encodes diseases (and procedures, as discussed later) hierarchically (see Figure 2-2). At the highest level ICD-9 codes fall into 1 of 19 categories [26]. Each row of the diagnoses table has an EID entry and an ICD-9 code. Since patient visits can be associated with multiple ICD-9 codes, there may be multiple rows corresponding to the same visit. In our data the average visit (including outpatient visits) is associated with two distinct ICD-9 codes. ICD-9 codes, widely used for billing purposes, can get coded well after a patient is discharged [27]. For this reason, we do not use the codes associated with a patient’s current visit in our model. Instead, we consider only the codes
from a patient's most recent previous hospital visit. In Section 2.5, we further describe how we preprocess ICD-9 codes.

001-139 Infectious and Parasitic Diseases
008 Intestinal Infectious Diseases
008.4 Intestinal infectious due to other organisms
008.45 Intestinal infection due to Clostridium difficile

Figure 2-2: Patient diagnoses are encoded using a hierarchical international disease classification system. This example illustrates the hierarchy for infection with Clostridium difficile.

Medications: Orders for medications are entered along with a visit identifier (EID), an 8-digit medication identifier, and a start/stop time. Figure 2-3 shows example entries (note: the EID is not shown here for privacy reasons). Each medication identifier is associated with a medication, a dosage and a form (e.g., in solution) as indicated by the “GiveCodeText” field in Figure 2-3. For example, in Figure 2-3 acetaminophen is represented using three different medication identifiers, depending on the dosage and the form. Since the dosage and form are encoded in the 8-digit medication identifier, we represent patient medications using only this identifier.

<table>
<thead>
<tr>
<th>EID</th>
<th>GiveCodeID</th>
<th>GiveCodeText</th>
<th>StartDateTime</th>
<th>EndDateTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>187</td>
<td>63622973</td>
<td>vancomycin 1.25 gm/250 mL 0.9% NaCl</td>
<td>2011-09-26 06:00:00.000</td>
<td>2011-09-26 05:57:14.000</td>
</tr>
<tr>
<td>188</td>
<td>63622245</td>
<td>cefAZolin 2 gr/100 mL 0.9% NaCl</td>
<td>2011-09-26 06:00:00.000</td>
<td>2011-09-26 12:27:01.000</td>
</tr>
<tr>
<td>190</td>
<td>63718894</td>
<td>bicacodyl 10 mg Supp</td>
<td>2011-09-27 18:00:00.000</td>
<td>2011-09-27 12:27:01.000</td>
</tr>
<tr>
<td>191</td>
<td>63610689</td>
<td>dextrose 5%/0.45% NaCl 1000 mL</td>
<td>2011-09-27 18:00:00.000</td>
<td>2011-09-27 12:27:02.000</td>
</tr>
<tr>
<td>192</td>
<td>63624621</td>
<td>insulin reg 100 unit/100 mL 0.3% NaCl</td>
<td>2011-09-28 12:27:00.000</td>
<td>2011-09-29 11:33:57.000</td>
</tr>
<tr>
<td>193</td>
<td>63623565</td>
<td>fentanyl 5.00 mcg/250 mL NaCl 0.9%</td>
<td>2011-09-28 12:27:00.000</td>
<td>2011-09-29 11:33:57.000</td>
</tr>
<tr>
<td>194</td>
<td>63605026</td>
<td>sodium chloride 0.5% 250 mL</td>
<td>2011-09-28 12:28:00.000</td>
<td>2011-09-29 11:33:58.000</td>
</tr>
<tr>
<td>195</td>
<td>63615984</td>
<td>Lactated Ringers 1000 mL</td>
<td>2011-09-29 12:28:00.000</td>
<td>2011-09-29 11:33:57.000</td>
</tr>
<tr>
<td>196</td>
<td>63604920</td>
<td>sodium chloride 0.45% 1000mL</td>
<td>2011-09-29 12:28:00.000</td>
<td>2011-09-29 11:33:57.000</td>
</tr>
<tr>
<td>197</td>
<td>63715775</td>
<td>acetaminophen 50 mg Supp</td>
<td>2011-09-29 12:28:00.000</td>
<td>2011-09-29 11:33:56.000</td>
</tr>
<tr>
<td>198</td>
<td>63700108</td>
<td>acetaminophen 325 mg Tab</td>
<td>2011-09-29 12:28:00.000</td>
<td>2011-09-29 11:33:56.000</td>
</tr>
<tr>
<td>199</td>
<td>63632503</td>
<td>acetaminophen 650 mg/20 mL Soln</td>
<td>2011-09-29 12:28:00.000</td>
<td>2011-09-29 11:33:56.000</td>
</tr>
<tr>
<td>200</td>
<td>63601477</td>
<td>ondansetron 4 mg/2 mL lirio</td>
<td>2011-09-29 12:28:00.000</td>
<td>2011-09-29 11:33:57.000</td>
</tr>
</tbody>
</table>

Figure 2-3: This is an excerpt from the medication table showing patient medication orders. Each row entry corresponds to a new medication order associated with an individual visit (EID). The ordered medication is identified by a unique 8-digit code (GiveCodeID).

We note that in early 2013 the coding system for medications changed to a
“PYXIS” coding, from an “AHFS” coding. This did not affect the methods we used, nor did it affect our how we represented medications. However, it is important to be aware of such changes, since if one learns a model based on one encoding scheme it will not automatically translate to data encoded with another scheme. To this end, we briefly explored the use of topic modeling to automatically learn classes of drugs based on the drug names alone [28]. Such an abstraction could prove useful if the amount of training data is limited, or if a coding system is not used.

**Locations:** Within the hospital, locations are represented as units and rooms. A unit can contain multiple rooms, and each room can contain multiple beds. The size of the units vary; some units contain no beds (e.g., operating rooms), whereas others may contain up to 30 beds (e.g., a patient care unit). For each hospital admission we have timestamped location data. Location data refer to the patient’s location within the hospital. Locations are collected at both the unit and the room level. Table 2.1 shows how we can trace a patient’s path through the hospital using the timestamped location data (note: in this table the dates and times were changed for de-identification purposes). From this we can infer when patients were co-located within a unit in the hospital, or who was in the room prior to that patient. In our data each location is represented by a separate binary variable. We incorporate this knowledge about a patient’s location into the model, but more importantly we use this information to extract information regarding patient exposure to *C. difficile*. This is further discussed in Section 2.4.

**Vitals:** Patient vitals (e.g., respiratory rate and heart rate) are encoded similarly to laboratory results. Each entry in the vitals table corresponds to a visit (EID), an observation identifier (e.g., “BPSYSTOLIC” for systolic blood pressure), an observation value, a reference range, an abnormal flag, and an observation date time. When extracting information about vitals for a patient we encode the observations the same way we encode laboratory results, i.e., as a concatenation
Table 2.1: The information contained in the EMR allows us to follow a patient’s physical trajectory through the hospital as he/she moves from room to room.

<table>
<thead>
<tr>
<th>Location Unit, Room</th>
<th>Time In</th>
<th>Time Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiac Cath Unit, 4Axx-P</td>
<td>8/15/03  13:19</td>
<td>8/15/03 22:40</td>
</tr>
<tr>
<td>Medicine Patient CU, 4NxxE</td>
<td>8/15/03  22:40</td>
<td>8/17/03 10:10</td>
</tr>
<tr>
<td>Main OR, MRxx-P</td>
<td>8/17/03 10:10</td>
<td>8/17/03 12:15</td>
</tr>
<tr>
<td>Cardiac Intensive CU, CRxx-P</td>
<td>8/17/03 12:15</td>
<td>8/18/03 10:56</td>
</tr>
<tr>
<td>Surgical Patient CU, 4Fxx-B</td>
<td>8/18/03 10:56</td>
<td>8/24/03 15:37</td>
</tr>
<tr>
<td>Surgical Patient CU, 4Fxx-A</td>
<td>8/24/03 15:37</td>
<td>8/28/03 9:14</td>
</tr>
</tbody>
</table>

of the observation identifier and an abnormal flag (e.g., “BPSYSTOLIC_H” for high systolic blood pressure).

Procedures: In the EMR, procedures are encoded using both Current Procedural Terminology (CPT) codes and ICD-9 procedure codes. Each row entry in the procedures table records a procedure, the corresponding visit and a procedure date and time. For some patients only one of the coding systems is used, for others both coding systems are used. Since both coding systems are used to describe procedures, in our analysis we consider both CPT and ICD-9 codes (however, we do not learn a mapping between them). We discuss how we represent procedure data in Section 2.5.

For each patient admission in our study population we extract knowledge pertaining to the EMR data described above, in addition to a binary label indicating whether or not the patient became infected with *C. difficile* during their hospital visit. The next section describes how we identify these cases and gives some additional background on the disease.

2.3 Identifying Positive Cases for *C. difficile*

*C. difficile* is an anaerobic gram-positive, spore-forming organism that produces two exotoxins (toxin A and toxin B). Infection with *C. difficile* is one of the most common causes of colitis (inflammation of the colon). In severe cases, *C. difficile* infection
can lead to toxic megacolon (i.e., life-threatening dilation of the large intestine) and eventually death.

*C. difficile* is transmitted from patient to patient through the fecal-oral route. Typically, patients become colonized after exposure to the organism and some disruption of normal gut flora (e.g., the receipt of antimicrobials). Once infected patients typically suffer from diarrhea, lower abdominal pain, fever and leukocytosis [29]. However, not all patients colonized with *C. difficile* become symptomatic. It is estimated that 20% of hospitalized adults are asymptomatic carriers of the disease [30].

The diagnosis of *C. difficile* infection requires both the presence of moderate to severe diarrhea and a positive stool test for toxigenic *C. difficile*. Details regarding testing protocol at the hospitals are given in Appendix B. For our analysis, time-stamped results of stool tests for toxigenic *C. difficile* were obtained from the laboratory database.

Current guidelines discourage repeated testing (if the initial result is negative for toxigenic *C. difficile*) and testing for cure [31]. This has important ramifications regarding ground truth. Some patients may continue to exhibit symptoms consistent with the disease despite a negative test result. However, we expect the number of false negatives to be low given the high sensitivity of the testing protocol e.g., 100% (95%CI 89.6%-100%) [32].

Many of the variables we extract for each patient visit (e.g., medications and diagnoses), are related to a patient’s underlying susceptibility to *C. difficile*. While necessary, susceptibility alone is not sufficient for infection with *C. difficile*. To become infected, a patient must also be exposed to the disease. We define how we measure patient exposure in the next section.

### 2.4 Colonization Pressure

Colonization pressure aims to measure the number of patients in a unit or hospital colonized or infected with a particular disease. Bonten *et al* were the first to study the relationship between colonization pressure in a medical intensive care unit (MICU)
and the acquisition of the healthcare-associated infection vancomycin resistant Enterococcus (VRE) [33]. They defined colonization pressure as the number of patients believed to have the disease on a given day divided by the number of patients treated in the MICU on that day. As highlighted by [34] there is considerable heterogeneity in how colonization pressure is measured among medical researchers. In some studies, in which disease is tested for on a daily basis, colonization pressure is simply defined as the daily fraction of patients diagnosed with disease through a positive lab result. In other studies, in which patients are not tested daily, patients contribute to the colonization pressure for several days after a positive laboratory test result [35]. A recent study measured the effect of colonization pressure on risk of C. difficile infection in severely ill patients [36]. The authors calculated the colonization pressure for every susceptible patient as the sum of the number of infected patients over each day spent in the ICU. The authors assumed infected patients contributed to the colonization pressure for 14 days after the initial positive stool sample. In all of these studies patients make a constant contribution to the colonization pressure over a certain time period. That is, the contribution each patient makes to the colonization pressure looks like a rectangular pulse in time.

In our work, we consider a slightly more complex notion of colonization pressure. In our analysis, a patient \( p \) makes a contribution to the colonization pressure, \( CPP(p, t) \), on day \( t \). This contribution depends on when the patient tests positive for the first and last time, \( t_f \) and \( t_l \), and when the patient is discharged from the hospital \( t_d \) (where time is measured in days from the day of admission). (Note that \( t_p \) and \( t_d \) depend on the patient, \( p \)). While the patient continues to test positive he or she contributes a constant amount to the colonization pressure. After the last positive test result (which is often the first positive test result, since testing for a cure is not recommended) a patient contributes to the colonization for no more than 14 days. During this time period, the patient is assumed to be receiving treatment (for the infection) or in isolation, and therefore we assume a linearly decreasing relationship between the patient's contribution to the colonization pressure and time. This function is summarized by Equation 2.1, and two examples are plotted in Figure 2-4.
A patient’s contribution to the colonization pressure is constant while the patient continues to test positive and decreases linearly after a positive test result, since the patient is assumed to be in isolation or undergoing treatment.

\[
CPP(p, t) = \begin{cases} 
1 & t \in [t_f, t_l] \\
- \frac{t}{14} + \frac{(t_i + 14)}{14} & t \in [t_l, \min(t_d, t_l + 14)] \\
0 & \text{otherwise}
\end{cases}
\]  

(B.1)

Because we have timestamped locations for each patient, we can calculate colonization pressure for each unit \( u \), as \( CPU(u, t) \). Equation 2.2 defines this quantity. The colonization pressure of a unit depends on each patient’s contribution to the colonization pressure on that day, \( CPP(p, t) \) and each patient’s length of stay in that unit on that day, in terms of the number of hours, \( LOS(u, p, t) \).

\[
CPU(u, t) = \sum_p CPP(p, t) * \frac{LOS(u, p, t)}{24}
\]  

(B.2)

When extracting the relevant unit-wide colonization pressure for a new patient on a given day, we sum the \( CPU(u, t) \) across all units in which that patient spent any time, i.e., \( \sum_{\{w:LOS(u, p, t) > 0\}} CPU(u, t) \). In contrast, the hospital-wide colonization pressure is calculated as \( \sum_u CPU(u, t) \). We calculate the hospital wide colonization pressure by summing across all units. As a result, the unit colonization pressure varies across patients for a given day, while the hospital wide colonization pressure does not.
Using the EMR data, we describe each patient visit in terms of exposure (e.g., colonization pressure), susceptibility (e.g., underlying disease), and outcome (e.g., whether or not the patient tested positive for toxigenic \textit{C. difficile}). In the next section, we describe how we preprocess all of these data once extracted from the databases.

### 2.5 Preprocessing the Data

The relevant EMR data are stored across several transact-SQL databases (described in detail in Appendix A). The data can be broadly grouped into two categories: time-invariant variables, and time-varying variables. Time-invariant variables are extracted \textit{once} for each patient admission, these include admission details (e.g., date and time of admission), diagnoses from the previous visit, and patient demographics (e.g., marital status). Time-varying variables are extracted for each \textit{day} of a patient admission, these include laboratory results, medications, patient locations, vitals, and procedures.

In our work, we chose to focus on events at the temporal resolution of a day, despite the fact that data are entered at different rates (e.g., vitals may get updated hourly, while medications are updated only twice a day). By discretizing time at the resolution of a day, we do not consider the order of events within a day. For our purposes this resolution suffices (though it may not be optimal), however it may not be suitable for all applications.

In our analysis, extracted data fell into one of the three categories: numerical, categorical or \textit{list feature}. Numerical features included variables like number of hospital visits in the past 90 days. Categorical features refer to features like financial class code. List features refer to variables for which a patient may have multiple values for a single day (e.g., medications). In the extracted data we represent each patient day as a single row, thus variables for which there might be multiple entries are represented by a list of observations. During the preprocessing stage, we parse these lists, and build global dictionaries mapping each possible observation to a unique binary
feature. This mapping is partly responsible for the high dimensionality of the feature space.

The majority of the mapping is data-driven; meaning the mapping from observation to binary feature is not predefined but determined directly from the data or from cutoffs based on the data (e.g., quintiles). The only mappings we hand code are age and diagnoses. We predefine the bins representing age because we want features corresponding to age to transfer readily across datasets. We map each diagnosis code to one of 20 binary features (19 features representing the highest level of the ICD-9 hierarchy and one representing no diagnoses). We consider only the highest level of the diagnoses codes to avoid a representation that is too sparse.

In the ICD-9 classification system there are 13,000 unique codes. For our application, this is quite high. Soon (if not already) hospitals will transition from ICD-9 to ICD-10, a coding system with 68,000 unique codes. This poses a problem to researchers working with EMR data, and has been the focus of some research in building useful abstractions [37]. This problem of abstraction is important and one that arises in other applications. However, we should remind the reader that ICD-9 codes are recorded for billing purposes. Depending on the billing protocol at the hospital, there can be large discrepancies between what is reported by the ICD-9 codes and the actual health of a patient. For this reason, ICD-9 codes alone should not be relied upon when building predictive models from EMR data.

As the authors of [38] state, there are many challenges that come with working with EMR data in research. Addressing these challenges requires careful consideration of the data and the use case of the work. Moreover, electronic health information systems will continue to change and therefore it is important that researchers take this into consideration when developing models based on EMR data. In our work, these changes motivated our simple, flexible, data-driven approach to extracting and representing the EMR data.
Chapter 3

Leveraging the Richness of the Electronic Medical Record

3.1 Introduction

Throughout this dissertation, we consider the task of predicting the risk that an inpatient will acquire an infection with *Clostridium difficile* (*C. difficile*). In this chapter, we make a single prediction for each patient admission 24 hours after the time of admission. We explain this choice later on in the chapter.

There have been several recent efforts on building prediction rules for identifying patients at high risk of *C. difficile*. While the notion of estimating risk may seem intuitively obvious, there are many different ways to define the problem. The precise definition has important ramifications for both the potential utility of the estimate and the difficulty of the problem. Reported results in the medical literature for the problem of risk stratification for *C. difficile* vary greatly, AUROC=0.63-0.88 [19–22]. The variation in classification performance arises from differences in the task definitions, the study populations, and the methods used to generate and to evaluate the predictions. These differences render simple comparisons among models based on performance measures uninformative. Thus, in a review of prior work, we shall focus on reported methodology rather than on reported performance.

Tanner et al., tested the ability of the Waterlow Score to risk stratify patients at
the time of admission for contracting *C. difficile* [19]. The Waterlow Score is used in the UK for predicting a patient's risk of developing a pressure ulcer. The score considers 10 variables available at the time of admission (build/weight for height, skin type/visual risk areas, sex and age, malnutrition, continence, mobility, tissue malnutrition, neurological deficit, major surgery or trauma). In [39], Dubberke et al., identify several key risk factors for *C. difficile* infection (CDI). More recently, the authors developed and validated a risk prediction model based on both variables collected at the time of admission and during the admission, to identify patients at high risk of CDI [20]. The final model included 10 different variables: age, CDI pressure (i.e., colonization pressure), times admitted to hospital in the previous 60 days, modified acute physiology score, days of treatment with high-risk antibiotics, whether albumin level was low, admission to an intensive care unit, and receipt of laxatives, gastric acid suppressors, or antimotility drugs. Other investigators have considered narrower study populations. For example, Garey et al., consider only hospitalized patients receiving broad-spectrum antibiotics, a known risk factor for *C. difficile* [21]. Considering only patients with exposure to broad-spectrum antibiotics, a population already at elevated risk makes the task of risk stratification more difficult but also results in a model that is less generalizable. The investigators develop a risk index based on the presence of 5 variables, age 50-80, age>80, hemodialysis, non-surgical admission, and ICU length of stay. Krapohl et al., study risk factors in adult patients admitted for surgical colectomy [22]. The authors identify mechanical ventilation and history of transient ischemic attack as being independently associated with *C. difficile*.

In all of the work discussed above, building risk-stratification models for *C. difficile* is a two-step process. In a first step, risk factors for *C. difficile* are identified. In previous work, risk factors are identified using either logistic regression or based on previously identified risk factors drawn from the literature. This initial step typically results in the use of fewer than a dozen different risk variables. A second step corresponds to constructing the prediction rule (i.e., the weights used to combine the factors into a risk score). This methodology is appropriate if the number
of available learning examples is small, if the variables must be extracted by hand, or if the prediction rule will be calculated by hand. Yet, many hospital databases now contain hundreds of thousands electronic medical records (EMRs). These data combined with regularization techniques can be used to learn more accurate hospital-specific risk-stratification models that take into consideration thousands of variables. The methods have the potential to yield risk stratification that is custom-tailored to the distributions and nuances of individual hospitals, and the approach can then be applied automatically to the data, rendering manual calculation unnecessary.

We propose a risk-stratification procedure based on over 10,000 variables automatically extracted from the electronic medical records. Using regularization techniques, we develop the model on 34,846 admissions and validate the model on a holdout set of data from the following year, a total of 34,722 admissions. We show that the inclusion of additional information automatically extracted from the EMR can lead to a significant improvement in performance compared to a model learned on the same data using the usual risk variables. While building and using such data-driven models is more complex than using a simple rule, we argue that the accuracy and hospital specificity makes them more appropriate. Moreover, the advent of electronic health information systems provides the necessary infrastructure to automate data-driven risk methods.

3.2 The Data

We consider all patients admitted to Hospital C on or after 2011-04-12 and discharged before 2013-04-12. This yields a total of 81,519 admissions. We excluded any patients less than 18 years of age at the time of admission (8,056), admissions less than 24 hours in length (3,770), and admissions for which the patient tested positive for C. difficile in the first 24 hours of the admission (125). The final study population consisted of 69,568 unique admissions. Table 3.1 summarizes the demographic and admission-related characteristics of the study population.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female gender, (%)</td>
<td>56.72</td>
</tr>
<tr>
<td>Age (%)</td>
<td></td>
</tr>
<tr>
<td>[18-25)</td>
<td>6.36</td>
</tr>
<tr>
<td>[25-45)</td>
<td>20.87</td>
</tr>
<tr>
<td>[45-60)</td>
<td>25.23</td>
</tr>
<tr>
<td>[60-70)</td>
<td>18.74</td>
</tr>
<tr>
<td>[70-80)</td>
<td>15.37</td>
</tr>
<tr>
<td>[80-100)</td>
<td>10.37</td>
</tr>
<tr>
<td>&gt;=100</td>
<td>2.97</td>
</tr>
<tr>
<td>Hospital Admission type (%)</td>
<td></td>
</tr>
<tr>
<td>Emergency</td>
<td>58.53</td>
</tr>
<tr>
<td>Routine Elective</td>
<td>19.36</td>
</tr>
<tr>
<td>Urgent</td>
<td>12.43</td>
</tr>
<tr>
<td>Term Pregnancy</td>
<td>9.41</td>
</tr>
<tr>
<td>Hospital Admission source (%)</td>
<td></td>
</tr>
<tr>
<td>Admitted from Home</td>
<td>79.34</td>
</tr>
<tr>
<td>Transferred from another health institution</td>
<td>12.02</td>
</tr>
<tr>
<td>Outpatient</td>
<td>6.20</td>
</tr>
<tr>
<td>Other*</td>
<td>2.42</td>
</tr>
<tr>
<td>Hospital Service (%)</td>
<td></td>
</tr>
<tr>
<td>Medicine</td>
<td>45.54</td>
</tr>
<tr>
<td>Cardiology</td>
<td>12.41</td>
</tr>
<tr>
<td>Surgery</td>
<td>11.41</td>
</tr>
<tr>
<td>Obstetrics</td>
<td>10.72</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>4.21</td>
</tr>
<tr>
<td>Other**</td>
<td>15.71</td>
</tr>
<tr>
<td>Hemodialysis performed (%)</td>
<td>5.02</td>
</tr>
<tr>
<td>Diabetic (%)</td>
<td>31.46</td>
</tr>
<tr>
<td>Medications (%)</td>
<td></td>
</tr>
<tr>
<td>Immunosuppressants (solid organ transplant)</td>
<td>1.84</td>
</tr>
<tr>
<td>Corticosteroids</td>
<td>11.31</td>
</tr>
<tr>
<td>Antimicrobials assoc.</td>
<td>36.67</td>
</tr>
<tr>
<td>Antimicrobials rarely assoc.</td>
<td>18.30</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>34.92</td>
</tr>
<tr>
<td>CDI (%)</td>
<td>1.05</td>
</tr>
<tr>
<td>Median LOS in days (IQR)</td>
<td>4.01</td>
</tr>
<tr>
<td>Previous Visit in last 90 days (%)</td>
<td>21.85</td>
</tr>
<tr>
<td>History of CDI, 1-year (%)</td>
<td>1.45</td>
</tr>
</tbody>
</table>

*Other includes Routine Admission (unscheduled), Transferred from a nursing home, Referred and admitted by family physician
**Other includes Burn, Gynecology, Neurosurgery, Open Heart Surgery, Oncology, Orthopedics, Trauma, Vascular

Table 3.1: This table summarizes descriptive characteristics of the study population under consideration.
Unique admissions to WHC between 2011-04-12 and 2013-04-12: 81519

Exclude patients, age criteria (8065, 9.88%)
   Less than 18 years of age: 8065

Exclude patients, LOS criteria (3770, 4.6%)
   LOS <24 hours: 3770

Exclude patients, C.difficile criteria (125, 0.15%)
   Positive for C. difficile <24hrs of adm.: 125

Final study population: 69568 (85.34%)

Figure 3-1: During the two year period from April 2011 to April 2013, we consider all adult patients with a length of stay (LOS) of at least 24 hours, who did not test positive for *C. difficile* in the first 24 hours of the admission.

### 3.3 Materials & Methods

#### 3.3.1 Feature Extraction

We seek to predict which patients will test positive for *C. difficile* during the current hospitalization. In this chapter, we focus on measuring solely the benefit of taking a data-centric approach to feature extraction, and thus we do not consider the temporal aspects of the prediction task. How risk changes over time will be considered in later chapters. For now, we make a single prediction for each patient about his/her risk of testing positive for *C. difficile* using information collected in the first 24 hours after admission. This prediction is indicated by the red arrow in Figure 3-2. We aim to make an informed prediction as early as possible about each patient arriving at the hospital. We make a prediction at 24 hours instead of at the time of admission since we consider variables not available in the EMR until 24 hours after admission (e.g., if the patient received dialysis). In prior studies on identifying novel risk factors for *C. difficile*, patients who test positive for *C. difficile* within 48 hours or 72 hours of admission are typically excluded from the analysis [20, 21]. Here, we do not exclude these patients from the analysis since our goal is to build a classifier that applies to all patients still present in the hospital 24 hours after admission. For each patient in the study population, we have laboratory data indicating if and when a patient tested positive for toxigenic *C. difficile*. We define the risk period of a patient as
the time between admission to the time of a positive test result or discharge time if the patient never tests positive. In our study population, all patients have a risk period greater than 24 hours. In the results section, we will measure our ability to risk stratify subsets of patients (e.g., patients with a risk period of greater than 48 hours).

Figure 3-2: We define the risk period differently depending on whether or not the patient tests positive for *C. difficile*.

For each admission in our study population, we extract two sets of variables from the hospital databases:

1. Curated Variables: Well-known clinical risk factors for *C. difficile* drawn from the literature, and readily available to physicians within 24 hours of admission [21, 39–50]

2. EMR Data: All patient data extracted automatically from the EMR within 24 hours of admission

These two sets of variables are described in detail in Table 3.2. The first set of variables in the table were selected by a team of collaborating physicians and represent well-known risk factors for *C. difficile* in the literature. This list does not include *all* known risk factors since we restricted this list to variables typically available to physicians (e.g., we do not consider colonization pressure). The second set of variables, described by the categories listed in the second half of Table 3.2, is a much larger set, consisting of structured variables that are easily extracted in an
Curated variables based on well known risk factors from the literature.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_70</td>
<td>(Time of Admission - Birthday) &gt;= 70 years [21, 40, 41]</td>
</tr>
<tr>
<td>admission_source:TE</td>
<td>Transfer From Nursing Home [51]</td>
</tr>
<tr>
<td>day90_hospit</td>
<td>Recent hospitalization in the previous 90 days [21, 42]</td>
</tr>
<tr>
<td>hist_cdi</td>
<td>previous C. difficile Infection within the last year [43]</td>
</tr>
<tr>
<td>hemodialysis</td>
<td>procedure code for dialysis (&lt;24hrs of Adm.) [42]</td>
</tr>
<tr>
<td>gastro_tube</td>
<td>procedure code associated with nasogastric or esophagostomy tube (&lt;24hrs of Adm.) [48, 52]</td>
</tr>
<tr>
<td>cesteroids</td>
<td>pharmacy order entry (POE) for corticosteroids (&lt;24hrs of Adm.) [42]</td>
</tr>
<tr>
<td>immunosuppresants</td>
<td>POE for solid organ transplant immunosuppresants (&lt;24hrs of Adm.) [42]</td>
</tr>
<tr>
<td>chemo_cdi</td>
<td>POE for chemotherapeutic agents associated with CDI (&lt;24 hrs of Adm.)</td>
</tr>
<tr>
<td>chemo_entero</td>
<td>POE for chemotherapeutic agents associated with enteropathy (&lt;24hrs of Adm.)</td>
</tr>
<tr>
<td>antimicrobials_assoc</td>
<td>POE for antimicrobials frequently associated with CDI (&lt;24hrs of Adm.) [41, 44-46, 53]</td>
</tr>
<tr>
<td>antimicrobials_rarely</td>
<td>POE for antimicrobials rarely associated with CDI (&lt;24hrs of Adm.) [39]</td>
</tr>
<tr>
<td>ppi</td>
<td>POE for proton pump inhibitors (&lt;24hrs of Adm.) [47, 49]</td>
</tr>
<tr>
<td>abdominal_surgery</td>
<td>procedure codes for abdominal surgery associated with CDI (&lt;24hrs of Adm.) [41, 50]</td>
</tr>
</tbody>
</table>

Categories of additional variables extracted from the EMR.

<table>
<thead>
<tr>
<th>Variable Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>previous visits</td>
<td>statistics on previous visit (within 90 days) lengths (total, max, avg.)</td>
</tr>
<tr>
<td>dxcodes</td>
<td>highest level of ICD9 codes coded during most recent visit</td>
</tr>
<tr>
<td>labresults</td>
<td>any lab test that was observed within 24hrs with flag (high, low, critical)</td>
</tr>
<tr>
<td>vitals</td>
<td>all vitals with flags (high, low) (&lt;24hrs of Adm.)</td>
</tr>
<tr>
<td>procedures</td>
<td>all procedure codes (&lt;24hrs of Adm.)</td>
</tr>
<tr>
<td>medications</td>
<td>all POE for previous visit and within 24hrs Adm.</td>
</tr>
<tr>
<td>admission_type</td>
<td>admission type</td>
</tr>
<tr>
<td>admission_source</td>
<td>admission source</td>
</tr>
<tr>
<td>hospital_service</td>
<td>hospital service</td>
</tr>
<tr>
<td>age</td>
<td>discretized into bins {15,25,45,60,70,80,100}</td>
</tr>
<tr>
<td>city</td>
<td>city the patient is from</td>
</tr>
<tr>
<td>colonization_pressure</td>
<td>unit and hospital wide colonization pressure on day of admission</td>
</tr>
</tbody>
</table>

Table 3.2: We describe each patient admission using two sets of variables. We refer to the first set of variables as *Curated*. The second set of variables consists of all additional data procured from the structured fields of patients electronic medical records.

automated manner from the EMR. While these data are available in most hospital database systems they are often overlooked when building prediction rules since the goal is typically simplicity (e.g., back of the envelope addition with a small number of factors) over accuracy. We will consider three models: one based on the small set of curated risk factors and two others that include a longer list of data extracted automatically from the EMR.

Most of the variables or features we consider are categorical (discrete) and several are continuous. We map all categorical variables to binary variables. For example, the binary variable admission_source:RA is 1 if the patient is admitted from home
and 0 otherwise. In a clinical setting, interpretability by clinicians may be important when learning predictive models from data. Therefore, we consider prediction rules based on a linear combination of features that promises to be more understandable than non-linear combination. Of course, in medicine many relationships between co-variates and outcomes are non-linear. In part, we handle this non-linearity by discretizing continuous features and mapping data to binary features. E.g., Age is binned as in Table 3.1 and each bin maps to a different binary feature. More details regarding data extraction and preprocessing are given in Chapter 2.

After the initial preprocessing, we have 14 curated features and 10,845 additional features extracted from the EMR.

### 3.3.2 Learning to Predict Risk

Given the data, the goal is to learn a model that can be used to predict a newly admitted patient’s probability of testing positive for *C. difficile* during the current admission. A patient’s probability is calculated based on a linear combination of features extracted from the EMR. To learn this linear combination, we use logistic regression, a computationally efficient linear classifier.

Our dataset is defined as follows:

\[ D = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \{-1, 1\}\}_{i=1}^{n} \]

where \( D \) represents the learning task and consists of a feature vector \( x_i \) and a binary label \( y_i \) indicating whether or not the patient tested positive for *C. difficile* during the current admission, and \( n \) represents the number of unique patient admissions in the dataset (note: we represent vectors in boldface).

With logistic regression, we seek a function \( f : \mathbb{R}^d \rightarrow [0, 1] \) of the form:

\[ f(x_i) = \frac{1}{1 + e^{-(b_0 + w^T x_i)}} \]

where \( w \in \mathbb{R}^d \) (and \( x_i \in \mathbb{R}^d \ \forall i \)). Solving for the regression coefficients \( w \) and \( b_0 \) is a maximum likelihood estimation problem. To improve generalizability, we employ L2-regularized logistic regression [54], where \( \lambda \) is a tuning parameter that controls
the tradeoff between the number of errors on the training set and the complexity of
the model.

$$\min_w \frac{\lambda}{2} \|w\|^2 + \sum_{i=1}^{n} \log(1 + e^{(-y_i w^T x_i)})$$  \hspace{1cm} (3.1)

We add an extra constant dimension to \( x \) and compute the offset \( b_0 \) implicitly.
The solution to optimization problem in 3.1 depends on the \( X \) and \( y \) employed in the
training. In the results that follow, we compare the performance of several models
that differ only in the set of observations considered as features.

### 3.4 Experiments and Results

We extract all of the features referenced in Table 3.2 for each patient admission, in
addition to a binary label indicating whether or not a patient tests positive for \( C.\)
difficile (and when so that we can be sure we are in fact predicting the infection).
Next, we split the patient admissions into a training set and a test set. The training
set is used to construct the model (e.g., solve for the set of feature weights), while the
test set is used to validate the model on patients that were not used in building the
classifier. The admissions are divided according to admission date. We construct the
risk-stratification models on admissions from the first year (Apr. 12, 2011 to Apr. 11,
2012) and validate the learned prediction rules on all admissions from the second year
(Apr. 12, 2012 to Apr. 11, 2013). Dividing the data in this way mimics how we might
expect a classifier learned on historical data to perform on future data. It provides
a more conservative estimate of the accuracy of the classifier than an experiment
in which the training and holdout sets are chosen at random from the entire set of
records.

- **Training Set:** 34,846 admissions (372 admissions test positive after 24 hrs)
- **Validation Set:** 34,722 admissions (355 admissions test positive after 24 hrs)

The training data are used in Equation 1 to find the optimal setting of \( w \). The hy-
perparameter \( \lambda \) in Equation 1 was found using 5-fold cross-validation on the training
set, sweeping the value from \( 2^1 \) to \( 2^8 \) in powers of two.
We learn three different models based on different sets of variables. The first model, EMR considers all features extracted from the EMR (i.e., all of the features in Table 3.2). The second model, EMR (filtered), uses a reduced feature space in which only those features that occur in at least 1% of the population are considered. This filtering step reduces the dimensionality of the problem from 10,859 to 1,017 variables. The third model, Curated, considers only the first 14 features listed in Table 3.2, risk factors for \textit{C. difficile} readily available to clinicians.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dimensionality</th>
<th>AUROC ALL (95%CI)</th>
<th>AUROC(95%CI) RP&gt;48</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMR</td>
<td>10,859</td>
<td>0.8140 (0.80-0.83)</td>
<td>0.7896 (0.76-0.81)</td>
</tr>
<tr>
<td>EMR (filtered)</td>
<td>1,017</td>
<td>0.8129 (0.79-0.83)</td>
<td>0.7886 (0.76-0.82)</td>
</tr>
<tr>
<td>Curated</td>
<td>14</td>
<td>0.7163 (0.69-0.75)</td>
<td>0.6900 (0.66-0.72)</td>
</tr>
</tbody>
</table>

Table 3.3: Performance of three models varying in complexity on the test data (34,722 patients). We measure performance in terms of area under the receiver operating characteristic curve (AUROC) of predictions applied to all of the patients present in the hospital 24 hours after admission (who have not yet tested positive for \textit{C. difficile}) and also a subset of patients with a risk period (RP) greater than 48 hours.

Applied to the validation data, we achieve the results in Table 3.3. This table presents the area under the receiver operating characteristic curve (AUROC) [55]. The 95% confidence intervals (CI) were calculated using 100 bootstrap samples of the test set [56]. In the third column of Table 3.3, we consider the performance on all patients in the study population. Note this includes patients who test positive or are discharged between 24 and 48 hours after admission. To measure the ability of the model to risk stratify patients who test positive (or who are discharged) later, in the last column of Table 3.3 we measure the performance of each model on the subset of admissions with a risk period >48 hours. This corresponds to a total 28,984 admissions (286 positive). In general, we expect the prediction task to become more challenging as we consider a population of patients with longer risk periods. Patients with longer risk periods have longer stays and are typically sicker than patients with shorter stays. Thus, distinguishing the high-risk patients from the low-risk patients becomes more difficult. Moreover, while the risk period increases the data we use to make the prediction remain the same (data from the first 24 hours). I.e., on average we are predicting further into the future.
Figure 3-3: Receiver operating characteristic curve for each of the three models listed in Table 3.3. Note that both models including additional variables gleaned from the EMR outperform the model based on only the Curated features. Despite the large difference in the dimensionality between the EMR and EMR(filtered) models, the ROC curves overlap significantly.

Related to the summary statistics given in Table 3.3, we plot the receiver operating characteristic curves generated using the test data in Figures 3-3. The dotted lines represent the 95% confidence bounds generated using 100 bootstrap samples from the
test data. Figure 3-3(a) shows the ROC curve generated on all admissions in the test data, while Figures 3-3 (b) and (c) focus on those patients with a risk period of at least 48 hours. In all three plots, we see a clear advantage of the models trained on all of the EMR data versus the model trained on the smaller subset of well-known risk factors, specifically in the region between 0.05 and 0.25 false positive rate (shown in Figure 3-3 (c)). There is significant overlap in the ROC curves for the EMR and the EMR (filtered) models. This suggests that a large fraction of the features are non-informative. Despite this, we were still able to learn a useful model in the higher dimensional space.

To further quantify each model's ability to risk stratify the patients in the validation set, we measure the calibration of each model, capturing how closely the predicted risk matches the actual risk [57]. In Figure 3-4, we show the extent to which each classifier is well calibrated. For each model, we sort the test examples by their predictions and group the examples into 10 groups of equal size. For each group, we calculate the median predicted probability of risk and the actual probability of risk (based on the number of examples in the group who test positive for \(C.\) difficile). We repeat this process for 100 bootstrap samples on the test data. The black dashed lines in the plots of Figure 3-4 represent a perfect calibration (a 45 degree line). While there is some variation in the data, the first two classifiers appear well calibrated. The smaller, Curated classifier, using only the 14 well-known variables, depicted in the last plot, underestimates the probability of testing positive for patients who are at high risk and overestimates the probability for patients who are at low risk. This suggests the model is underfit to the data.

Next, we compare the EMR (filtered) model and Curated model using a performance measure called Net Reclassification Improvement (NRI) [58]. The NRI is the reclassification improvement among those patients who test positive and those who test negative. We consider grouped NRI instead of continuous NRI since the Curated model is not well calibrated, and calibration can greatly affect the continuous NRI measurement [59]. We split test admissions into two groups based on the classifier output: low-risk and high-risk. We plot the NRI for different cutoffs (based
Figure 3-4: For characterizing calibration, we measure how well the predicted probability of testing positive for *C. difficile* generated by each of the three predictive models agrees with the actual probability of testing positive for *C. difficile*, among patients in the validation set.

In general we see an approximate 15% improvement in the reclassification of positive examples (i.e., the EMR (filtered) classifier does a better job at classifying the patients who eventually test positive for *C. difficile*). As the cutoff approaches 100, the performance of the two classifiers converges (i.e., NRI=0). This is expected since those patients with the highest predicted risk are also
Figure 3-5: The net reclassification improvement (NRI) of using the EMR(filtered) model to classify patients as high/low risk versus the Curated model.

Figure 3-6: The confusion matrices result from a decision threshold based on the 95th percentile.

the easiest to classify. Additionally, we give in Figure 3-6 the confusion matrix for each classifier using a decision threshold based on the 95th percentile. This comparison is based on a single threshold, so to further show the benefit of the EMR(filtered) approach over the Curated approach we plot the positive predictive value as we seep the decision threshold from the 50th percentile to the 100th percentile in Figure 3-7.
3.5 Summary & Conclusion

We examined two different approaches to learning models for the risk stratification of patients for *C. difficile* 24 hours after admission. First, we explored the predictive performance of a traditional approach, in which we identified a set of risk factors from the literature, and then learned a model (based on, in this case, 14 variables). Second, we took a data-driven approach to generating features in which we learned two different models (varying in complexity) using data automatically extracted from the EMR.

Despite the high-dimensionality of the problem, we showed that with regularization techniques it is possible to learn meaningful risk-stratification models for *C. difficile*. Furthermore, we showed that such a methodology can lead to better performance than the traditional approach. Applied to a set of held-out test patient admissions, the EMR-based models successfully identified a subset of patients who were at approximately 6 times the relative risk of the reference population.

However, this improvement in performance comes at a cost. The high dimensionality of the models makes them inherently less interpretable. In our work we chose a model that is linear in the covariates, in an effort to maintain some interpretability. The weights assigned to each variable shed light on the importance of the variables in task of predicting risk of *C. difficile*. For example, we noted in our analysis several important variables including: the unit location of the patient, the receipt of laxatives, receipt of IV vancomycin, and several blood serum levels (e.g., lipase and albumin). However, because of the correlation among covariates in our model, it is difficult to discern clinical significance from the regression weights alone. Such relationships would need to be studied in a follow-up study in order to confirm the clinical importance of these factors.
Figure 3-7: Applied to the validation set for varying decision thresholds, the EMR(filtered) model identifies a population at higher risk for *C. difficile* compared to the Curated model. Risk-stratification models can help identify a subset of patients at increased risk of acquiring *C. difficile* compared to the reference population (1%).
Chapter 4

Patient Risk Stratification as a Time-Series Classification Task

4.1 Introduction

A patient’s risk for infection with Clostridium difficile (C. difficile) is affected by temporal processes including the nature and timing of diagnostic and therapeutic activities, and the evolution of the patient’s pathophysiology over time. As both a patient’s state and hospital conditions change so will a patient’s risk of contracting specific diseases. In this chapter, we focus on estimating patient risk as it evolves over the course of a hospital admission.

In contrast to the work presented in the previous chapter where we made a single prediction for each patient admission, here we make a prediction on each day of a patient’s hospital admission. Concatenated over time, daily estimates of patient risk produce a one-dimensional time series, a risk profile. We hypothesized that these risk profiles could be used to further distinguish high-risk patients from low-risk patients, i.e., that there are patterns of risk that are associated with infections. To test this hypothesis we framed the problem of identifying hospitalized patients at high risk for infection with C. difficile as a time-series classification task. In this chapter, we propose and motivate the study of patient risk profiles to model the evolution of risk over the course of a hospital admission.
To the best of our knowledge, representing and studying the risk of acquiring an infection as a time series has not previously been done. However, in other domains researchers working in machine learning have proposed many different ways of taking into account the temporal dynamics of a problem.

Many argue that Hidden Markov Models (HMMs) are the most natural tool for dealing with sequential data. However, because the method is unsupervised one has limited control over what the hidden states correspond to [60]. When working with high-dimensional imbalanced data it can be especially difficult to learn relevant states in an unsupervised manner. Thus, despite the fact that HMMs seem to naturally fit the sequential data problem, many prefer support vector machines (SVMs) because of the ability to learn generalizable classifiers even when working with complex high-dimensional data.

When dealing with temporal data for classification, researchers commonly apply SVMs in one of two ways. The first technique maintains order by concatenating features from the last $n$ observations [61]. This approach works well when the event of interest always occurs within the same time-frame (i.e., within $n$ observations). This is not the case in our task; a patient may test positive any number of days into the admission and incubation periods can vary. The second technique involves time-dependent features. For example in financial time series forecasting, where the goal is to predict a stock price for the next day, features are commonly based on metrics calculated using past data [60]. We consider a variation on this approach, including both time-varying and time-invariant variables in our feature vectors.

We propose a two-stage risk-stratification method that extracts patient risk profiles and aims to identify patterns of risk that are more likely to lead to adverse outcomes. Our approach uses SVMs to first reduce the high-dimensional feature space. Concatenated together, the outputs of the SVM produce the risk profile for a patient. Once patient risk profiles are extracted, the problem of risk stratification becomes that of time-series classification. In this chapter, we explore a variety of different methods including classification using similarity metrics, feature extraction, and hidden Markov models.
Here, we consider patients with at least a 7-day hospital admission, who do not test positive for *C. difficile* before day 7. We chose to focus on this population since we are interested in the evolution of risk over time. This group of patients is already at an elevated risk for acquiring *C. difficile* because of the duration of the hospital stay. Focusing on this group makes the problem more difficult than predicting risk for the general population.

A direct comparison with the reported results in the literature for *C. difficile* risk prediction is difficult because of the differences in the precise problem definition and the population considered. Thus, to measure the added value of considering the temporal dimension, we implemented the standard approach as represented in the related literature [20]. This approach, referred to as the snapshot approach, risk stratifies patients based on their current state, ignoring the evolution of patient risk. Snapshot approaches presented in the literature consider the patient at either the time of admission (similar to the previous chapter) or some constant number of days before the index event [19]. In this chapter we show how our method, which incorporates the evolution of patient risk over the course of the visit, leads to a significant improvement over the snapshot approach.

### 4.2 The Data

For this analysis, we considered all inpatient admissions to Hospital C during the year of 2010. We included all stays at least 7 days in length. To ensure that we are in fact predicting the acquisition of *C. difficile* during the current admission, we excluded patients who tested positive for *C. difficile* in the 60 days preceding admission or, if negative, following the current admission [39]. In addition, we removed patients who tested positive before day 7 of the admission. The final population consisted of 9,751 hospital admissions (177 admissions with a positive test result after day 7).
4.3 Methods

Time-series data exist in many different domains, including medicine, finance, information retrieval and weather prediction. Much research has been devoted to the analysis and classification of such signals [62, 63]. In recent years, researchers have had great success in identifying temporal patterns in such time series and developing methods that forecast the value of variables. In most applications there is an explicit time series, e.g., ECG signals, stock prices, audio recordings, or daily average temperatures.

In contrast to the examples listed above, patient risk is not a directly measurable time series. Our proxy for patient risk is a laboratory test with a binary outcome, a test that is only administered to symptomatic patients and typically only once. This is a rather poor approximation for patient risk, since it does not vary over the course of a hospitalization whereas we expect risk to fluctuate with time. Thus, we propose a two-stage approach to risk stratification. We first extract patient risk profiles (one dimensional time-series approximations of patient risk) and then we apply time-series classification techniques to those profiles. We describe both stages in detail in Section 4.3.2 and Section 4.3.3. We begin by describing the extracted features.

4.3.1 Feature Extraction

In the previous chapter we extracted data pertaining to the first 24 hours of each patient admission. In this chapter, we extract data for each day of a patient’s admission. Consequently, whereas in the previous chapter we had a single feature vector per patient admission, here we have multiple feature vectors for each patient admission.

We extracted more than 10,000 features for each day of each hospital admission. Approximately half of the features, 5,055 features, are based on data collected at the time of admission. These features remain constant throughout the stay. The remaining 5,545 features are collected over the course of the admission and may change on a daily basis. The types of features are listed in Table 4.1.

We note that at the time of data extraction, the hospital was using a version of
Amalga that differed from the current version (e.g., a different coding system—the American Hospital Formulary System—was employed for medications). Therefore the features extracted in this chapter differ slightly from the features extracted in the previous chapter (which used a later version of Amalga). However, the data extraction and preprocessing has remained fairly consistent across versions. Categorical variables were processed as described in Chapter 2, except for ICD-9 codes which we considered as coded (it was later that we decided to map all diagnosis codes to the highest level of the hierarchy). In addition to the variables given in Chapter 2, here we also had access to the admission complaint and admission procedure, which were recorded as unstructured text entries. We mapped these entries to categories corresponding to ICD-9 codes using look-up tables based on commonly used terms. In this work,

<table>
<thead>
<tr>
<th>Time-Invariant Features</th>
<th>Time-Varying Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• prev. adm. ICD-9 codes (1727)</td>
<td>• lab results (2012)</td>
</tr>
<tr>
<td>• prev. adm. primary ICD-9 code (786)</td>
<td>• procedures (1293)</td>
</tr>
<tr>
<td>• home medications (761)</td>
<td>• location room (1209)</td>
</tr>
<tr>
<td>• prev. adm. medications (655)</td>
<td>• medications (872)</td>
</tr>
<tr>
<td>• patient's city (535)</td>
<td>• vitals (95)</td>
</tr>
<tr>
<td>• attending md (443)</td>
<td>• location unit (61)</td>
</tr>
<tr>
<td>• hospital service(39)</td>
<td>• hospital wide colonization pressure (1)</td>
</tr>
<tr>
<td>• admission source (22)</td>
<td>• unit colonization pressure (1)</td>
</tr>
<tr>
<td>• financial class code (19)</td>
<td>• day of admission(1)</td>
</tr>
<tr>
<td>• admission complaint (18)</td>
<td></td>
</tr>
<tr>
<td>• admission procedure (16)</td>
<td></td>
</tr>
<tr>
<td>• patient's race (10)</td>
<td></td>
</tr>
<tr>
<td>• patient age (9)</td>
<td></td>
</tr>
<tr>
<td>• marital status (5)</td>
<td></td>
</tr>
<tr>
<td>• patient sex (1)</td>
<td></td>
</tr>
<tr>
<td>• expected surgery (1)</td>
<td></td>
</tr>
<tr>
<td>• ER admission (1)</td>
<td></td>
</tr>
<tr>
<td>• dialysis (1)</td>
<td></td>
</tr>
<tr>
<td>• diabetic (1)</td>
<td></td>
</tr>
<tr>
<td>• history of C. difficile (1)</td>
<td></td>
</tr>
<tr>
<td>• num. hospital visits (90 days) (1)</td>
<td></td>
</tr>
<tr>
<td>• avg., max., total LOS (90 days) (3)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Variables considered in our analysis can be grouped into two broad categories: time-invariant variables and time-varying variables. Here, we list the variable types that fall into each category with the number of resulting features in noted in parentheses.
colonization pressure and statistics from the hospital admissions in the last 90 days were left as continuous variables. The resulting feature vectors were sparse; on average less than 1% of each feature vector was non-zero.

4.3.2 Extracting Patient Risk Profiles

As described in the previous section, each day of a hospital admission is associated with a feature vector representing both time-varying and time-invariant data for a visit. We refer to a feature vector for a single day as a *snapshot*. In this section, we describe how multiple snapshots are used to estimate the risk profile for a patient. We use SVMs here instead of L2-regularized logistic regression (which we used in the previous chapter). These learning methods are similar, but differ in the definition of the loss function. SVMs use the hinge loss while L2-regularized logistic regression relies on the log loss. In our work, we actually started out using SVMs and later switched to using L2-regularized logistic regression. We do not expect this switch to lead to significant differences in classification performance. The main reason we switched to logistic regression was for the ease of generating a probabilistic interpretation of risk. Using SVMs we learn a model mapping each snapshot (i.e., feature vector) to a scalar value related to the patient's daily risk of eventually testing positive for *C. difficile*.

We thought carefully when deciding how to label the data for learning this mapping. We do not have ground-truth labels for each day of a patient's admission. We only know whether or not a patient eventually tests positive for *C. difficile*, not what their risk is on each day. We briefly explored a labeling in which we assumed that a patient was only at high risk *n* days before a positive index event, but choosing *n* proved difficult since this approach assumes that *n* is constant across patients. Events influencing a patient's risk can happen at any time. Looking only *n* days back may not capture this event. Since it is possible that any part of a patient's admission may contribute to their risk of later acquiring *C. difficile*, we assign each day of an admission in which the patient eventually tests positive as positive, even though the patient may not have actually been at high risk on each of those days. In doing so, we hope to identify high-risk patients as early as possible. Since a patient's risk does
not remain constant during an entire admission, there will be noise in the training labels. For example, there may be some days that look almost identical in the feature space but have different labels. To handle this noise we use a soft-margin SVM, one that allows for misclassifications in the training data. As long as our assumption does not lead to more incorrect labels than correct labels, it should be possible to learn a meaningful classifier.

Using these labeled feature vectors, we learn a linear SVM $f(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$, where $d$ is the dimensionality of the feature space. Applied to a test patient, $\mathbf{x}$ (where $\mathbf{x} \in \mathbb{R}^d$) the SVM produces a scalar that can be used to classify the patient as high risk or low risk. The distribution of these risk scores on the training data is shown in Figure 4-1. As one can see, these scores do a reasonable job of separating positive and negative cases.

This method ignores the temporal evolution of risk. The resulting classification is based only on a snapshot of the patient. Thus, we do not use the SVM as a classifier but instead concatenate the daily continuous predictions made by the SVM, $\mathbf{w}^T \mathbf{x} - b$ (i.e., the distance to the decision boundary) to produce a risk profile. In Figure 4-2, we show an example risk profile for a positive and for a negative patient. In the next section we explore the possibility of using these risk profiles to further distinguish high-risk from low-risk patients.
4.3.3 Classifying Patient Risk Profiles

Given the risk profiles for each patient, the risk-stratification task becomes a time-series classification task. Time-series classification is a well-investigated area of research, with many proposed methods. For an in-depth review of sequence classification we refer the reader to [63]. Here, we explore three different approaches to the problem: classification based on feature vectors, similarity measures, and finally HMMs. We first describe each method, and later present results about their performance in Section 4.4.

Classification using Feature Extraction

There are many different ways to extract features from time series. In the literature many have proposed time-frequency representations extracted using various Fourier or wavelet transforms [64]. Given the small number of samples in our time-series data, we were wary of applying such techniques. Instead we chose an approach inspired by the combination of classifiers in the text domain using reliability indicators in [65]. We define a feature vector based on different combinations of the predictions made in the first stage. We list the features in Table 4.2. These features aim to summarize different temporal patterns of risk (but are not comprehensive e.g., we do not consider the skewness).

Features 2-4 are averages; Features 3 and 4 weight days closer to the time of classification more heavily. Features 6-10 are different measures for the amount of fluctuation in the time series. Features 5 and 11 capture information about the most recent states of the patient. Features 12 and 13 identify runs in the data, i.e., periods.
Table 4.2: Univariate summary statistics for observation vector $x = [x_1, x_2, ..., x_n]$ of time where the patient is consistently at high or low risk. Finally, Features 14-17 summarize information regarding global maxima and minima in the time series.

Given these feature definitions, we map each patient admission risk profile to a fixed-length feature vector. These summarization variables allow us to compare time series of different lengths, while still capturing temporal information (e.g., when the maximum risk occurs relative to the time of prediction). Given this feature space, one can learn a classifier to identify high-risk patients. This approach is described in Figure 4-3.

### Classification using Similarity Metrics

In the previous section, we learned a second classifier based on features extracted from the time series. In this section, we consider classifiers based directly on the time series.

SVM classification is based on a notion of similarity between pairs of examples.
Given \( m \times n \) admission records, where \( m \) is the number of observations for each day and \( n \) is the number of days, estimate daily risk \( x_i \) based on the observed admission records for each day \( p_{i} \) for \( i=1...n \).

Concatenate predictions and extract feature vectors \( x' \) based on time series \( x \).

Classify each admission based on the \( x' \) to predict whether or not \( p \) will test positive for C. difficile.

Figure 4-3: A two-step approach to risk stratification where features summarizing the temporal patterns of the data are extracted from the time-series data.

(e.g., \( x_i \) and \( x_j \)). This similarity measure is defined by a kernel function. One of the most common non-linear kernels is the Gaussian radial basis function kernel:

\[
k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2).
\]

Its output is dependent on the Euclidean distance between examples \( x_i \) and \( x_j \). This distance measure requires vectors of the same length. We consider two approaches to generating vectors of the same length: (1) linear interpolation and (2) truncation. In the first approach we linearly interpolate between points. In the second approach we consider only the most recent 5 days of data, \( x_{n-4}, x_{n-3}, ..., x_n \).

Euclidean distance is a one-to-one comparison. In contrast, the dynamic time warping (DTW) distance is a one-to-many comparison [66]. DTW computes the distance between two time series by finding the minimal cost alignment. Here, the cost is the absolute distance between aligned points. We linearly interpolate all time series to have the same length, the length of the longest admission within the dataset (54). To ensure that the warping path does not contain lengthy vertical and horizontal segments, we constrain the warping window (how far the warping path can stray from the diagonal) using the Sakoe-Chiba band with a width equal to 10% of the length of the time series [67]. We learn an SVM classifier based on this distance metric, by replacing the Euclidean distance in the RBF kernel with the DTW distance,

\[
k(x_i, x_j) = \exp(-\gamma DTW(x_i, x_j))
\]

as in [68]. It is important to note that this kernel is not guaranteed to be positive semidefinite [69]. This has important theoretical implications regarding the convexity of the SVM objective function, and the solution is no longer guaranteed to converge to a global optimum. Despite this, researchers
have been experimenting with indefinite kernels for some time [70]. Various strategies for transforming an indefinite kernel matrix into one that is positive semi-definite, e.g., by shifting all the negative eigenvalues by a constant, have been proposed [71]. In our work, we did not find it necessary to employ any of these strategies.

**Classification using Hidden Markov Models**

In our application, we can make observations about a patient on a daily basis, but we cannot directly measure whether or not a patient is at high risk (our labels are only approximate). By applying HMMs we assume there is a sequence of hidden states, $x_1, x_2, ..., x_n$ that govern the observations $y_1, y_2, ..., y_3$. Here, the observations are the predictions made by the SVM. We consider a two-state HMM where each state, $s_1$ and $s_2$, is associated with a mixture of Gaussian distributions over possible observations. At an intuitive level, one can think of these states as representing low and high risk. Using the data, we learn and apply HMMs in two different ways.

In the first, we explore a classic approach to HMMs for classification, where one separates the training examples by class, ultimately learning two separate models. In the second we learn a single left-to-right HMM on the training data. The HMM applied in this way detects if and when a state transition occurs.

**Classification via Likelihood**

We hypothesize that there exist patterns of risk that are more likely to lead to a positive test result. To test this hypothesis, we first consider the classic approach to classification using HMMs (described in Section VI-B of [72]). We learn two separate HMMs: one using only observation sequences from positive patients and another using only observation sequences from negative patients. We initialize the emission probabilities for each model based on the data and initialize the transition probabilities as uniform probabilities. Given a test observation sequence, we apply both models and calculate the log-likelihood of the data given each model using the forward-backward algorithm. We stratify the patients, based on the ratio of the log-likelihoods.

**Classification via Posterior State Probabilities**

As we saw in Figure 4-2, the SVM output for a patient may fluctuate greatly
from day to day. While large fluctuations in risk are not impossible, they should not be common. Recall that in our initial calculation while the variables from time of admission are included in the prediction, the previous day’s estimated risk is not. The predictions produced by the SVM are independent. HMMs allow us to model the observations as a sequence and induce a temporal dependence in the model in which the current state $x_t$, depends on the previous state, $x_{t-1}$.

We learn an HMM on a training set. We consider a two-state model. We initialize the emission probabilities as $p(y_t|x_t = s_1) = N(\mu_1, 1)$, $p(y_t|x_t = s_2) = N(\mu_2, 1)$ $\forall t$ where $\mu_1 = -1$ and $\mu_2 = 1$. Based on this initialization, $s_1$ and $s_2$ correspond to “low-risk” and “high-risk” states as mentioned above. A key decision was to use a left-to-right model where once a patient reaches a “high-risk” state they remain there. I.e., the observations can trigger a single transition from low risk to high risk. This decision was motivated by the clinical use case in which identifying a patient as high risk triggers an intervention, such as moving the patient to a private room. All remaining transition probabilities were initialized uniformly. Applied to a test example we compute the posterior probabilities $p(x_t|y_1, ..., y_n)$ for $t = 1...n$ using the forward-backward algorithm. Because of the left-to-right assumption, if enough high-risk observations are made it will trigger a transition to the high-risk state. Figure 4-4 shows two examples of risk profiles and their associated posterior state probabilities $p(x_t = s_2|y_1, ..., y_n)$ for $t = 1...n$.

Note that all of the predictions in Figure 4-4, except the prediction on the last day, are non-causal since data from $t = 1...n$ are used to calculate the posterior. Therefore, we do not use these earlier state probabilities to classify patients; we classify each patient according to the probability of being in a high-risk state on the most recent day i.e., $p(x_n = s_2|y_1, ..., y_n)$. The reason we calculate the state for the earlier days is to illustrate how this method can help identify when a patient transitioned from low risk to high risk. E.g., in Figure 4-4 (b) the patient became high risk around day 11 but did not test positive until day 24.
Figure 4-4: Given all of the observations from $y_1, \ldots, y_n$ (in blue) we compute the posterior probability of being in a high-risk state for each day (in red).
4.4 Experiments & Results

This section describes a set of experiments used to compare the methods outlined in this chapter for predicting a patient's risk of acquiring *C. difficile* during the current hospital admission. We start by describing the experimental setup, which is maintained across all experiments, and then present the results.

4.4.1 Experimental Setup

In this chapter, we consider the task of predicting which patients are at high risk two days before an index event. For patients who eventually test positive, we defined the index event as the time of the positive test result. For patients who never test positive, researchers typically use the discharge day as the index event [39]. However, this can lead to deceptively good results because patients nearing discharge are typically healthier than patients not nearing discharge. To avoid this problem, we define the index event for negative examples as either the halfway point of their admission, or 5 days into the admission, whichever is greater. We consider a minimum of 5 days for a negative patient since 5 days is the minimum amount of data we have for any positive patient (e.g., a patient who tests positive on day 7).

In our problem we have severe class imbalance: 9,574 negative examples and 177 positive examples. In recent years, there has been much work investigating how to best deal with class imbalance [73–77]. In general, these techniques result in a shift of the decision threshold away from the minority class and toward the more prevalent class. The offset of the decision threshold is important when considering calibration results, however in this chapter we focus on the AUROC which depends solely on the ordering of patients from low risk to high risk. This ordering is not affected by a change in the offset of the decision threshold. Therefore, to reduce the severity of the class imbalance and speed up computation, we randomly subsampled the negative class, selecting 10 negative examples for each positive example. To offset the remaining class imbalance, we employed asymmetric cost parameters as in [78].

After subsampling the data we removed outliers, those patients with admissions
longer than 60 days. Next, we randomly split the data into stratified training and test sets with a 70/30 split. The training set consisted of 1,251 admissions (127 positive), while the test set was composed of 532 admissions (50 positive). This split was maintained across all experiments. In all of the experiments, the training data was used for parameter selection. For training and classification, we employed SVMlight [79] and Kevin Murphy’s HMM Toolbox [80].

### 4.4.2 Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>AUROC</th>
<th>95% CI</th>
<th>F-Score</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Snapshot</strong></td>
<td>0.69</td>
<td>0.62-0.76</td>
<td>0.28</td>
<td>0.19-0.38</td>
</tr>
<tr>
<td><strong>RP+Average</strong></td>
<td>0.75</td>
<td>0.69-0.82</td>
<td>0.32</td>
<td>0.21-0.41</td>
</tr>
<tr>
<td><strong>RP+Similarity_{Euc,5days}</strong></td>
<td>0.73</td>
<td>0.67-0.80</td>
<td>0.27</td>
<td>0.18-0.37</td>
</tr>
<tr>
<td><strong>RP+HMM_{likelihood}</strong></td>
<td>0.74</td>
<td>0.68-0.81</td>
<td>0.30</td>
<td>0.20-0.38</td>
</tr>
<tr>
<td><strong>RP+Similarity_{Euc,interp.}</strong></td>
<td>0.75</td>
<td>0.69-0.82</td>
<td>0.31</td>
<td>0.22-0.41</td>
</tr>
<tr>
<td><strong>RP+Similarity_{DTW}</strong></td>
<td>0.76</td>
<td>0.69-0.82</td>
<td>0.31</td>
<td>0.22-0.41</td>
</tr>
<tr>
<td><strong>RP+HMM_{posterior}</strong></td>
<td>0.76</td>
<td>0.70-0.82</td>
<td>0.30</td>
<td>0.21-0.41</td>
</tr>
<tr>
<td><strong>RP+Features</strong></td>
<td>0.79</td>
<td>0.73-0.85</td>
<td>0.37</td>
<td>0.24-0.49</td>
</tr>
</tbody>
</table>

Table 4.3: Predicting a positive test result two days in advance using different classifiers. **Snapshot** represents the traditional approach to risk stratification, and is the only classifier that is not based on patient risk profiles (RP).

![Graph](image)

Figure 4-5: Results of predicting a patient’s risk of testing positive for *C. diff* in the held-out test set using **RP+Features**.
Table 4.3 compares the performance of eight different classifiers applied to the held-out test data. The first classifier is our baseline approach. In this approach patients are classified after the first stage based on their most recent snapshot. The second classifier $RP + \text{Average}$ is an initial improvement on this approach, and classifies patients based on the average value of their risk profile. The remaining classifiers are all based on time-series classification methods. $RP + \text{Similarity}_{\text{Euc.5days}}$ classifies patients using a non-linear SVM based on the Euclidean distance between the most recent 5 days of the risk profiles. $RP + \text{Similarity}_{\text{Euc.interp.}}$ uses the entire risk profile by interpolating between points. These two methods in addition to $\text{DTW}$ were described in Section 4.3.3. The difference between $RP + \text{HMM}_{\text{likelihood}}$ and $RP + \text{HMM}_{\text{posterior}}$ was described in Section 4.3.3. $RP + \text{Features}$ classifies patients based on a linear combination of the average and other summary statistics (described in Section 4.3.3) of the risk profile. For all of the performance measures we compute 95% pointwise confidence intervals by bootstrapping the held-out test set 100 times.

There is large overlap in the confidence intervals for many of the results reported in Table 4.3, in part because of the paucity of positive examples. Still, based on the mean performance, all of the techniques that incorporate the risk profile outperform the $\text{Snapshot}$ approach in terms of both the area under the receiver operating characteristic curve (AUROC) and the F-Score. Interestingly, the one approach that did not outperform the $\text{Snapshot}$ approach to the same extent, $RP + \text{Similarity}_{\text{Euc.5days}}$, considers only the most recent 5 days of a patient visit. This result supports our hypothesis regarding the lasting effect of earlier days on patient outcome, and our decision not to limit ourselves to training data from only $n$ days before a positive event.

Figure 4-5 gives the receiver operating characteristic curve (ROC) curve for the best method, $RP + \text{Features}$. It is difficult to interpret the performance of a classifier based on these results alone. Figure 4-6 gives the confusion matrix for the mean performance of $RP + \text{Features}$. To further convey the ability of the classifier to risk stratify patients, we split the test patients into quintiles (as is often done in clinical studies) based on the continuous output of the classifier. Each quintile contains
approximately 106 patients. For each quintile we calculated the probability of a positive test result, based on those patients who eventually test positive for \textit{C. difficile}. Figure 4-7 shows that the probability increases with each quintile. The difference between the 1st and 5th quintiles is striking. Relative to the 1\textsuperscript{st} quintile, patients in the 5\textsuperscript{th} quintile are at more than a 25-fold greater risk.

\begin{center}
\textbf{Predicted Outcome}
\end{center}
\begin{center}
\begin{tabular}{c|c|c}
\textbf{Actual Outcome} & \textbf{p} & \textbf{n} \\
\hline
\textbf{p'} & 26 & 24 \\
\textbf{n'} & 72 & 410 \\
\end{tabular}
\end{center}

Figure 4-6: Confusion Matrix Using the best approach, \textit{RP+Features}, we achieve a sensitivity of 50\% and a specificity of 85\% on the held-out data.

\begin{center}
\begin{tikzpicture}
\begin{axis}[
    title={Fraction of Patients who test Positive},
    ylabel={Fraction of Patients who test Positive},
    xlabel={Quintile},
    ybar, ymajorgrids, xtick=data,
    xticklabels={1st, 2nd, 3rd, 4th, 5th},
    bar width=10pt,
    enlarge x limits=0.5,
    symbolic x coords={1st, 2nd, 3rd, 4th, 5th},
    nodes near coords,]
\addplot[ybar, fill=red] coordinates {
(1st, 0.005) (2nd, 0.05) (3rd, 0.1) (4th, 0.25) (5th, 0.3)
};
\end{axis}
\end{tikzpicture}
\end{center}

Figure 4-7: Test patients with \textit{RP+Features} predictions in the 5\textsuperscript{th} quintile are more than 25 times more likely to test positive for \textit{C. difficile} than those in the 1\textsuperscript{st} quintile.

To get a sense of the importance of each feature used by \textit{RP+Features}, we used repeated sub-sampling validation on the training set. We randomly subsampled 70\% of the training data 100 times and learned 100 different SVMs; this resulted in 100 different sets of feature weights. The results of this experiment are shown in Figure
Figure 4-8: Feature weights from SVM learned using different folds of the training set. The definition of features is given in Table 4.2

4-8. The most important features are: the length of the time series (Feature 1) which corresponds to the time in the hospital, the fraction of the time for which the patient is at positive risk (Feature 9), and the maximum risk attained (Feature 14). The only two features with significantly negative weights are Features 10 and Features 13: the overall fraction of time a patient has a negative risk, and the longest consecutive period of time that a patient has negative risk.

4.5 Conclusion

In this chapter, we consider the risk of acquiring an infection as a time-series problem. This results in a two-stage approach to patient risk stratification. We first extracted patient risk profiles and then used them as an input to a classifier. We explored three different approaches to classification: similarity metrics, feature vectors, and hidden Markov models. The majority of the methods performed as well as if not better than the previous approach of classifying patients based on patient snapshots. We are encouraged by these results, which suggest that posing the risk-stratification problem as a time-series classification task can provide more accurate models.
Chapter 5

Risk Stratification with Time-Varying Coefficients: A Multitask Approach

5.1 Introduction

In previous work, risk-stratification models for *C. difficile* infection considered no more than a dozen risk factors identified by experts, and many of the risk factors considered pertained to time-invariant features (e.g., patient history) and researchers ignored changes in patient risk over time. [19–22]. In contrast to previous work, in Chapter 3 we leveraged the entire structured contents of the electronic medical record (EMR) and in Chapter 4 we incorporated the evolving risk profile of a patient when calculating patient risk.

These two extensions led to significant improvements in patient risk stratification. In this chapter, we describe a third extension that further distinguishes our risk-stratification approach from previous work. In previous work (and in the previous two chapters of this dissertation), predictive models for *C. difficile* were static over the course of the hospital admission. That is, although the patient changed over time the model used to compute the patient’s risk did not. However, in addition to changes
in patient state and hospital conditions, the relative importance of risk factors may change during an admission. E.g., the effect of patient history may diminish as the patient spends more time in the hospital. In this chapter, we extend our risk model to allow for this type of temporal flexibility.

We develop a novel approach to learning the time-varying effects of risk factors. We propose a multitask SVM framework based on the domain adaptation techniques presented in [81]. Time-varying coefficient models have been studied in other contexts like survival analysis [82]. Over the years, standard approaches to survival analysis, like Cox proportional hazards, have been extended to include time-dependent coefficients [83]. Extensions typically involve the addition of interaction terms between features and time-varying functions [84–88]. In many cases, the user must specify these functions. Researchers have developed non-parametric extensions, but these methods can be computationally inefficient for large, complex datasets. Often, researchers end up partitioning time into intervals and analyze each time period with a simple model. Our proposed approach is similar in the sense that we break the problem up into multiple tasks. However, instead of learning the models independently, we learn the models jointly using a multitask learning framework.

Multitask learning is a popular branch of machine learning that leverages the intrinsic relatedness among different tasks [89]. It has been studied extensively in many different applications [90]. Zhou et al. employed a multitask learning framework in their work on Alzheimer's disease progression [91]. In predicting the cognitive state of patients at different time points in the future, the authors considered each time point of prediction as a single regression task. They learned the tasks jointly using a temporal group Lasso regularization. In our work, we also treat each time-point of prediction as a single task; however instead of predicting multiple points into the future we predict risk each day based on an updated set of variables collected that day (i.e., the variables are also time-varying).

In the next section we describe our proposed approach in detail. In Section 5.4 we present the validation results on a holdout set of patients and compare the performance of a model with time-varying coefficients to one with time-invariant coefficients.
5.2 The Data

For this analysis we considered all adult inpatients admitted to Hospital C on or after 2011-04-12 and discharged on or before 2013-04-12 (n=73,454). We exclude admissions in which the patient was discharged or tested positive for *C. difficile* before the end of the third day (n=24,389) and admissions for which the patient had a positive test result for *C. difficile* within 14 days of the admission (n=59). This resulted in 49,006 unique admissions.

Here, we remove many predictable low-risk patients with shorter stays, and focus on those patients who we believe acquire the infection during the current hospital visit (as opposed to those who are already infected at the time of admission). Choosing to define the cut-off as the end of the third day, rather than in terms of hours from the time of admission was an important design decision. If we had done the latter, we would be making predictions at many different times in a day, since patients are admitted throughout the day. Instead, we set up the problem in a way that allows for simultaneous predictions for every patient (at the end of each day). A uniform time of prediction makes sense from a clinical perspective, since it streamlines the risk-stratification process. E.g., each morning we could produce a list of high-risk patients and circulate it to the healthcare staff before making patient rounds.

5.3 Methods

5.3.1 Feature Extraction

As in the previous chapter, we represent each day of a patient’s admission with a single feature vector composed of both the time-invariant features collected at the time of admission and the time-varying features collected over the course of each day. These variables are listed in Table 5.1. This work was done using data from the current version of Amalga. We extracted and processed the data using the same techniques described in Chapter 2. In contrast to the work in Chapter 4, here we map all continuous variables to binary features. Doing so allows us to later capture
some of the nonlinear relationships that may be present in the data without using a nonlinear classifier. We discretize all continuous variables (except for age) using cutoffs based on quintiles from the training data. After preprocessing we have close to 10,000 binary variables for each patient day. To reduce the dimensionality, we filter features that do not occur in at least 1% of the training set. Ultimately, each day is represented as a vector of 905 binary variables.

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invariant</td>
<td>previous visits</td>
<td>statistics on previous hospitalizations (within 90 days)</td>
</tr>
<tr>
<td>Invariant</td>
<td>dxcodes</td>
<td>highest level of ICD9 codes coded during most recent visit</td>
</tr>
<tr>
<td>Invariant</td>
<td>hist.cdi</td>
<td>previous C. difficile Infection within the last year and earlier</td>
</tr>
<tr>
<td>Invariant</td>
<td>previous medications</td>
<td>all ordered medications from the most recent visit</td>
</tr>
<tr>
<td>Invariant</td>
<td>admission_type</td>
<td>admission type</td>
</tr>
<tr>
<td>Invariant</td>
<td>admission_source</td>
<td>admission source</td>
</tr>
<tr>
<td>Invariant</td>
<td>hospital_service</td>
<td>hospital service at the time of admission</td>
</tr>
<tr>
<td>Invariant</td>
<td>age</td>
<td>age at time of admission</td>
</tr>
<tr>
<td>Invariant</td>
<td>city</td>
<td>city in which the patient resides</td>
</tr>
<tr>
<td>Varying</td>
<td>lab results</td>
<td>all lab tests with flags (high, low, critical, normal)</td>
</tr>
<tr>
<td>Varying</td>
<td>vitals</td>
<td>all vitals with flags (high, low, normal)</td>
</tr>
<tr>
<td>Varying</td>
<td>procedures</td>
<td>all procedure codes</td>
</tr>
<tr>
<td>Varying</td>
<td>medications</td>
<td>all ordered medications</td>
</tr>
<tr>
<td>Varying</td>
<td>locations</td>
<td>all unit and room locations within the hospital</td>
</tr>
<tr>
<td>Varying</td>
<td>colonization_pressure</td>
<td>unit- and hospital-wide colonization pressure (see Chapter 2)</td>
</tr>
</tbody>
</table>

Table 5.1: We describe each patient admission using both time-invariant and time-varying variables.

### 5.3.2 Learning to Predict Daily Risk

In the previous chapter we came up with a daily risk score proportional to a patient’s probability of acquiring an infection with C. difficile during a hospitalization. This estimate was updated daily and incorporated information about previous days’ estimates (i.e., risk profile). However, the initial model we learned (to extract the risk profiles) was static. Here, we extend the work to incorporate time-varying coefficients (i.e., feature weights).

Our dataset is defined as follows:

\[
\mathcal{D} = \left\{ \{(x_i^j, y_i^j)\}_{j=1}^{m_i} \right\}_{i=1}^{n} \quad \text{where,}
\]
$D$ represents the dataset and consists of a feature vector $x_i^j \in \{0, 1\}^d$ where $d$ is the number of variables we consider (corresponding to patient $i$ on day $j$), and a binary label $y_i^j$ indicating whether or not the patient tested positive for $C.\difficultile$ during the current admission. As in the previous chapters, we label each day of an admission in which the patient eventually tests positive as $+1$ and $-1$ otherwise. $n$ represents the number of unique patient admissions in the dataset, and $m_i$ is the number of admission days we have for each patient admission.

Pooling all training examples $(x_i^j, y_i^j)$ together and learning a single model ignores the value of $j$. As a patient spends more time in the hospital, we expect the factors contributing to patient risk to change. Thus, we aim to learn a patient risk-stratification model that changes as the patient spends more time in the hospital. We begin by splitting $D$ into separate tasks according to $j$, the day of the visit.

$$D_t = \{(x_i^j, y_i^j)\}_{i=1}^n \text{ for } j = t,$$

We chose the number of tasks, and the timing intervals $t$ based on the number of learning examples available (see Figure 5-1). For our data, this resulted in 6 distinct tasks:

$$D_1, D_2, D_3, D_{[4,5]}, D_{[6,9]}, D_{[10,\infty)}$$

We could learn a separate model for each task independently of the others, but in doing so we would be limiting ourselves to one sixth of the available training data. Moreover, the tasks themselves are related in time and thus are not independent. While coefficients might vary from one day to another, we do not expect large fluctuations in time. Therefore, we use a multitask learning framework to leverage the inherent relatedness among the different tasks and take advantage of the entire corpus of training data.

We learn the different models jointly using a multitask $L2$-regularized logistic regression framework. We employ domain adaptation techniques from [81], remapping each feature vector $x_i^j$ to a feature vector 7 times the dimensionality of the original vector using the mapping function $\Phi(x_i^j)$:
Figure 5-1: We divide the problem of daily patient risk stratification into six different tasks based on the number of available training examples.

The new feature vectors consist of two copies of the original feature vector, padded with zeros \( \mathbf{0} = [0_1, 0_2, 0_3, ..., 0_d] \). We can estimate the risk of a new patient day \( \mathbf{x}_i^j \) using:

\[
f(x_i^j) = \frac{1}{1 + e^{-w^T\Phi(x_i^j)}}
\]  

(5.1)

We learn the regression coefficients \( \mathbf{w} \) using L2-regularized logistic regression, replacing \( \mathbf{x}_i^j \) with \( \Phi(x_i^j) \) in the objective function. The result is a \( \mathbf{w} \in \mathbb{R}^{7d} \) where \( d \) was the dimensionality of the original problem, before the remapping.

Consider \( \mathbf{w} = [\mathbf{w}_0, \mathbf{w}_{[1]}, \mathbf{w}_{[2]}, \mathbf{w}_{[3]}, \mathbf{w}_{[4,5]}, \mathbf{w}_{[6,9]}, \mathbf{w}_{[10,\infty]}] \), where each \( \mathbf{w}_t \in \mathbb{R}^d \), \( \mathbf{w}_0 \) corresponds to a vector of shared feature weights since it is based on data from all days, \( \mathbf{w}_{[1]} \) is based on only data from day 1 and so on. Using Eq. 5.1, we note that
the risk of patient \(i\) on day 1 is proportional to \(w_0^T x_i^1 + w_{[1]}^T x_i^1\), or \((w_0 + w_{[1]})^T x_i^1\). In general, we can estimate the risk of a new patient day \(x_i^j\) using Eq 1, but replacing \(w\) with \(w'_j\), without having to remap the feature vector. Writing the function this way shows how learning the models jointly results in six different models all with a shared component \(w_0\).

\[
\begin{align*}
   w'_j &= \begin{cases} 
   w_0 + w_{[1]} : j \in [1] \\
   w_0 + w_{[2]} : j \in [2] \\
   w_0 + w_{[3]} : j \in [3] \\
   w_0 + w_{[4,5]} : j \in [4, 5] \\
   w_0 + w_{[6,9]} : j \in [6, 9] \\
   w_0 + w_{[10,\infty)} : j \in [10, \infty) 
   \end{cases}
\end{align*}
\]

Applying the model described above to a patient’s data results in a single estimate of risk for each day \(\hat{y}_j\). We incorporate risk estimates from previous days using a cumulative moving average. That is the predicted risk on day \(j\) is calculated as \(\text{risk}_j = \frac{\hat{y}_1 + \ldots + \hat{y}_j}{j}\) as in the \(RP+Average\) approach presented in Chapter 4. In terms of classification performance, this was not the best approach presented in Chapter 4; however it significantly outperformed the \(Snapshot\) approach and was by far the simplest of the non-snapshot approaches.

5.4 Experiments and Results

Employing the methods and data described above we learned and validated a risk stratification model for identifying inpatients at high-risk of acquiring an infection with \(C.\ difficile\) throughout their hospital admissions. We split the data into a training set and a holdout set temporally, training on data from the first year, and validating our model on data from the second year.

Our training data consisted of 190,675 visit days pertaining to 24,607 unique visits. Within the training data, 258 admissions had a positive test for \(C.\ difficile\) resulting in 2,608 training days with a positive label. To mitigate the effect of patients already
showing symptoms, we removed patient days corresponding to the day of and the day before the positive test result. In addition, when training the classifier we randomly subsampled the data such that no patient contributed more than 3 days worth of the data to the training set. If we had not done this some patients would have been represented up to 10 times more often than other patients. Given the small number of positive examples, patients with longer visits could have significantly biased the classifier. In turn, this could have resulted in overfitting to those patients with longer visits. In the previous chapter, this was not an issue since all patients had visits of at least 7 days, and thus the bias was not as severe.

As stated in previous chapters, L2-regularized logistic regression has a hyperparameter ($\lambda$) that controls the tradeoff between the loss and the regularization terms in the objective function. We performed repeated 5-fold cross validation on the training data to find the optimal setting of this hyperparameter. We swept $\lambda$ from $2^1$ to $2^{10}$ in powers of 2 and chose the setting which maximized the AUROC in cross validation. The regression coefficients, i.e., $[w_1, w_2, w_3, \ldots, w_7d]$, were solved for using LIBLINEAR [92].

In our final model, the relative importance of risk factors (i.e., the regression coefficients) is allowed to vary over time. Figure 5-2 (a) shows the extent to which the weights vary across time. The columns correspond to the different time periods (i.e., tasks) and the rows correspond to the different features. The features are sorted in descending order according to their weight on the first day. The color of each cell is related to the weight of that feature for the specified task. We consider the normalized feature weight i.e., we divide each weight by the sum of the absolute value of all weights for a task. All cells corresponding to features that have high positive weight are red, and those with high negative weight are dark blue. If the relative importance of the weights did not change over time, each column would look identical to the first column. However as Figure 5-2 (a) shows, the relative importance of features does change.

Figure 5-2 (b) shows that even among the top ten features (on Day 1) there are changes in the relative importance of features over time. In Figure 5-2 (b) the first
Figure 5-2: The changing relative importance of features over time. For each time period (i.e., task), the features are ranked according to the feature weight for the Day 1 task. The color represents the normalized feature weight for each task.

two features become less important over time, while the third feature becomes more important. Table 5.2 lists the five features with the greatest positive weight on Day 1 through Day 10+. Note that initially the most important feature is the patient’s one year history of infection with \textit{C. difficile}. As a patient spends more time in the hospital, this feature loses importance relative to the location of the patient in the hospital.
<table>
<thead>
<tr>
<th>Ranking</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1</td>
<td>1 Yr. hist. of CDI</td>
<td>Hist. of CDI</td>
<td>Daily Units:2NE</td>
<td>Temp.:High</td>
<td>Prev. Meds:Sevelamer</td>
</tr>
<tr>
<td>Day 2</td>
<td>1 Yr. hist. of CDI</td>
<td>Hist. of CDI</td>
<td>Daily Units:2NE</td>
<td>Temp.:High</td>
<td>Service:MED</td>
</tr>
<tr>
<td>Day 3</td>
<td>1 Yr. hist. of CDI</td>
<td>Hist. of CDI</td>
<td>Daily Units:2NE</td>
<td>Temp.:High</td>
<td>Prev. Meds:Sevelamer</td>
</tr>
<tr>
<td>Day [4,5]</td>
<td>1 Yr. hist. of CDI</td>
<td>Hist. of CDI</td>
<td>Daily Units:2NE</td>
<td>Temp.:High</td>
<td>Mean Platelet Vol.:normal</td>
</tr>
<tr>
<td>Day [6,9]</td>
<td>Meds: Pantoprazole</td>
<td>1 Yr. hist. of CDI</td>
<td>Daily Units:2NE</td>
<td>Hist. of CDI</td>
<td>Mean Platelet Vol.:normal</td>
</tr>
<tr>
<td>Day [10,+]</td>
<td>Daily Units:2NE</td>
<td>1 Yr. Hist. of CDI</td>
<td>Temp.:High</td>
<td>Hist. of CDI</td>
<td>Service:MED</td>
</tr>
</tbody>
</table>

Table 5.2: Features with greatest weight on Day 1 through Day 10+. We use color to highlight certain trends.

In Table 5.3, we note the 25 features with the greatest weight according to the shared model i.e., \( w_0 \). Not too surprisingly, patient history of \( C. \) difficile infection appears at the top of this list. The medications that appear in Table 5.3 include drugs administered to patients receiving kidney dialysis, drugs for the treatment of high blood pressure and heart disease, and proton pump inhibitors. When interpreting these weights it is important to keep in mind that many of the features in our model are highly correlated. These features may be directly or indirectly linked with an increased risk of \( C. \) difficile infection. However, further analysis could generate hypotheses about causal relationships that could be tested in a clinical setting.

### 5.4.1 Evaluating the Model

In previous work, risk stratification models were evaluated at a single point in time (e.g., typically 2 days before an index event). In Chapter 4 we modified the definition of index event for negative patients, since the task of distinguishing between patients about to test positive for \( C. \) difficile and patients about to be discharged from the hospital provides deceptively good results. Here, we further modify the evaluation methodology to better fit the context in which the models will be used. Evaluating models 2 days before an index event can help you compare classifier performance, but does not yield an accurate representation of how the model will perform in a clinical setting, since clinicians have no way of knowing when a patient is 2 days from an index event. In practice, we expect to apply the risk stratification model to each day of a patient’s visit. This results in multiple predictions for each patient, one corresponding to each day of the admission.
Table 5.3: Features with greatest “shared” weight.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature Index</th>
<th>Feature Name</th>
<th>Feature Description</th>
<th>Shared Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>905</td>
<td>OneYear_History</td>
<td>positive test for toxigenic C. diff in past year</td>
<td>0.2472</td>
</tr>
<tr>
<td>2</td>
<td>904</td>
<td>AllHistory</td>
<td>positive test for toxigenic C. diff ever in the past</td>
<td>0.2314</td>
</tr>
<tr>
<td>3</td>
<td>904</td>
<td>daily_units:XX</td>
<td>medicine patient care unit (22 beds)</td>
<td>0.2084</td>
</tr>
<tr>
<td>4</td>
<td>427</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1885</td>
</tr>
<tr>
<td>5</td>
<td>950</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1708</td>
</tr>
<tr>
<td>6</td>
<td>44</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1666</td>
</tr>
<tr>
<td>7</td>
<td>685</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1508</td>
</tr>
<tr>
<td>8</td>
<td>94</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1418</td>
</tr>
<tr>
<td>9</td>
<td>234</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1466</td>
</tr>
<tr>
<td>10</td>
<td>475</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1418</td>
</tr>
<tr>
<td>11</td>
<td>209</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1418</td>
</tr>
<tr>
<td>12</td>
<td>433</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1418</td>
</tr>
<tr>
<td>13</td>
<td>418</td>
<td>daily_vitals:temporal_h</td>
<td>temperature oral high</td>
<td>0.1418</td>
</tr>
<tr>
<td>14</td>
<td>74</td>
<td>attending_doctor_number:xxxx</td>
<td>nephrologist</td>
<td>0.1094</td>
</tr>
<tr>
<td>15</td>
<td>615</td>
<td>prev meds:63708390</td>
<td>lisinopril 10 mg Tab</td>
<td>0.1088</td>
</tr>
<tr>
<td>16</td>
<td>398</td>
<td>prev meds:63715254</td>
<td>vitamin B comp w/C, FA Tab</td>
<td>0.1077</td>
</tr>
<tr>
<td>17</td>
<td>434</td>
<td>daily_vitals:temporal_h</td>
<td>temperature axillary</td>
<td>0.1077</td>
</tr>
<tr>
<td>18</td>
<td>292</td>
<td>daily_vitals:temporal_h</td>
<td>temperature axillary</td>
<td>0.1077</td>
</tr>
<tr>
<td>19</td>
<td>587</td>
<td>daily_vitals:temporal_h</td>
<td>temperature axillary</td>
<td>0.1077</td>
</tr>
<tr>
<td>20</td>
<td>591</td>
<td>daily_vitals:temporal_h</td>
<td>temperature axillary</td>
<td>0.1077</td>
</tr>
<tr>
<td>21</td>
<td>441</td>
<td>daily_vitals:temporal_h</td>
<td>temperature axillary</td>
<td>0.1077</td>
</tr>
<tr>
<td>22</td>
<td>506</td>
<td>prev meds:63008947</td>
<td>pharmacy comment</td>
<td>0.1036</td>
</tr>
<tr>
<td>23</td>
<td>719</td>
<td>prev meds:63713239</td>
<td>pharmacy comment</td>
<td>0.1036</td>
</tr>
<tr>
<td>24</td>
<td>267</td>
<td>prev meds:63713239</td>
<td>pharmacy comment</td>
<td>0.1036</td>
</tr>
<tr>
<td>25</td>
<td>688</td>
<td>prev meds:6370897</td>
<td>pharmacy comment</td>
<td>0.1036</td>
</tr>
</tbody>
</table>

One could imagine a validation scheme in which we evaluated the performance of a classifier on each day independently. However, this does not provide an informative estimate of the model's performance since it is not clear, at least from a clinical perspective, what it means to classify a patient correctly \( m \) days out of a visit of \( n \) days. Again, we believe that the evaluation of a model should be driven by the use case of the model. In practice, we expect the model to identify high-risk patients who should receive some form of intervention (e.g., relocated to a private room). We assume that such an intervention will last for a time period determined by the physician (e.g., 10 days or for the remainder of the visit). So long as a patient is at low risk, the decision to intervene (or not) is made daily.

Thus, while the model’s daily predictions are allowed to fluctuate (from low to high and high to low), when evaluating the model we choose a single decision threshold. We apply this decision threshold to each day of a patient’s visit (up to the day before a positive test result or the day of discharge). If the patient’s daily estimated risk ever exceeds the decision threshold they are classified as high risk, otherwise they...
are classified as low risk. Thus, we consider it a more meaningful evaluation, which accurately mimics how we expect the model to be used in practice. By taking the maximum prediction for each patient and sweeping the decision threshold, we can generate a receiver operative characteristic curve.

We applied the model to the held-out validation set. The training set consisted of data from Apr. 12, 2011 to Apr. 11, 2012. The validation set consisted of data from the year Apr. 12, 2012 to Apr. 11, 2013 and was composed of 24,399 patient visits of which 242 had a positive test result for *C. difficile*. When validating our model we did not subsample the test data as we did with training data. Each patient admission in the validation set had at least 3 daily predictions of risk (since we considered only patients who were still present in the hospital at the end of the 3rd day).

Figure 5-3 plots the receiver operating characteristic curve, and the precision recall curve for the held-out validation set. Our model results in an area under the receiver operating characteristic curve (AUROC) of 0.81 (95%CI 0.78-0.84) and an area under the precision recall curve (AUPR) of 0.04 (95%CI 0.03-0.05). The AUPR appears low however, this is significantly better than a baseline classifier, (recall that the incidence of infection is approximately 1% in the study population).

To evaluate the ability of our model to distinguish high-risk patients from low-risk patients in the validation set, we picked a decision threshold based on the 95th percentile. We chose this cut-off to limit the number false positives, since infections with *C. difficile* are relatively rare events. Given this decision threshold, we correctly identify 69 patients out of 242 as high risk, and achieve a sensitivity of 0.28, a positive predictive value of 0.06, an F-score 0.09, and an odds ratio of 7.97 (confusion matrix TP=69 TN=23,006 FN=173 FP=1151). Of course, lowering the decision threshold will increase the sensitivity, however it will also increase the number of false positives. Ultimately, the choice of decision threshold depends on the intervention one intends to apply to high-risk patients. In the next chapter, we further explore the clinical ramifications of our prediction model through a cost-benefit analysis.

Figure 5-4(a) illustrates how far in advance we can predict positive test results. We note that in approximately half of the cases correctly identified, we identify them
Figure 5-3: We plot two performance curves generated by applying the risk stratification method described in the previous section to our holdout set of patient visits. We achieve an AUROC of 0.81 (95% CI 0.78-0.84) and a AUPR of 0.04 (95% CI 0.03-0.05).
(a) Fraction of correctly identified cases we can predict at least $x$ days in advance.

(b) Distribution of time to positive test from admission (in days)

Figure 5-4: Classification performance resulting from a classifier based on the 95th percentile.
at least 7 days in advance. Figure 5-4(b) illustrates when patients are testing positive. While we identify more patients who test positive earlier, the fraction of patients we correctly identify increases as the length of stay increases.

Figure 5-5: We compare a single-task classifier, where the model is time-invariant, to a multitask classifier where the model varies over time. The difference between the two classifiers is apparent for those patients who test positive later during the admission. These are the patients we are more interested in.

Finally, in Figure 5-5 we illustrate the performance of our model (the Multi-Task approach) compared to that of a model learned by simply pooling all the data (the Single-Task approach), in terms of the AUROC. Applied to all patients in the test set (risk period >3 days) the classifiers perform almost identically. However, when the patients are divided into subsets based on their risk period the difference becomes apparent. For patients with a longer risk period, the Multi-Task approach results in an improvement in performance over the Single-Task approach. While the difference is small, it is consistent. This difference is relevant, since the potential to intervene in a timely manner is greater for patients who test positive later in a visit. Therefore, the ability to identify such cases accurately is of considerable importance.
5.5 Conclusion

In this chapter we extended our patient risk stratification model to allow for time-varying coefficients. We segmented the problem into several temporal intervals and employed a multitask learning framework.

We evaluated the model by considering the predictions made on each day of a patient’s admission up to the point of a positive test result or discharge. Prior to this work, such models have only been evaluated at a single point during a hospital visit (e.g., at the time of admission or \(n\) days before the index event). We argue that evaluating the model over the entire course of the admission is a more accurate approximation of how the model would perform in practice.

When compared to a model with time-invariant coefficients, our proposed model performed better on patients with longer risk periods. While consistent, the difference was not significant; this may be due to the fact that the number of cases with longer risk periods is small. We have approximately 5,000 visits in the validation set with a risk period greater than 10 days, and only a small fraction of those patients end up testing positive. Additionally, in our formulation, we consider the same decision threshold everyday. However, a variable decision threshold could possibly lead to better results. In future work, we plan to further investigate how to improve the multitask learning approach. Still, we prefer this model to the static model presented in Chapter 4, since it allows for changes in risk factors over time.

In this work, we investigated the ranking of features across tasks and observed changes over time in the relative effects of risk factors. Our proposed method could be used to further investigate the temporal effects of risk factors over the course of a hospital admission, perhaps shedding new light on the relationship between risk factors and time, and in turn improving our understanding of the disease.
Chapter 6

Transferring Knowledge Across Hospitals

6.1 Introduction

Up to this point, all of the work presented in this dissertation has focused on using data from a single hospital, Hospital C, to predict infections with Clostridium difficile (C. difficile) at that hospital. In this chapter, we focus on the challenge of predicting infections at two neighboring hospitals within the same hospital network. The problem is made challenging by the fact that these hospitals engage with patient populations roughly one third the size of the population served by the much larger Hospital C; the ability to learn accurate models for predicting patient outcomes at specific hospitals typically hinges on the amount of training data available.

According to the American Hospital Association more than half of all hospitals registered in the US have fewer than 100 beds [93]. An average length of stay of 4.8 days [94] results in fewer than 8,000 admissions per year (this estimate assumes 100% capacity and is therefore an upper bound). Given a goal of learning models to predict rare events, (e.g., an event occurring in less than 1% of the population) the smaller institutions can collect no more than 80 positive training examples per year. In medicine, where relationships between covariates and outcomes are usually complex, 80 positive training examples is usually too few to learn a predictive model
that generalizes to new cases. This lack of training data makes it difficult to build hospital-specific models.

Global models, developed for general use across multiple hospitals, have been developed for some areas of medicine; statistical models (as well as heuristic risk scores) have been developed and tested on data accessed from large national registries (e.g., AC NSQIP [95]). However, as discussed in [96], these models often perform poorly when applied to specific institutions, because they do not take into account institutional differences.

In this chapter, we investigate an approach to building predictive models that involves augmenting data from individual hospitals with data from other hospitals. We consider a set of three hospitals, all belonging to the same hospital network. The data collected at each hospital contains hospital-specific distinctions, e.g., the labels used to refer to units and rooms within the hospital. Moreover, the hospitals differ in the types of patients admitted. These differences contribute to differences in the sets of variables that characterize relationships among observations and outcomes.

Applying data from multiple hospitals to predictions at a single target hospital presents an opportunity for transfer learning i.e., the leveraging of evidential relationships in one or more related source tasks for making predictions in a target task. Here, the target task aims to predict which admissions to a specific hospital will result in a positive test result for *C. difficile*. Labeled datasets from the other two hospitals make up the source tasks.

We explore three different solutions for building predictive models for a specific institution, in the context of using the base statistical methodology of *L*2-regularized logistic regression. The methods vary in terms of what training data are used and the details of the evidential features considered. The results suggest practical approaches to moving beyond a reliance on only local data to build institution-specific models for small institutions or rare events.
6.1.1 Background on Transfer Learning

Transfer learning tackles the problem of leveraging data from a related source task to improve performance on a target task. There are different flavors of transfer learning depending on how the source and target tasks differ and the distribution of labeled training data across source and target tasks. See [97] for an in-depth review of transfer learning.

Most of the studies performed to date in transfer learning have addressed differences in underlying distributions across tasks but often assume that all of the data lie in the same observation or feature space [96], [98–102]. In our problem formulation the outcome of interest, a *C. difficile* infection, is the same across all tasks. However, the datasets lie in distinct but overlapping feature spaces.

In prior work, Evgeniou and Pontil generalized regularization-based methods from single-task to multitask learning [98]. Their proposed solution is a natural extension of existing kernel-based learning methods that builds on ideas from hierarchical Bayesian modeling [99,100]. They assume a solution for each predicted outcome (task) of the form \( w_t = w_0 + v_t \), where \( w_0 \) represents a common solution shared among all tasks and \( v_t \) represents the task-specific variation from the common solution. Their method learns both common and task-specific components simultaneously, but assumes that all data are sampled from the same feature space.

Similar transfer learning approaches have been applied successfully to medical data [96,103]. In [96], the authors explore transfer learning for adapting surgical models to individual hospitals. Like us, they hypothesize that models learned in a straightforward manner from pooled data fail to reflect individual variations across hospitals. Their approach is two-step: using cost-sensitive support vector machines, they first train a model on the source data, and then learn a model for the target data while regularizing the model parameters toward those of the source model. Their experiments show a significant improvement over other methods such as ones learned only from target data or from source data. However, their work assumes that there is no missing data and that all data lie in an identical feature space. The omission
of hospital-specific features is typical in multi-center studies. In reality, transferring
data across hospitals can be messy because many of the observational variables such
as staff, protocol, and locations are hospital-specific. In Section 6.3, we show how
important these hospital-specific variables can be.

Researchers have investigated the task of transferring knowledge across different
feature spaces in other contexts. Bel et al. explored the problems that arise when
classifying documents in different languages [104]. Common solutions to this problem
either translate one document to the target language or map both documents to a
language-independent feature space, analogous to either mapping the source data
into the target domain or mapping both to a shared representation. Previous work
has also proposed approaches to translate auxiliary data into the target space from
one medium (e.g., text) to another (e.g., an image) [105]. Such translated learning
applies to a different scenario than ours; in translated learning there is no explicit
correspondence between the source feature space and the target feature space.

Several researchers have investigated transfer learning in the context of linear
classifiers [101,102]. Previous work has explored modifications to the SVM objective
function to include consideration of the loss associated with both target and source
data, but exclude the source data from either the constraint or the set of support vec-
tors. In comparison, we consider methods based on L2-regularized logistic regression
that do not require explicit modification of the objective function.

6.2 Methods

6.2.1 Data and Preprocessing

Our data comes from three hospitals within the same hospital network: Hospital A,
Hospital B, and Hospital C. All three hospitals are described below.
Hospital A is the smallest of the three hospitals. Hospital A has approximately 180 beds and sees just over 10,000 admissions a year. Hospital B is an acute care teaching hospital. It has approximately 250 beds and 15,000 inpatient visits per year. Hospital C is a major teaching and research hospital with over 900 beds and more than 40,000 inpatient visits per year.

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Number of Admissions</th>
<th>C diff cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First Year (Apr 2011-Apr 2012)</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>11,380</td>
<td>82</td>
</tr>
<tr>
<td>B</td>
<td>14,675</td>
<td>161</td>
</tr>
<tr>
<td>C</td>
<td>39,467</td>
<td>426</td>
</tr>
<tr>
<td></td>
<td>Second Year (Apr 2012-Apr 2013)</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>10,579</td>
<td>94</td>
</tr>
<tr>
<td>B</td>
<td>14,640</td>
<td>157</td>
</tr>
<tr>
<td>C</td>
<td>42,112</td>
<td>428</td>
</tr>
</tbody>
</table>

Table 6.1: The amount of available data varies significantly across the three different institutions. The outcome we consider occurs in approximately 1% of the population, resulting in low numbers of positive examples at smaller institutions.

Hospital A and Hospital B are located in the same city only 10 miles apart, whereas Hospital C is located in a different city about 50 miles away. Despite the large differences in size and location, Hospital C overlaps with Hospital A and B in terms of many of the services provided. Table 6.2 describes the population of patients admitted to each hospital over the same two-year time period.

As in previous chapters, we are interested in risk stratifying patients for infection with *C. difficile* during the current admission. Since in this chapter we are interested in understanding the effects of external data on the target task, we consider the task of risk stratifying patients at the time of admission (instead of the more complex task of continuously stratifying patients throughout the admissions as in the previous chapter). We consider all inpatient visits for the two-year period between April 2011 and April 2013. This results in a total of 132,853 admissions and 1,348 positive cases of *C. difficile* (see Table 6.1 for the distribution across hospitals).
<table>
<thead>
<tr>
<th></th>
<th>Hospital A (%)</th>
<th>Hospital B (%)</th>
<th>Hospital C (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 21,959</td>
<td>n = 29,315</td>
<td>n = 81,579</td>
<td></td>
</tr>
<tr>
<td>Female gender</td>
<td>62.34</td>
<td>50.29</td>
<td>55.97</td>
</tr>
<tr>
<td>Age:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[0, 2)</td>
<td>14.38</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>[2, 10)</td>
<td>0.75</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[10, 15)</td>
<td>0.8</td>
<td>0.07</td>
<td>0</td>
</tr>
<tr>
<td>[15, 25)</td>
<td>7.23</td>
<td>3.77</td>
<td>6.73</td>
</tr>
<tr>
<td>[25, 45)</td>
<td>21.27</td>
<td>15.46</td>
<td>19.05</td>
</tr>
<tr>
<td>[45, 60)</td>
<td>21.28</td>
<td>30.98</td>
<td>22.77</td>
</tr>
<tr>
<td>[60, 70)</td>
<td>13.16</td>
<td>21.19</td>
<td>16.78</td>
</tr>
<tr>
<td>[70, 80)</td>
<td>10.79</td>
<td>15.97</td>
<td>13.74</td>
</tr>
<tr>
<td>[80, 100)</td>
<td>8.11</td>
<td>10.2</td>
<td>9.24</td>
</tr>
<tr>
<td>≥100</td>
<td>2.25</td>
<td>2.36</td>
<td>2.67</td>
</tr>
<tr>
<td>Hospital Admission Type:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newborn</td>
<td>13.13</td>
<td>0</td>
<td>8.74</td>
</tr>
<tr>
<td>Term Pregnancy</td>
<td>7.53</td>
<td>0</td>
<td>8.89</td>
</tr>
<tr>
<td>Routine Elective</td>
<td>15.87</td>
<td>31.28</td>
<td>17.39</td>
</tr>
<tr>
<td>Urgent</td>
<td>7.53</td>
<td>7.84</td>
<td>11.26</td>
</tr>
<tr>
<td>Emergency</td>
<td>10.79</td>
<td>15.97</td>
<td>13.74</td>
</tr>
<tr>
<td>Hospital Service:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicine</td>
<td>51.18</td>
<td>49.15</td>
<td>40.85</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>5.61</td>
<td>18.76</td>
<td>1.54</td>
</tr>
<tr>
<td>Surgery</td>
<td>7.53</td>
<td>5.97</td>
<td>10.28</td>
</tr>
<tr>
<td>Obstetrics</td>
<td>13.97</td>
<td>0</td>
<td>10.09</td>
</tr>
<tr>
<td>Cardiology</td>
<td>0</td>
<td>2.99</td>
<td>11.36</td>
</tr>
<tr>
<td>Newborn</td>
<td>13.15</td>
<td>0</td>
<td>9.01</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>0</td>
<td>13.11</td>
<td>3.7</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>3.06</td>
<td>5.32</td>
<td>6.76</td>
</tr>
<tr>
<td>Diabetic</td>
<td>24.44</td>
<td>32.73</td>
<td>33.59</td>
</tr>
<tr>
<td>CDI (C. difficile infection)</td>
<td>0.8</td>
<td>1.08</td>
<td>1.05</td>
</tr>
<tr>
<td>Previous Visit in last 90 days</td>
<td>5.87</td>
<td>7.43</td>
<td>5.54</td>
</tr>
</tbody>
</table>

Table 6.2: Descriptive statistics comparing the study population across the three different institutions
### Variable Type

<table>
<thead>
<tr>
<th></th>
<th>Set1</th>
<th>Set2</th>
<th>Set3</th>
<th>Set4</th>
<th>Set5</th>
<th>Set6</th>
<th>Set7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admission Details*</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>A,B,C</td>
<td>A,B</td>
<td>A,C</td>
<td>B,C</td>
</tr>
<tr>
<td>Patient Demographics**</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>28</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Patient History***</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>52</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Previous Visit Statistics (LOS)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Medications from Previous Visit</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>137</td>
<td>30</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Home Medications</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Attending Doctor Identification Number</td>
<td>27</td>
<td>30</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Location Units</td>
<td>10</td>
<td>14</td>
<td>62</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td>45</td>
<td>46</td>
<td>177</td>
<td>256</td>
<td>31</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>

*includes complaint, source, hospital service, expected surgery, month
** includes age, marital status, race, sex, city and financial class
*** includes previous diagnoses (ICD9 codes), and history of *C. difficile*

Table 6.3: The variables considered can be grouped into the broad categories given here. Each column gives the number of features within a category pertaining to the feature subset (see Figure 6-1).

![Venn Diagram](image)

1. Specific to A : **45**
2. Specific to B : **46**
3. Specific to C : **177**
4. Common to A, B & C : **256**
5. Common to A & B (only) : **31**
6. Common to A & C (only) : **12**
7. Common B & C (only) : **11**

Figure 6-1: The data for each hospital lie in a different feature space. Here we give the amount of overlap among the different institutions.

Because we are interested in stratifying patients by risk at the time of admission, we consider only data available at admission. For admissions from all three hospitals, we extract observations or features pertaining to the categories listed in Table 6.3. We map all features to binary-valued observations and remove variables that do not occur in at least 1% of at least one hospital’s population. This preprocessing results
in 578 binary features (summing the last row of Table 6.3): 256 shared by all three hospitals. The remaining features are specific to either a single hospital or shared by two hospitals. Figure 6-1 shows a labeling of the sets of shared and specific features across the different hospitals. Table 6.3 gives more detail regarding the types of features present across the three different hospitals.

6.2.2 Risk Stratification

To preserve interpretability, we consider learning in the context of linear classifiers.

We formulate the problem as follows. We have $N$ datasets:

$$D_j = \{(x_{j,i}, y_{j,i}) | x_{j,i} \in X_j, y_{j,i} \in \{-1, 1\} \}$$

where $j = 0, \ldots, (N-1)$. $D_0$ represents the target task and $D_1, \ldots, D_{(N-1)}$ represent the source tasks. $n_j$ represents the number of labeled examples available from each task. The binary classification goal is the same for each task. However, we must contend with different sets of variables, which we refer to as feature spaces $X_0, \ldots, X_{(N-1)}$.

We assume that there is some overlap between the features spaces for each of the source tasks and the target task under consideration, i.e., $\forall 1 \leq i \leq N-1, X_0 \cap X_i \neq \emptyset$.

Figure 6-2 depicts the possible intersection among feature spaces for a specific target task when $N = 3$. In logistic regression, we seek a function $f : \mathbb{R}^d \rightarrow [0, 1]$ of the form:

$$f(x_i) = \frac{1}{1 + e^{-\langle b_0 + w^T x_i \rangle}}$$

where $w \in \mathbb{R}^d$ (and $x \in \mathbb{R}^d$). Solving for the regression coefficients $w$ and $b_0$ is a straightforward maximum likelihood estimation problem. To improve generalizability, we consider $L2$-regularized logistic regression, where $\lambda$ is a tuning parameter.

$$\min_w \frac{\lambda}{2} ||w||^2 + \sum_{i=1}^{n} \log(1 + e^{(-y_i w^T x_i)})$$  \hspace{1cm} (6.1)

Note that we add an extra constant dimension to $x$ and compute the offset $b_0$ implicitly. The solution to (2) depends on the $X$ and $y$ employed in the training. Here we describe three potential solutions:

**Target-Only** This approach is a standard single-task approach and uses only data from the target task (i.e., $D_0$). This results in a solution of the form $w \in X_0$. 

94
When \( N = 3 \) the target feature space contains four different sets of features: features common to all tasks, target-specific features, and features shared only between the target and one of the sources.

This is identical to the approach employed in Chapter 3 and can easily overfit if the number of training examples from the target task is small.

**Source-Only** Given only labeled data from the source tasks, this approach seeks a solution by exploiting the shared features of each of the source tasks and the target tasks. The solution to this approach lies in the union of the intersections of the target feature space and each of the source features spaces (e.g., regions 4, 5 and 6 in Figure 6-2). The solution is of the form \( \{w_c, v_1, v_2, \ldots, v_{(N-1)}\} \), where \( w_c \) represents the common features shared among all tasks (i.e., region 4 in Figure 6-2, and \( \forall 1 \leq i \leq (N - 1) \), \( v_i \) represents the features shared only between the target and the source \( D_i \) (i.e., regions 5 and 6 in Figure 6-2). We rewrite the objective function in (6.1) to incorporate data from different sources. Here source data are mapped to the common feature space shared with the target task by removing all source-specific features.

\[
\min_{w_c, v_1, \ldots, v_{N-1}} \frac{\lambda_c}{2} \|w_c\|^2 + \sum_{j=1}^{N-1} \left( \frac{\lambda_j}{2} \|v_j\|^2 + \sum_{i=1}^{n_j} \log(1 + e^{-y_j[w_c; v_j]x_i})) \right)
\]

Because this solution depends on only source data, target-specific features (e.g., region 1 in Figure 6-2) will have no effect on the classification of a test patient.
**Source+Target** With Source+Target, we extend the solution described above to incorporate the target data, and the target-specific features.

\[
\min_{\mathbf{w}_c, \mathbf{v}_0, \ldots, \mathbf{v}_{N-1}} \frac{\lambda_c}{2} \|\mathbf{w}_c\|^2 + \frac{\lambda_0}{2} \|\mathbf{v}_0\|^2 + \sum_{i=1}^{n_0} \log(1 + e^{-y_{0_i}[w_{c_1}; v_{0_1}; \ldots; v_{N-1}]^T x_{0_i}})
\]

\[
+ \sum_{j=1}^{N-1} \left( \frac{\lambda_j}{2} \|\mathbf{v}_j\|^2 + \sum_{i=1}^{n_j} \log(1 + e^{-y_{j_i}[w_{c_1}; v_{j_1}; \ldots; v_{N-1}]^T x_{j_i}}) \right)
\]

This approach assumes a solution of the form \([w_{c_1}; v_0; \ldots; v_{(N-1)}]\) where \(v_0\) pertains to target specific features. The final solution \([w_{c_1}; v_0; \ldots; v_{(N-1)}] \in X_0\) as in the Target-Only, approach but incorporates data from all tasks. Note that if \(\forall j \lambda = \lambda_c = \lambda_j\) we can rewrite the objective function of Equation 6.3 as:

\[
\min_{\mathbf{w}_t} \frac{1}{2} \|\mathbf{w}_t\|^2 + \sum_{i} \log(1 + e^{-y_i w_t^T x_i})
\]

where \(\mathbf{w}_t = [w_{c_1}; v_0; \ldots; v_{(N-1)}], \ y = [y_0; \ldots; y_{(N-1)}], \ X = [X_0; X'_1; \ldots; X'_{(N-1)}].\)

The target data are used in their original form while the source data undergo two transformations. First, it is mapped to the common feature space \(X_j \rightarrow X'_j\) (removing source-specific features) and then mapped to the target feature space \(X'_j \rightarrow X''_j\) (by augmenting with zeros). Transforming the data in this way renders the objective function analogous to (6.1). Note that a similar transformation can be applied to the data of (6.2). These target-specific transformations to the data allow for transfer of knowledge across hospitals.

### 6.3 Experiments and Results

In this section we present a series of experiments in which we investigate the applicability of each of the learning approaches described in the previous section. In each subsection we analyze different aspects of the problem in order to gain insight into how and when to transfer knowledge across hospitals (see Table 6.4).
### 6.3.1 Including Source Data Helps

Eventually, we will consider risk stratification for *C. difficile* at each of the three hospitals. To start, we consider the task of risk stratification at the smallest hospital (i.e., Hospital A). I.e., Hospital A is the target task and Hospital B and Hospital C are the source tasks. We split the data for each hospital temporally into data from the first year and data from the second year (see Table 6.1 in Section 6.2.1). In all of the experiments, we train on data from the first year and test on data from the second year.

As described in Section 6.2.2, depending on the scenario considered, the training set consists of all or only a subset of the available training data. The dimensionality of the classifier learned depends on the origin of the training data. Both the Target-Only and Source+Target approaches incorporate data from the target task, so the dimensionality is that of the target task. When data from only the source tasks (Source-Only) is used, the features that occur only at the target hospital (target-specific) are ignored and the solution lies in a lower dimension (see Figure 6-2).

<table>
<thead>
<tr>
<th>Section</th>
<th>Training Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.1</td>
<td>Target-Only</td>
<td>Compares the three approaches presented in the previous section.</td>
</tr>
<tr>
<td></td>
<td>Source-Only</td>
<td>Measures the importance of the target specific features to each target task. We keep the training and test data constant but vary ( d ), the dimensionality of the solution ( D_1 &gt; D_2 &gt; D_3 ).</td>
</tr>
<tr>
<td></td>
<td>Target-Only:</td>
<td>Measures the effect of having a small amount of data from the target task versus twice as much data from the source tasks.</td>
</tr>
<tr>
<td>6.3.2</td>
<td>( d = D_1 )</td>
<td>Source-Only: Investigates the relative contribution each source (Source1 and Source2) makes to the target task, by considering each source independently.</td>
</tr>
<tr>
<td></td>
<td>( d = D_2 )</td>
<td>Source2</td>
</tr>
<tr>
<td></td>
<td>( d = D_3 )</td>
<td>Source1</td>
</tr>
</tbody>
</table>

Table 6.4: Outline of experiments presented in the remainder of this section. All experiments are repeated three times such that each hospital is considered as the target task.
Using LIBLINEAR [92], we learn three different risk prediction models based on the approaches described in the Section 6.2.2. We apply each classifier to the same holdout set (data from the second year) from Hospital A. We select hyperparameters using 5-fold cross-validation on the training set. We set the hyperparameters equal to one another, as in Equation 6.4. While this assignment is not optimal, it makes training a model more efficient, because otherwise optimization would require a search over three dimensions to find $\lambda_0, \lambda_1$, and $\lambda_2$.

![Graph showing AUROC results for Target Task A, B, and C](image)

Figure 6-3: Results of applying all three approaches to each of the target tasks. In each case the source data pertain to data from the other two hospitals. AUROC, area under the receiver operating characteristic (curve)

The results of this initial experiment are shown in Figure 6-3(a) and Table 6.5 (denoted by Target Task A). We give the performance on the holdout set in terms of the AUROC curve, the area under the precision recall curve (AUPR), the breakeven point where precision=recall, and finally the odds ratio (using a cutoff based on the 95th percentile). We calculated the 95% confidence intervals (CI) using 100 bootstrapped samples from the holdout set. Comparing the performance of three classifiers in Figure 6-3(a), we see that the classifier learned solely on data from the target task (i.e., Target-Only) performs the worst. When data from hospitals B and C is included in the training set, we see a significant improvement in performance. These results demonstrate how auxiliary data can be used to augment hospital-specific models. Hospital A has only 82 positive training examples, compared to hospitals B and C with a combined 587 positive training examples. These additional positive examples help the model generalize to new data.
In Figure 6-3(a), Source-Only and Source+Target perform almost identically. This could be because 1) the relatively small amount of added data when training the Source+Target classifier is not be enough to have a significant influence on the performance, and/or 2) the target task (Hospital A) does not differ significantly from the source tasks (Hospital B and Hospital C).

To explore how the amount of available training data from the target task affects the relative performance of the three approaches, we repeat the experiment described above but using hospitals B and C as target tasks. The results of these additional experiments are displayed in Figures 6-3(b) and 6-3(c) and Table 6.5. Figure 6-3(b) shows the improvement that comes with using the Source+Target approach over the Source-Only approach. This difference is amplified in Figure 6-3(c) where the Source-Only approach does significantly worse compared to the other two approaches.

<table>
<thead>
<tr>
<th>Target Task</th>
<th>Approach</th>
<th># Training Examples (pos)</th>
<th>AUROC (95% CI)</th>
<th>AUPR (95% CI)</th>
<th>Breakeven Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target-Only</td>
<td>Source-Only</td>
<td>54,142 (587)</td>
<td>0.8242 (0.80-0.85)</td>
<td>0.0656 (0.02-0.10)</td>
<td>0.1383</td>
</tr>
<tr>
<td></td>
<td>Source+Target</td>
<td>65,522 (669)</td>
<td>0.8239 (0.79-0.86)</td>
<td>0.0638 (0.03-0.09)</td>
<td>0.1489</td>
</tr>
<tr>
<td>A</td>
<td>Target-Only</td>
<td>14,675 (161)</td>
<td>0.8110 (0.78-0.85)</td>
<td>0.0664 (0.04-0.09)</td>
<td>0.1274</td>
</tr>
<tr>
<td></td>
<td>Source-Only</td>
<td>50,847 (508)</td>
<td>0.7907 (0.75-0.82)</td>
<td>0.0557 (0.03-0.08)</td>
<td>0.1146</td>
</tr>
<tr>
<td></td>
<td>Source+Target</td>
<td>65,522 (669)</td>
<td>0.8219 (0.80-0.85)</td>
<td>0.0699 (0.04-0.10)</td>
<td>0.1656</td>
</tr>
<tr>
<td>B</td>
<td>Target-Only</td>
<td>39,467 (426)</td>
<td>0.8142 (0.80-0.83)</td>
<td>0.0526 (0.04-0.06)</td>
<td>0.0958</td>
</tr>
<tr>
<td></td>
<td>Source-Only</td>
<td>26,055 (243)</td>
<td>0.7428 (0.72-0.77)</td>
<td>0.0356 (0.03-0.04)</td>
<td>0.0818</td>
</tr>
<tr>
<td></td>
<td>Source+Target</td>
<td>65,522 (669)</td>
<td>0.8114 (0.79-0.83)</td>
<td>0.0518 (0.04-0.06)</td>
<td>0.1051</td>
</tr>
</tbody>
</table>

Table 6.5: Result of applying the three approaches to each hospital.

These results show that the relative performance of the three approaches differs depending on the target task at hand. When ample data from the target task is available (e.g., target task C), ignoring the target data can significantly hurt performance. This result highlights the importance of including available target-specific data when training a model.
6.3.2 Target-Specific Features Are Important

Across all three hospitals (i.e., target tasks), we note that the Source+Target approach performs at least as well as the other classifiers. The Source+Target approach jointly incorporates all of the available training data and all of the features relevant to the target task. In the next set of experiments, we investigate how the inclusion or exclusion of target-specific features affects classifier performance.

For each task, we learn three different classifiers on the same training data but in different feature spaces. First, we train a Target-Only classifier, as in the previous experiments using the available target training data for each of the target tasks. Next, we learn two additional Target-Only classifiers but in a lower dimensionality compared to the first classifier. For example, consider target task A, the first classifier (A1) learns a solution using all of the features available to task A (i.e., the union of sets 1, 4, 5, and 6 in Figure 6-2), the second classifier (A2) ignores the target-specific features (i.e., it uses sets 4, 5, and 6), while the final classifier (A3) considers only features common to all tasks (i.e., set 4). In doing so, we control for the amount of training data and any changes in underlying distributions that could influence performance on the holdout data (e.g., relationship between the conditional or marginal distributions of the source and target data). For the three classifiers, the training data and test data are identical except for the set of features considered.

The results of this experiment are shown in Figure 6-4. The trend across all three tasks is the same, fewer features leads to worse performance. The detrimental effect of removing the target specific features is most noticeable for target task C. Hospital C has 177 hospital-specific features not found at the other two hospitals. Ignoring these target-specific features leads to a significant drop in performance from an AUROC of 0.81 (95% CI 0.80-0.83) to an AUROC of 0.78 (95% CI 0.75-0.79). The removal of the target-specific features at the other two hospitals has less of an impact on performance. For hospitals A and B there are fewer target-specific features (45 and 46 features respectively) and less target-specific training data. This could explain why there is no significant difference between the AUROC achieved by the Source-Only
Figure 6-4: Here the amount of training and test data are kept constant but the dimensionality of the solution varies.

and Source+Target approaches for these two target tasks (see Table 6.5).

In a follow-up experiment, we learned a single classifier by pooling all of the data and searching for a solution in the common feature space (i.e., region 4 in Figure 6-1). Applied to the held-out data from task A, B, and C we achieve an AUROC of 0.82 (95% CI 0.78-0.86), 0.79 (95% CI 0.76-0.84), and 0.77 (95% CI 0.74-0.79) respectively. This straightforward approach, ignores the target-specific features and as we might expect results in a worse performance relative to the Source+Target approach.
6.3.3 More Data is Not Always Better

In our next experiment we compared three different models for each hospital (1) a Target-Only model using a random sample of 5,000 admissions, (2) a Target-Only model at each hospital using a random sample of 10,000 admissions, (3) a Source-Only model using a random sample of 5,000 admissions from each of the two hospitals. The average performance across ten repetitions is shown in Figure 6-5.

For Hospitals B and C having a small amount of data from the target task is better than having twice as much data from the source task. However, for Hospital A the Target-Only approach does not outperform the Source-Only approach despite the same amount of training data. These two approaches seek solutions in different feature spaces. The Target-Only approach seeks a solution in a higher dimensionality. Moreover, the Target-Only approach samples from a much smaller pool of samples compared to the Source-Only approach. Therefore, the high dimensionality combined with the smaller pool of samples likely contributes to our inability to learn more informative feature weights using the Target-Only approach.

![Figure 6-5: The results of experiments from 'More data are not always better', applied to each target hospital. In two out of three hospitals it is better to have data from the target hospital than twice as much data from the source hospitals.](image)
6.3.4 Not All Transfer is Equal

Comparing the results for target tasks A and B, the Source-Only approach appears to work better for target task A than it does for target task B. The amount of training data used in training a classifier for Hospital A is only slightly greater than Hospital B (54,142 versus 50,847). This raises the question of whether the discrepancy in performance is simply due to the effect of having 6.5% more data or due to differences in the underlying similarities between the source and target tasks. Data from Hospital C are included in the training data for both target tasks, but it may be the case that data from Hospital C transfer more readily to target task A than to target task B.

To investigate this question, we apply the Source-Only approach to each of the three target tasks. However, instead of combining data from the two available source hospitals we learn a model for each source independently (while controlling for the amount of training data) and apply it to the target task. The results of this experiment are shown in Figure 6-6. The source of the training data is denoted along the x-axis in Figure 6-6. We control for the amount of training data available at each hospital by randomly undersampling data from the larger hospital.

Figure 6-6: Here only data from a single source task are used to learn a model, which is then applied to the target task. The ' indicates that the amount of data used in training was limited to the amount of data available from the other source.
These results suggest that data from Hospital C might transfer more readily to Hospital A than data from Hospital B, despite the fact that Hospital A and Hospital B have more features in common. This observation is supported by the last pair of bars in Figure 6-6: for target task C a classifier trained on only data from Hospital A outperforms a classifier trained on only data from Hospital B. This suggests that out of the three hospitals, Hospital B is the most different. And this could explain why, despite the large amount of training data, the Source-Only approach performs relatively poorly for target task B, but performs well for target task A (see Figures 6-3(a)(b)). Additionally, as alluded to earlier, these relationships could explain why the performance of the Source-Only and Source+Target approaches are almost identical for target task A.

6.4 Discussion & Conclusion

In the previous section, our experiments were limited to three hospitals ($N=3$). With over 5,000 registered hospitals in the US alone, larger numbers of $N$ are feasible. Opportunities for scaling raise several important considerations and implications. First, when $N$ grows the number of features common to all hospitals will shrink and therefore the number of hospital-specific features will increase. Limiting models to only the shared feature set (as in [96]) risks ignoring possibly crucial hospital-specific information. Second, as $N$ grows the variation among hospitals will increase. Even for hospitals within the same network, we found that the transferability of knowledge was neither equal nor symmetric among hospitals. As the variation among tasks increases, it is plausible that including auxiliary data when training a model could actually diminish performance on the target task. Future work is needed to investigate how to best select source data from a large pool of hospital databases. Depending on the task, this could mean selecting the best subset of hospitals, or the best subset of data from each hospital. Third, as $N$ grows the number of hyperparameters grows. Each hyperparameter controls the extent to which data from each hospital contributes to the final model. Procedures for identifying an optimal setting for hyperparameters can
quickly become inefficient with increasing \( N \), posing new challenges and opportunities in machine learning.

It is important to note that this work focuses on how tasks vary in terms of the definition of the feature space, and not explicitly on how data may be distributed differently across these feature spaces. As mentioned in the section on related work, researchers have proposed many different strategies to address the latter issue. In future work, methods for dealing with differences in conditional and marginal distributions across tasks could be employed, not instead of, but in addition to the solution we presented here for transferring data across different but overlapping feature spaces.

We have presented methods and experiments using data from three hospitals to understand the potential gains and challenges associated with leveraging data from external hospitals in building hospital-specific predictive models for \textit{C. difficile} infections. While there is no global model for the prediction task we considered, the inconsistent performance of the Source-Only approach across target tasks is indicative of why national models often perform poorly when applied to specific institutions [96]. Unsurprisingly, auxiliary data tends to have the biggest impact when the number of target training examples is small, the number of shared features is large and there is significant overlap in the shared feature space. When ample data from the target space is available, our results demonstrate the importance of including target-specific data, and target-specific features when training hospital-specific risk-stratification models.

Our findings highlight the promise of leveraging external data for building models at specific hospitals at which predictions will be used. Moving forward, we believe that further study of techniques that facilitate the incorporation of all available data across hospitals and databases should be a top priority in efforts to construct and harness predictive models in healthcare.
Chapter 7

Translation to Clinical Practice

7.1 Introduction

Beyond causing significant morbidity and mortality, each case of *C. difficile* infection is associated with the addition of thousands of dollars in hospital costs for primary infections and tens of thousands of dollars per case for recurrent infections [106]. Despite much effort, *C. difficile* rates in the US have increased in recent years [51]. This is partly due to the lack of cost-effective interventions and of means for focusing interventions based on risk.

In the absence of effective risk stratification, widespread implementation of known interventions (such as isolating patients) is prohibitively expensive. Only a small fraction of patients become infected with *C. difficile* during their hospitalization. Thus, any intervention applied to the entire population is likely to have no effect in at least 99% of the admissions, and therefore not cost-effective. Clinical risk tools, such as the one proposed in our work, could allow clinicians and epidemiologists to target high-risk patients with a number of different interventions. We believe that augmenting current guidelines with specific interventions aimed at patient’s with a risk exceeding a specific threshold could reduce the incidence of *C. difficile* in a cost-effective manner.

In addition to helping match high-risk patients with the appropriate intervention, our risk-stratification tool could be used in the design of clinical trials. To date, there
have been numerous studies investigating the effectiveness of different interventions for preventing *C. difficile* infections. However, because of the complexity and relative rarity of the disease, these studies have been underpowered and inconclusive. Focusing on only patients at high risk would allow clinicians to efficiently test interventions on the population most likely to benefit from an intervention.

As a first step towards having this kind of clinical impact, we have integrated a version of our model (the multitask model from Chapter 5) into the health record system at Hospital C. The risk score is computed by a scheduled task that runs once a day (shortly after midnight) on patient data automatically drawn from the hospital’s health information system. We provide an updated risk score that is displayed as part of the patient information and available throughout Hospital C. Figure 7-1 is a screenshot of the Amalga clinical interface, showing the risk score of those patients at greatest risk for infection on a day in the winter of 2014. Please note that despite the header “cDiff Probability”, this value is technically a percentage (the probability was multiplied by 100).

We are now exploring ways in which this risk tool could be incorporated into current clinical practice in order to have the greatest positive impact. In this chapter, we begin by reviewing the current guidelines for prevention and control of *C. difficile* infections, and suggest ways in which our tools could be used to augment current efforts. At the end of the chapter, we present a cost-benefit analysis illustrating the potential impact our model could have if used to guide interventions.

### 7.2 Review of Current Clinical Guidelines

Current guidelines proposed by the Association for Professionals in Infection Control (APIC) aim to reduce the recurrence of the disease in patients with a *C. difficile* infection (CDI), and limit the transmission of the disease [107]. In general, the APIC recommendations apply to patients who have already been identified as infected with *C. difficile*, instead of focusing on those at risk of becoming infected with *C. difficile*.

In recent years, researchers have also proposed prophylactic interventions that
We have integrated a version of our risk-stratification algorithm into the health information system at Hospital C. This figure shows a screenshot of the electronic health record system employed by the physicians. For privacy reasons we have blacked out the account number, the admit date and the patient name.

aim to reduce a patient’s risk of CDI. In this section, we review the efficacy of several interventions commonly proposed in the clinical literature, including: antimicrobial stewardship programs, cessation of proton pump inhibitors (PPIs), environmental cleaning, contact precautions, and probiotics. We present the current APIC guidelines as they apply to the intervention, in addition to an overview of relevant clinical studies. At the end of each subsection, we describe how our risk-stratification model could be used to either guide the intervention or design more effective clinical trials.

7.2.1 Antimicrobial Stewardship

Current APIC Guideline: recommends the implementation of a program that supports the prudent use of antimicrobial agents.
There is strong evidence for the link between antibiotic exposure and acquisition of CDI [108–112] [113] (e.g., clindamycin, cephalosporins, and fluoroquinolones are known risk factors for the disease). These studies suggest that antimicrobial stewardship programs should focus on antimicrobial selection, dosing and duration. Antimicrobial stewardship programs are defined as overarching programs that aim to change the usage of antimicrobials [114].

Current guidelines recommend discontinuing any antimicrobial agents (unless absolutely indicated) if the patient has a CDI [115]. These recommendations were based on strong evidence from a meta-analysis of 12 observational studies and randomized controlled trials (RCTs), showing that the continued use of antimicrobials for infections other than CDI is significantly associated with recurrence of CDI [116].

There have been many different studies investigating the relationship between antimicrobial stewardship programs and outcomes. A recent interventional study found that restricting the use of high-risk antibiotics led to a significant change in the incidence trend of CDI (compared to the incidence prior to the intervention). However, there was no significant change in the incidence rate [117]. In a separate study, a policy banning the routine use of ceftriaxone and ciprofloxacin led to a reduction in the incidence of CDI of 70% [118]. In this study, empirical prescription of ceftriaxone for systemic sepsis and surgical prophylaxis was swapped for amoxicillin, gentamicin and metronidazole. (Note, metronidazole is also a treatment for CDI.) Another study considered the effect of introducing guidelines for replacing high-risk broad-spectrum antimicrobials with low-risk antimicrobials (e.g., penicillin, clarithromycin, doxycycline, gentamicin, vancomycin, trimethoprim, and nitrofurantoin) [119]. The change led to a significant reduction in the use of fluoroquinolones and cephalosporins, and led to a significant decrease in CDI.

In addition to programs focused on substituting low-risk antimicrobials for high-risk antimicrobials, there have been a number of studies investigating the relationship between duration of treatment with antimicrobials and CDI. Prolonged exposure to antimicrobials has been linked to higher risk of CDI [113]. A possibly effective antimicrobial stewardship program is one that stresses the adherence to shorter durations
of therapy [120]. It has also been suggested that one switch from parenteral therapy to oral therapy (in part to shorten the duration of antimicrobial therapy) [120]. A study investigating the effect of nonrestrictive measures to optimize antibiotic usage and duration, showed a marked reduction in CDI by 60% [121].

While the positive effect of swapping out high-risk antimicrobials for lower-risk antimicrobials is clear, the negative consequences are not as straightforward. While increasing the risk of CDI may be an undesirable side effect of these drugs, the risk may be outweighed by the benefit of the primary effect. Therefore, at the level of the individual patient, careful consideration should be taken when modifying patient treatment. Our risk-stratification tool could help guide antimicrobial stewardship programs, by alerting physicians to high-risk patients receiving high-risk antimicrobials.

7.2.2 PPI Cessation

Current APIC Guideline: acknowledges the existing research but makes no recommendations regarding the cessation or restriction of PPIs

There has been some exploration into the pathophysiology of PPI receipt and CDI [122]. It is hypothesized that gastric acid suppression leads to the survival of \textit{C. difficile} in vegetative form. However, there are no RCTs to solidify the causality of the effect, because of the ethical concerns surrounding such a trial. While no RCTs have investigated the relationship between PPIs and CDI there has been retrospective research aimed at exploring this relationship.

Recently, a meta-analysis investigated the relationship between \textit{C. difficile}-associated diarrhea (CDAD) and receipt of PPIs. The analysis showed a 65% increase in the incidence of CDAD among patients receiving PPIs [47]. More recently, a study of the epidemiology of community-associated CDI explored the potential sources of \textit{C. difficile} acquisition in the community [123]. In an analysis of 984 patients with community-acquired CDI, 31% of patients without antibiotic exposure received a
PPI. The study concluded there was a strong association between PPI receipt and CDI. Finally, a retrospective cohort study found recurrent CDI was more common in those exposed to PPIs than in those not exposed [49].

These retrospective studies are limited since they cannot fully control for underlying disease [124]. Patients who receive PPIs and patients who develop CDI are sicker than those who do not. Further investigation is needed before we can conclude anything regarding a causal link between PPI exposure and CDI.

Nevertheless, these studies suggest that prescribing PPIs to high-risk patients warrants careful consideration [125]. Recommendations in the literature range from completely discontinuing PPIs during hospitalization [126], to lowering the dosage while including other protective precautions [127]. Since it has been suggested that PPIs may affect patients with minimal or no antimicrobial exposure, cessation or restriction of PPIs could benefit patients who are not receiving antimicrobials, but are still classified as high risk by our model.

### 7.2.3 Environmental Cleaning

| Current APIC Guideline: | recommends EPA-approved germicides for routine disinfection during non-outbreak situations, during outbreaks a 1:10 dilution of 5.25% sodium hypochlorite solution is recommended for disinfecting CDI rooms. |

Without any exposure to the *C. difficile* organism, primary CDI cannot occur regardless of exposure to antimicrobials or PPI. Transmission from contaminated surfaces to patients can happen directly or through the contaminated hands of healthcare workers. Proper cleaning to eliminate *C. difficile* spores from the hospital environment is important in the effort to reduce the incidence of CDI [128–130].

Cleaning of environmental surfaces typically involves the manual application of a detergent disinfectant. Cleaning contaminated rooms with dilute bleach upon discharge was associated with a 48% reduction in CDI [131]. Newer non-touch disinfection methods include hydrogen peroxide vapor, and automated germicidal ultraviolet
irradiation. Cleaning via hydrogen peroxide vapor has recently been associated with an 80% reduction in acquisition of Vancomycin Resistant Enterococcus (another type of healthcare-associated infection) among previously contaminated rooms [132]. The study also showed a reduction in the number of cases of CDI, however the result was not significant.

A separate prospective study at a VA hospital considered 3 sequential interventions [133]. These included 1) fluorescent markers to provide monitoring and feedback on thoroughness of cleaning facility-wide, 2) addition of an automated ultraviolet radiation device for disinfection of CDI rooms, and 3) enhanced standard disinfection of CDI rooms. The proposed interventions led to a reduction in positive cultures of CDI by 14%, 48%, and 89% with each intervention respectively. These results suggest enhanced sequential disinfection of CDI rooms as a valuable tool in reducing the incidence of CDI.

However, such enhanced cleaning programs can be costly and labor intensive. Moreover, hypochlorite solutions or highly concentrated hydrogen peroxide solutions are associated with corrosion and pitting of equipment/surfaces over time, and are known to cause respiratory difficulties. Therefore APIC does not recommend their use for the cleaning of all CDI rooms. However, our risk model could perhaps be used to identify the rooms or units most likely to benefit from enhanced daily cleanings.

In a study comparing the sequencing of isolates obtained from 1250 symptomatic patients infected with C. difficile, it was concluded that the majority of cases of CDI were not transmitted from another symptomatic patient [134]. This is an important finding, since it suggests transmission might be occurring through asymptomatic patients or through some other environmental reservoir. In our proposed risk model we attempt to capture exposure by including colonization pressure as a variable. However, the results of this study suggest tracking exposure might be more difficult than initially thought.
7.2.4 Isolation/Contact Precautions/Hand hygiene

**Current APIC Guideline:** recommends presumptive isolation and contact precautions in symptomatic patients while awaiting test results. Infected patients should be isolated in a room with a private bathroom until 2 days after diarrhea stops. Use of alcohol-based hand rubs (ABHRs) for hand hygiene during routine infection prevention is recommended. However, traditional hand washing is preferred when hands are visibly soiled.

Standard precautions are recommended for all patients regardless of diagnosis, however as the APIC guidelines recommend, infected patients should be moved to a private room. Double rooms are significantly associated with higher risks of CDI compared to private rooms [135]. Current guidelines suggest discontinuing special contact precautions (gloves and gowns before entering the patient’s room) after symptoms disappear. However, during an outbreak (or a period of high risk) this duration may be increased.

The impact of hand hygiene on infection rates has been studied extensively [136]. In recent years there has been some controversy regarding ABHRs versus traditional soap and water hand washing. It has been shown the traditional hand washing is more effective at removing spores compared to ABHRs [137]. However, completely replacing ABHRs with traditional hand washing is not recommended, since it has a lower adherence rate (because it is more time consuming) [138]. Guidelines recommend the use of soap and water during outbreaks when caring for patients with CDI. With the incorporation of a risk-stratification tool it may be possible to efficiently extend these guidelines to all high-risk patients (not only those already infected).
7.2.5 Probiotics

**Current APIC Guideline:** does not recommend the use of probiotics for the prevention of primary or recurrent *C. difficile*.

*C. difficile* takes over the gut when normal flora that might compete with the harmful bacteria get wiped out. Thus, administration of probiotics has been proposed as a possible intervention. There have been several studies in recent years investigating the effectiveness of probiotics in preventing CDI [41,139–142].

The authors of [139] conducted a multi-institution randomized double blind placebo controlled study including 135 hospital patients taking antibiotics. The intervention was a 100g drink containing *Lactobacillus casei*, *Lactobacillus bulgaricus*, and *Streptococcus thermophiles*. Results seemed promising, as no one in the probiotic group developed CDI and 17% of the placebo group did. However, the study has been criticized for its strict exclusion criteria, including excluding all patients receiving high-risk antibiotics. The exclusion criteria removed approximately 72% of the eligible patients, and therefore the generalizability of the results has been questioned.

A meta-analysis reviewing 20 RCTs testing the effectiveness of probiotics (*Bifidobacterium*, *Lactobacillus*, *Saccharomyces*, or *Streptococcus*) suggests that probiotics reduce the incidence of CDAD by 66% [41]. The authors conclude there is moderate-quality evidence suggesting that probiotic prophylaxis results in a reduction of CDAD without an increase in adverse events. Of the 20 included trials, 13 excluded immunocompromised patients, limiting the population for which the conclusions apply.

In a separate meta-analysis, authors considered the efficacy of probiotics for reducing recurrent CDI [142]. The analysis considered six RCTs (354 patients with *C. difficile*). Two of the six trials reported a significant reduction of recurrent CDI, while the other four did not show any significant difference. Of the different probiotics considered, only *Saccharomyces boulardii* was associated with a significant reduction in the recurrence of CDI. Still, the authors concluded that treatment with probiotics leads to a significant reduction in recurrent CDI, relative risk 0.59 (95%CI 0.41-0.85).
A separate meta-analysis focused on randomized placebo-controlled efficacy studies of probiotic use for the prevention of primary CDI (instead of recurrent CDI) [140]. Of the 11 RCTs reviewed, most were underpowered. Still, in 9 out of 11 studies, the trend was towards protection. The analysis found a consistent and significant beneficial trend for *Lactobacillus acidophilus*+*Lactobacillus casei* formulations. Among studies using *Lactobacillus acidophilus* and *Lactobacillus casei* the risk ratio was 0.21 (95% CI 0.11-0.42). In the studies using *Saccharomyces boulardii* the trend was the same, but not significant, risk ratio 0.70 (95%CI 0.29-1.69).

Recently, a large RCT (perhaps the largest of its kind to date), considered patients ≥ 65 years of age, exposed to one or more oral/parenteral antibiotics [141,143]. The study identified a trend toward reduced CDI, but ultimately the study was underpowered. There were just 12/1470 cases of CDI in the probiotic group and 17/1471 in the placebo, resulting in a risk ratio of 0.71 (95%CI 0.34-1.47).

Given the risks associated with probiotics (e.g., acquisition of probiotic-related infections particularly among immunocompromised patients [140]), and the large number of studies with inconclusive results [144], further study is needed before prophylactic probiotics can be used broadly in high-risk patients for either primary or recurrent CDI. Current guidelines stress that these organisms be used cautiously among patients with immune suppression [115].

A large majority of studies investigating the effect of probiotics on the incidence of CDI have been underpowered. By focusing on a subset of patients at higher risk of CDI, we hypothesize that our risk-stratification tool could help researcher design more cost-effective/conclusive clinical trials.
7.3 A Cost-Benefit Analysis

Risk-stratification models could be used to identify those patients most likely to benefit from the interventions described in the previous section. By targeting high-risk patients we could reduce the number of patients receiving an intervention who are unlikely to benefit, while reducing the total cost of applying the intervention. In this section, we present a hypothetical cost-benefit analysis to study the financial trade-offs that come with applying a targeted intervention.

We explore the problem in terms of the net financial gain. The net financial gain $G$ of a given intervention depends on several variables, listed in Table 7.1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>the number of patients in the reference population</td>
</tr>
<tr>
<td>$K$</td>
<td>the financial cost of each case of CDI</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>the cutoff of the risk-stratification metric (i.e., the fraction of $N$ receiving the intervention)</td>
</tr>
<tr>
<td>$\sigma_\gamma$</td>
<td>the positive predictive value of the risk-stratification metric</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>the effectiveness of the intervention</td>
</tr>
<tr>
<td>$\beta$</td>
<td>the cost of the intervention</td>
</tr>
</tbody>
</table>

Table 7.1: The net financial gain of applying an intervention to a targeted population depends on several variables.

Using these variables, we define the net financial gain of an intervention applied to a subset of patients (identified as high-risk by our risk-stratification model) using Equation 7.1.

$$G = \gamma N(\sigma_\gamma \alpha K - \beta)$$ (7.1)

The values of $N$ and $K$ vary across regions and hospitals, $\alpha$ and $\beta$ are determined by the chosen intervention, and the values of $\gamma$ and $\sigma_\gamma$ depend on the risk-stratification metric.

We consider applying our model for risk (from Chapter 5) at Hospital C over a one-year period in which there were 24,399 admissions ($N = 24,399$). During that time period 242 admissions were associated with a positive test result for $C. \text{difficile}$. Assuming $K=$$10,000, in Figure 7-2, we plot the maximum $\beta$ (i.e., cost of the intervention per patient) that would result in a net financial gain of $0$. We sweep
Figure 7-2: We plot the maximum $\beta$ (i.e., cost of an intervention) that would result in a net gain of $0 in terms of hospital costs. $\alpha$ is the effectiveness of the intervention.

the decision threshold along the x-axis. In addition, we vary $\alpha$, the effectiveness of the intervention ($\alpha = 0.1$ indicates that an intervention is effective in 10% of patients who would otherwise become infected with *C. difficile*). This analysis illustrates how targeting high-risk patients allows for more expensive interventions compared to the scenario in which an intervention is applied to the entire population. Note, this analysis considers only the net gain in terms of immediate hospital costs, and does not include the benefit to patients or the beneficial feedback caused by a reduction in colonization pressure.

Figure 7-3: We plot the estimated savings in terms of millions of dollars in hospital costs, assuming each case of *C. difficile* costs an additional $10,000 and the cost of the intervention is $20 per patient.

In Figure 7-3, we consider a more concrete example where we estimate the savings
in the millions of dollars that come with applying an intervention that costs $20 per patient (the approximate cost per visit of probiotics) to those patients at Hospital C in the top 5% as estimated by our risk-stratification model. As discussed above, the efficacy of probiotics in preventing infection with \textit{C. difficile} is still being investigated. Since recent studies have shown roughly a 30\% reduction in relative risk in the treatment group compared to the control group [41,143,145,146], we set \( \alpha = 0.3 \).

Applied to the 40\% of patients at highest risk of acquiring \textit{C. difficile} at Hospital C, such an intervention would cost approximately $195,200 a year. However, we would expect an overall 25\% reduction in the number of cases of \textit{C. difficile} and a savings of approximately $620,000 per year. This results in a net gain of $424,800 per year and a significant reduction in patient suffering.

Our analysis is optimistic in that it assumes that the intervention i.e., the receipt of probiotics, is not contraindicated in any of the patients identified by the model as high risk. However, as the literature reviewed in the previous section indicates, the intervention is not recommended in some immunocompromised patients.

In Figure 7-3 when the effectiveness of the intervention exceeds 20\% we note a savings even when applying the intervention to the entire population. Operating in this regime (i.e., giving every patient the intervention) would make the risk tool irrelevant. However, as mentioned above, interventions may have negative side-effects (not accounted for here) that make it infeasible or dangerous to apply at a population level. Moreover, the effectiveness of the intervention itself could depend on the number of patients to which it applies. For example, a recent study showed that placing 40\% of patients under contact precautions was a tipping point for noncompliance [147].

In addition to the cost of the intervention itself, future analyses should include the expected burden to the patient and to the healthcare staff.

Here, we studied a scenario in which only a single intervention is considered, however hospitals may want to consider bundles or sequences of interventions (e.g., isolating patients and enhanced environmental cleanings). Each intervention would be associated with a different decision threshold (or risk cutoff), depending on the cost and the effectiveness of the intervention.
The cost-benefit analysis presented above optimizes for monetary savings, however this may not be our goal. Instead, we may be interested in reducing the incidence of *C. difficile* disease with limited resources. For example, many hospitals have a limited number of private rooms. Our risk model could potentially help healthcare workers determine how to best distribute patients throughout the hospital.

In our work, we have focused on the clinical impact of identifying the high-risk patients, however our model also identifies a population of patients at very low risk of *C. difficile* infection. For MRSA (another healthcare-associated infection), it has been shown that actively screening for MRSA clearance leads to a significant increase in the discontinuation of MRSA contact precautions [148]. I.e., passive screening results in the unproductive, or even counter-productive, use of contact precautions. Depending on how hospital protocol for dealing with suspected *C. diff* changes over time, the ability to identify a population of patients at very low risk could prove clinically useful.

### 7.4 Summary & Conclusion

In this chapter, we showed how our model for patient risk stratification can be incorporated into a hospital’s health information system. We integrated our algorithm into the health record system at Hospital C. The risk of each patient in the hospital is automatically recomputed daily and made available to hospital workers. This has led us to explore different ways in which the risk tool could help guide clinical practice with the ultimate goal of reducing the incidence of *C. difficile*.

After reviewing the current APIC guidelines for preventing infections with *C. difficile*, we highlighted how our model could be used to target high-risk patients with prophylactic interventions.

Through a cost-benefit analysis, we illustrated the tradeoff between the cost of an intervention and the chosen risk cutoff (i.e., the risk at which physicians might decide to intervene), in terms of net monetary gain. The results of our analysis suggest that the selective targeting of high-risk patients, as identified by our model, could serve as
a cost-effective approach to reduce the incidence of \textit{C. difficile} in hospitals.

Beyond the targeting of high-risk patients, we believe that our risk-stratification model could aid in the design of clinical trials. In the current clinical literature, many of the randomized controlled trials end-up ultimately underpowered and inconclusive. However, our model could help identify a better study population (i.e., the population of patients most likely to benefit from the intervention), thereby reducing the total cost of the study while increasing the statistical power of the analysis.

While our proposed risk model can help identify patients who are at high-risk of acquiring an infection with \textit{C. difficile}, it does not explicitly state \textit{why} a patient is at high risk. In a clinical setting, knowing why a patient is at risk is important, since this can affect a physician’s decision of how to intervene. The development of interpretable models that provide explanations along with risk estimates is an important future research direction.
Chapter 8

Summary and Conclusion

In this dissertation we developed novel data-driven solutions for predicting potentially avoidable bad outcomes. In particular, we focused on the task of stratifying inpatients according to their risk of acquiring an infection with *Clostridium difficile* (*C. difficile*), a common type of healthcare-associated infection. This problem is made challenging by the large number of factors contributing to patient risk, the temporal aspects of the problem (i.e., both patient state and hospital conditions vary over time), the relatively small number of positive learning examples, institutional differences, and the lack of ground truth.

For some time now, researchers have directed efforts towards learning accurate risk-stratification models for *C. difficile*. In previous work, risk-stratification models for *C. difficile* were based on a small number of expert-chosen clinical risk factors. In addition, researchers ignored the temporal aspects of the problem and considered only time-invariant models. Finally, evaluations of such models have been limited to a single point in time. In contrast, in our work we took a data-centric approach to the problem, leveraging structured EMR data from over 100,000 admissions and addressing the temporal aspects of the problem.

In Chapter 3, we showed the advantage of taking a data-centric approach over a conventional expert-driven approach. Our model, based on over 10,000 variables extracted automatically from an EMR, significantly outperformed a model based on features chosen by experts. For comparison, consider the case where both models
correctly identified half of the positive cases correctly: our model resulted in over 3,000 fewer misclassified patients. One problem with our method is that the high-dimensionality and the correlation among the covariates leads to a model that is inherently less interpretable. In a clinical setting, interpretability can be important. Knowing why a patient is at high risk, is almost as critical as knowing if a patient is at high risk. Investigating techniques for improving the interpretability of the model without sacrificing classification performance will be a priority in future research.

In Chapter 4, we explored the benefit of posing the problem as a time-series classification task. We grouped the risk factors into two broad categories: time-invariant risk factors, which are collected at the time of admission and do not change over the course of the hospitalization (e.g., patient gender), and time-varying risk factors, which can change over the course of the hospitalization (e.g., medications). Our risk-stratification approach had two stages. In the first stage we extracted a time series of independent daily risk estimates for each patient, which we referred to as a risk profile. In the second stage we further distinguished high-risk patients from low-risk patients based on their risk profiles. Incorporating the evolving risk profile in this way led to significantly better predictions compared to the snapshot approach. Moreover, by framing the problem as a time-series classification task we raised the issue of measuring patient risk at multiple times during a visit. It is important to note that this approach was developed for/on patients with longer visits (of at least 7 days). Because of the temporal nature of the approach, we do not expect there to be a significant advantage for short visits.

In Chapter 5, we continued to explore the temporal aspects of the problem of patient risk stratification and broadened the study population (to include patients with shorter visits). We extended the model presented in Chapter 4 to allow for time-varying model coefficients. We segmented the problem into multiple tasks, where each task corresponded to a different time period during the visit (e.g., days 4 through 6). Next, through an adaptation of transfer learning techniques, we learned the tasks jointly. By learning the models jointly, we leveraged the inherent relatedness among the different tasks. We demonstrated a consistent improvement in the classification
performance using our multitask approach over the single-task approach (particularly among patients with longer risk periods). While consistent, the improvement was small. In future work, such models could be improved by considering additional temporal smoothness constraints.

In our work, we incorporated variables related to both susceptibility (i.e., the state of the patient) and exposure (i.e., the hospital conditions) into the risk model. In future work, it would be interesting to further study the separate effects of susceptibility and exposure on patient risk over time.

In Chapters 3 through 5 we considered data from a single hospital. In Chapter 6, we expanded our analysis to consider two additional hospitals within the same hospital network. In this work, we investigated the transferability of models across hospitals. By leveraging data across three hospitals (varying in size from 180 to over 900 beds), we performed a comparative analysis of the value of using external data to enhance hospital-specific predictions. This setup, where we train on data from a different hospital than the one we test on, violates the assumption that the training data and future data lie in the same feature space and share the same distribution. In Chapter 6, we focused on the problem of the data from different hospitals lying in different but overlapping feature spaces. We proposed a simple modification to the objective function that allowed us to leverage data from all three hospitals while seeking a solution in the target-feature space. Compared to hospital-specific and global models, our proposed approach achieved consistent performance across all three hospitals. We also demonstrated the importance of hospital-specific features (e.g., locations within the hospital). There is still a lot of work to be done on this front and going forward we hope that studying techniques for the incorporation of data across hospitals and databases will be a top priority in efforts to construct and harness predictive models in healthcare.

We proposed a new evaluation scheme motivated by a clinical use case in which the model is applied daily to evaluate each patient in the hospital. Evaluating the model in this way leads to a good approximation of how we expect the model to perform in practice. We also evaluated our final model in terms of how many days in
advance we could predict high-risk cases. In a clinical setting, how far in advance one can predict an infection is as relevant as classic performance metrics like sensitivity and specificity. In addition, we were careful to split our data temporally, learning on data from one year and evaluating the model on data from the next year. When evaluating predictive models in medicine it is especially important to ensure that all examples in the training set precede all examples in the test set. Failure to do so, which is all too common in the literature, can lead to misleading results.

We note that many of the issues that arise when transferring knowledge across hospitals, arise when transferring knowledge across time at an individual hospital. Over time, hospital populations, physical plants, tests, protocols, and staff change. Furthermore, electronic medical records change both in terms of what is collected and the precise meanings of variables. Investigating methods for incorporating past data into current models is an important future direction.

In Chapter 7, we discussed some of the practical implications surrounding the incorporation of our risk-stratification metric into clinical practice. We showed that our algorithm can be integrated into the health information system at a large hospital. Presently, our model is online and producing daily risk estimates for all inpatients. In addition, in Chapter 7 we reviewed the current guidelines on the prevention and control of \textit{C. difficile} and discussed their relationship to predictive models. Current guidelines focus on lowering risk by applying interventions at the population level. We proposed extending current efforts with more targeted interventions. To this end, we presented a hypothetical cost-benefit analysis, illustrating how our model could be used to identify those patients most likely to benefit from an intervention. We showed that by focusing efforts in prevention on high-risk patients, hospitals can reduce the incidence of infection while achieving a significant cost-savings.

Our analysis considered a single intervention, however there are many possible interventions and combinations. Studying the optimization of how and when to intervene is a challenging direction for future research.

In our work, we selected the model hyperparameters with the goal of maximizing the AUROC. In doing so, we sought the best “general” classifier, i.e., a classifier
that achieved good classification results at many different thresholds. However, in future work one might select a specific operating regime based on the hospital and the chosen intervention e.g., a limited number of private rooms may place a constraint on the allowed number of false positives. Given these constraints, one might select hyperparameters that optimize the classification performance for a region of the ROC, rather than the AUROC.

When building our classifiers we labelled the training data as positive if the patient eventually tested positive for toxigenic \textit{C. diff} and negative otherwise, with the goal of identifying high-risk patients as early as possible during the hospital visit. However, the problem could also be framed as a regression or time-to-event task. Predicting \textit{when} a patient will acquire an infection is more difficult than predicting \textit{if} a patient will become infected.

The main contributions of our work are in the form of feature engineering, problem formulation and evaluation methodology. In this dissertation, we focused on developing predictive models for infections with \textit{C. difficile}. However, our contributions extend to building models for other types of healthcare-associated infections (e.g., Methicillin-resistant \textit{Staphylococcus Aureus} and Vancomycin-resistant \textit{Enterococcus}), and other patient outcomes influenced by patient care (e.g., in-hospital mortality). In particular, our work has utility in other problems where the data are high-dimensional and time-varying, where a time-varying model may be more suitable than a static one, or where the data lie in multiple overlapping feature spaces.

One potential application of the insights and techniques presented in this dissertation is the problem of \textit{failure to rescue}. When patients are admitted to a healthcare facility some improve dramatically, while others deteriorate to the point of death or permanent disability. Developing tools in machine learning and data mining, like the ones presented in this dissertation, to improve our understanding of the complex dynamics governing a patient’s health during a hospital admission is an important area of future research.

Ultimately the techniques developed in this dissertation have the potential to effect important changes at the patient level. Widespread development and use of such data-
driven models promises to enable cost-effective, targeted prevention strategies that improve patient care.
Appendix A

Database Details

<table>
<thead>
<tr>
<th>Table</th>
<th>Database Name</th>
<th>Table Name</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>azADT</td>
<td>WHC_PV1601</td>
<td>Admission details</td>
</tr>
<tr>
<td>2</td>
<td>azADT</td>
<td>WHC_PID601</td>
<td>Patient demographics</td>
</tr>
<tr>
<td>3</td>
<td>azADT</td>
<td>AEID201</td>
<td>Mapping of EID to OID</td>
</tr>
<tr>
<td>4</td>
<td>azLAB</td>
<td>V_OBX701</td>
<td>Laboratory results across all hospitals</td>
</tr>
<tr>
<td>5</td>
<td>azLAB</td>
<td>WHC_OBX701</td>
<td>Laboratory results at Hospital C</td>
</tr>
<tr>
<td>6</td>
<td>azADT</td>
<td>V_DG1601</td>
<td>Patient visit diagnoses (all hospitals)</td>
</tr>
<tr>
<td>7</td>
<td>azMEDS</td>
<td>V_RXE701</td>
<td>Medication orders (all hospitals)</td>
</tr>
<tr>
<td>8</td>
<td>azMEDS</td>
<td>V_RXC701</td>
<td>Custom med. orders (all hospitals)</td>
</tr>
<tr>
<td>9</td>
<td>azUSERMODULE</td>
<td>HIS611</td>
<td>Confirmed home medications</td>
</tr>
<tr>
<td>10</td>
<td>azADT</td>
<td>WHC_LOC601</td>
<td>Patient locations within Hospital C</td>
</tr>
<tr>
<td>11</td>
<td>azCLINDOC</td>
<td>WHC_OBX703</td>
<td>Patient vitals at Hospital C</td>
</tr>
<tr>
<td>12</td>
<td>azADT</td>
<td>WHC_PR1601</td>
<td>Procedures at Hospital C</td>
</tr>
</tbody>
</table>

Table A.1: EMR data are stored in tables across several databases. Here we list the relevant tables for extracting data for Hospital C.

In the paragraphs that follow we describe in more detail how patient data are represented in V2 of Amalga (one of the EMR systems currently used by the hospital network). Each database contains institution-specific tables, in addition to views that a user can query to retrieve information across all institutions. Here we focus on the EMR data for Hospital C since it is the largest hospital. The data for the other two hospitals are stored similarly.

**Admission Details** Admission details for patients admitted to Hospital C are found in Table 1. This table contains information regarding patient admissions e.g.,
date and time of admission, date and time of discharge, and type of visit. In this work we consider only healthcare-associated infections, therefore we focus on *inpatient* visits. Table 1 also contains other information pertaining to the admission such as the financial class code, the source of the admission, the hospital service, and the attending doctor. We extract all this information for each visit included in our analysis.

**Patient Demographics** Information pertaining to patient demographics is found in Table 2. Each entry (i.e., row) in this table corresponds to a different OID, rather than an EID. Therefore, we retrieve patient demographic information through a mapping of EID to OID (via Table 3). This table contains information about each patient such as date of birth, gender, race, marital status, and city (or neighborhood) of residence. Aside from the date of birth, all data in this table is categorical. Patient age at the time of admission is calculated using the admission date and the date of birth. In Section 2.5 we describe in more detail how we map this information to a feature vector representing each day of a patient admission.

**Laboratory Results** We consider both Table 4 and Table 5 when processing patient laboratory information. The former consists of lab results across all hospitals in the network, while the latter consists of only lab results for Hospital C. Having lab results across all institutions is useful in determining whether or not a patient has a history of *C. difficile*. Later in this chapter we give more detail describing how positive cases are identified using the laboratory data. In addition to tests for *C. difficile* we have data pertaining to all ordered/observed laboratory tests. Laboratory test are associated with unique identifiers. The database contained 2,237 unique laboratory tests (since 2011). Some of these tests appear in the database only a single time, however a significant number of the tests occur on a regular basis (e.g., 911 tests had a frequency of at least 100, since 2011).

Figure A-1 gives an example of a row in the laboratory table. Each row in the
Each laboratory test is also associated with a reference range (e.g., 120-200 for cholesterol). If the observed value lies outside the normal range for that measurement, an abnormal flag is entered. Abnormal flags are either H=high, L=low, C=critical, or empty=normal. These flags are coded based on the reference ranges (defined by experts) in the health record system. When extracting laboratory results for patients we extract the observation identifier and the flag associated with the observation, as shown in Figure A-1. Measurements pertaining to vital signs are encoded in a similar way.

**Row in database**

<table>
<thead>
<tr>
<th>ObservationIdentifier</th>
<th>ObservationName</th>
<th>ObservationValue</th>
<th>ReferenceRange</th>
<th>AbnormalFlags</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREAT</td>
<td>Creatinine</td>
<td>2.03</td>
<td>0.52-1.04</td>
<td>H</td>
</tr>
</tbody>
</table>

Figure A-1: Laboratory results are represented by the observation identifier and the flag associated with the result.

**Diagnoses** Diagnoses are extracted from Table 6. Diagnoses for a patient are encoded by ICD-9 codes. ICD-9 is a coding system developed by the International Statistical Classification of Diseases and Related Health Problems that encodes diseases (and procedures, as discussed later) hierarchically (see Figure A-2). At the highest level ICD-9 codes fall into 1 of 19 categories [26]. Each row of Table 6 has an EID entry and an ICD-9 code. Since patient visits can be associated with multiple ICD-9 codes, there may be multiple rows in Table 6 corresponding to the same visit. In our data the average visit (including outpatients) is associated with two distinct ICD-9 codes. ICD-9 codes, widely used for billing purposes, can get coded well after a patient is discharged [27]. For this reason, we do not use the codes associated with a patient’s current visit in our model. Instead, we consider only the codes from a patient’s most recent previous hospital visit. We refer to Table 6 since it includes diagnoses codes from all visits.
to any hospital within the network. In Section 2.5, we further describe how we preprocess ICD-9 codes.

001-139 Infectious and Parasitic Diseases
  008 Intestinal Infectious Diseases
    008.4 Intestinal infectious due to other organisms
      008.45 Intestinal infection due to *Clostridium difficile*

Figure A-2: Patient diagnoses are encoded using the hierarchical international disease classification system, ICD-9 codes.

**Medications** We consider medication records from three different Table 7, Table 8 and Table 9. Table 7 contains the medication orders from all hospitals. Table 8 contains information regarding custom combination orders (e.g., a beta-blocker with a diuretic). By joining Table 8 with Table 7 we can retrieve all of the medications ordered for a patient during a visit. Table 9 contains information regarding medications a patient takes at home. In this table, information regarding home medication is recorded for only Hospital C.

Table 7 contains the majority of the information regarding medication orders. An example entry (with the relevant fields) for this table is shown in Figure A-3. As Figure A-3 shows each row is an entry associated a visit (note: the EID is not shown here for privacy reasons), an 8-digit medication ID, and a start time and stop time. Each medication ID is associated with a medication, a dosage and a form as indicated by the “GiveCodeText” field in Figure A-3. For example, in the table *acetaminophen* is represented using three different medication IDs, depending on the dosage and the form. Since the dosage and form are encoded in the medication ID, we represent patient medications using only the ID (and the start/stop times).

We note that in early 2013 the coding system for medications changed to a “PYXIS” coding, from an “AHFS” coding. This did not affect the methods we used, nor did it effect our representation of the medication. However, it is important to be aware of such changes, since if one learns a model based on
Figure A-3: An excerpt from the medication table showing patient medication orders. Each entry corresponds to a new medication order, where the ordered medication is described by a medication ID.

To this end we briefly explored the use of topic modeling to automatically learn classes of drugs based on the drug names alone [28]. Such an abstraction could prove useful if the amount of training data is limited, or if a coding system is not used.

Locations In the database the hospital is divided into units and rooms. A unit can contain multiple rooms, and each room can contain 1-2 beds. The size of the units vary; some units contain no beds (e.g., operating rooms), whereas others may contain up to 30 beds (e.g., a patient care unit). For each hospital admission we have timestamped location data. Location data refers to the patient’s location within the hospital. It is collected at both the unit and the room level. Table 10 contains entries associated with each patient admission. Each entry indicates room, unit and two timestamps indicating when the patient entered and left that room. Table A.2 shows how we can trace a patient’s path through the hospital using the timestamped location data (note: in this table the dates and times were changed for de-identification purposes). From this we can infer when patients were co-located within a unit in the hospital, or who was in the room prior to that patient. In our data each location is represented by a separate binary variable. We incorporate this knowledge about a patient’s location...
into the model, but more importantly we use this information to extract information regarding exposure to *C. difficile*. This is further discussed in Section 2.4.

<table>
<thead>
<tr>
<th>Location Unit, Room</th>
<th>Time In</th>
<th>Time Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiac Cath Unit, 4Axx-P</td>
<td>8/15/03 13:19</td>
<td>8/15/03 22:40</td>
</tr>
<tr>
<td>Medicine Patient CU, 4NxxE</td>
<td>8/15/03 22:40</td>
<td>8/17/03 10:10</td>
</tr>
<tr>
<td>Main OR, MRxx-P</td>
<td>8/17/03 10:10</td>
<td>8/17/03 12:15</td>
</tr>
<tr>
<td>Cardiac Intensive CU, CRxx-P</td>
<td>8/17/03 12:15</td>
<td>8/18/03 10:56</td>
</tr>
<tr>
<td>Surgical Patient CU, 4Fxx-B</td>
<td>8/18/03 10:56</td>
<td>8/24/03 15:37</td>
</tr>
<tr>
<td>Surgical Patient CU, 4Fxx-A</td>
<td>8/24/03 15:37</td>
<td>8/28/03 9:14</td>
</tr>
</tbody>
</table>

Table A.2: Mapping out a Patient’s Visit

**Vitals** We extract vital signs (e.g., respiratory rate and heart rate) from Table 11. In the EMR, vital signs are encoded similarly to laboratory results. Each entry in the vitals table corresponds to a visit (EID), an observation identifier (e.g., BPSYSTOLIC), an observation value, a reference range, an abnormal flag, and an observation date time. When extracting information about vitals for a patient we encode the observations the same way we encode laboratory results, i.e., as a concatenation of the observation identifier and an abnormal flag (e.g., BPSYSTOLIC.H).

**Procedures** At Hospital C, procedures are recorded in Table 12. Procedures are encoded using both the 5-digit *Current Procedural Terminology* (CPT) code and a ICD-9 procedure codes. Each row (i.e., entry) in the procedures table records a procedure, the corresponding visit and a procedure date and time. For some patients only one of the coding systems is used, for others both coding systems are used redundantly (it is not clear why this is done). Since both coding systems are used to describe procedures, in our analysis we consider both CPT and ICD-9 codes (however, we do not learn a mapping between them). We discuss how we represent procedure data in Section 2.5.
Appendix B

Testing Protocol for *C. difficile*

<table>
<thead>
<tr>
<th>Test</th>
<th>Toxin</th>
<th>Antigen</th>
<th>Start Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>C diff Quik Chek Complete</td>
<td>A &amp; B</td>
<td>Glutamate dehydrogenase</td>
<td>6/18/2013</td>
</tr>
<tr>
<td>Illumigene <em>C. diff</em></td>
<td>A &amp; B DNA Amplified Assay</td>
<td>Glutamate dehydrogenase</td>
<td>3/18/2011</td>
</tr>
<tr>
<td>C diff Chek 60</td>
<td></td>
<td></td>
<td>12/1/2009</td>
</tr>
<tr>
<td>C diff Toxin A/B Quik Chek</td>
<td>A &amp; B Antibody</td>
<td></td>
<td>12/1/2009</td>
</tr>
<tr>
<td>C diff Premiere Toxin A &amp; B</td>
<td>A &amp; B Qualitative enzyme immunoassay</td>
<td></td>
<td>8/25/2008</td>
</tr>
</tbody>
</table>

Table B.1: *Clostridium difficile* testing history

At the three hospitals stool tests for *C. difficile* are only performed on patients showing symptoms (e.g., unformed stools). At these hospitals the testing protocol for *C. difficile* has changed over the years. These changes are outlined in Table B.1. In 2008, Hospital C employed an enzyme immunoassay to detect the presence of *C. difficile* toxins A & B in the stool. Later, this test was replaced with a two-step testing algorithm including an antigen test for glutamate dehydrogenase (GDH). GDH is an enzyme produced by *C. difficile* organisms. Assays for GDH are very sensitive and thus can be used to accurately rule out the presence of *C. difficile* in stool samples, however other bacteria can also produce GDH. Therefore, from 12/1/2009 to 3/18/2011 all positive antigen tests were followed up with a toxin test to confirm the presence of toxigenic *C. difficile*. In 2011, the hospital switched to a one step DNA amplification assay for the direct detection of *C. difficile* toxins A & B in the stool. Most recently, in 2013, this was replaced with a test for the simultaneous detection of GDH antigen and toxins A and B. The current algorithm for identifying positive cases for toxigenic *C. difficile* is described in Figure B-1.
Figure B-1: *C. difficile* diagnosis algorithm currently employed at Hospital C (and other hospitals within the same network).
References


A Linsky, K Gupta, E Lawler, J Fonda, and JA Hermos. Proton pump inhibitors and risk for recurrent Clostridium difficile infection. Archives of Internal Medicine, 170(9):772-778, 2010.


142


