Improving Clinical Decisions Using Correspondences Within and Across Electronic Health Records

by

Jen Jian Gong

A.B., Harvard University (2012)
S.M., Massachusetts Institute of Technology (2014)

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

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

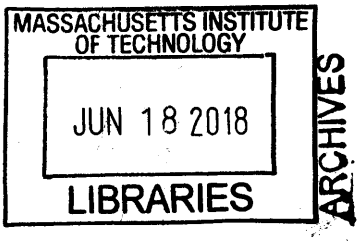
Author ....
Department of Electrical Engineering and Computer Science
May 23, 2018

Signature redacted

Certified by .....
John V. Guttag
Dugald C. Jackson Professor of
Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science Chair,
Department Committee on Graduate Students



Improving Clinical Decisions Using Correspondences Within and Across Electronic Health Records

by

Jen Jian Gong

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Electronic Health Record (EHR) adoption and retrospective analyses of health care data are part of a broader conversation about health care quality and cost in the United States. Machine learning in health care can be used to develop clinical decision-making aids and assess quality of care. This can help improve quality of care while lowering cost. In this thesis, we present three methods of using different kinds of data in health care records to aid clinicians in making care decisions. We focus on the critical care environment, where patient state can rapidly change, and many care decisions need to be made in short periods of time.

First, we introduce a method to use correspondences between structured fields from two different EHR systems to a shared space of *clinical concepts* encoded in an existing domain ontology. We use these correspondences to enable the transfer of machine learning models across different or evolving EHR systems. Second, we introduce a method to learn correspondences between structured health record data and topic distributions of clinical notes written by care team members. Finally, we present a method to characterize care processes by learning correspondences between *observations* of patient state and *actions* taken by care team members.

Thesis Supervisor: John V. Guttag
Title: Dugald C. Jackson Professor of
Electrical Engineering and Computer Science

Acknowledgments

This Ph.D. experience would not have been what it has been without the supportive mentors, colleagues, friends, and family that I have had.

I want to thank Professor John Guttag, who has been an incredible mentor and advisor. I am grateful to him for the feeling of community in our group that he fosters, his enduring optimism and critical insights in research conversations, and the example he sets as a mentor and advisor. I am also grateful for his ability to find (almost!) every run-on sentence and misuse of modal verbs in my writing.

The other members of my committee, Professor Jenna Wiens and Professor Collin Stultz, have also been supportive mentors over the years. I want to thank Jenna for her mentorship and research advice, both as a senior graduate student in John's lab when I was first starting the Ph.D. program, and as a member of my thesis committee. Her feedback has been invaluable to the work in this thesis. I'd also like to thank Collin, who has always pushed for precise hypotheses and problem statements, for clear clinical use cases in our research, and for the plurality of "data."

I cannot thank the members of the Guttag lab enough for their support and guidance over the years. Jenna, Garthee, Anima, Yun, Adrian, Guha, Joel, Amy, Davis, Maggie, Harini, Jose, Katie, Matt, Akhil and Divya have all been incredibly supportive. I'm so grateful for the many spontaneous conversations (research and otherwise), and for the support during late-night (and, too often, early-morning) paper deadlines.

On a similar note, the projects I've worked on during my Ph.D. have been made more rewarding because of the people I've worked with. These include Tristan Naumann (Ph.D.-twin!) and Professor Peter Szolovits at MIT, and Professors Jordan Green and Tiffany Hogan at the MGH Institute of Health Professions. I've also had a great time working with two amazing undergraduate and Master's students over the years, Dina Levy-Lambert and Maryann Gong.

I'd also like to thank the friends I've made at MIT. In addition to those listed above, I'd like to thank Marzyeh Ghassemi, Frank Wang, and Petra Lindovska for

their friendship and encouragement through all of the ups and downs in the last 6 years.

Finally, I'd like to thank my family. My parents, Weibo Gong and Ping Yuan, and my brother, Steven, have always been there for me. I can't thank them enough for the sacrifices they have made in order for me to be where I am today, and for their support and encouragement through the years. Last, but certainly not least, I want to thank my partner, Daniel Reichert. Despite the thousands of miles between us in the last four years, you have always been there for me. Your support, love, and thoughtfulness mean the world to me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Health care quality and spending in the United States are urgent national issues. In 2016, U.S. health care spending accounted for 17.9% of the GDP [1]. In addition, preventable medical errors in inpatient settings are estimated to account for over 250,000 deaths per year in the U.S [2], and have been estimated to account for $17.1 billion of health care costs annually [3]. To improve care and lower costs, clinical decision-making aids can be used to help manage care, and quality assessment tools can be used to evaluate and regulate provided care.

Increasing volumes of healthcare data enable researchers to better understand individual health (e.g., health conditions, disease progression, risk factors), and the effects of patient interactions with the health care system. Since the introduction of the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009, data storage in electronic health records (EHRs) in the U.S. has exploded [4]. A 2016 report shows that over 95% of hospitals eligible for the Medicare and Medicaid EHR Incentive Program have adopted EHR systems that have achieved "meaningful use" [5]. These record systems are used during the course of care for structured information entry and retrieval, and communication between care team members about patient status in clinical notes.

The availability of such data enables secondary analyses that can lead to actionable metrics. These metrics are important for 1) clinical decision-making aids that help clinicians make better informed decisions by summarizing vast amounts of data, and

2) quality of care assessment tools for improved transparency and accountability. An important step to achieving these goals is disentangling systematic *care* factors from patient-specific *health* factors. Machine learning approaches can be used to derive actionable, data-driven insights in both of these areas.

Machine learning has successfully been applied to health care data to predict adverse events in patients, including mortality, hospital-acquired infections, and intervention administration. These models take large amounts of data collected during the course of care, and identify patterns in the data that are predictive of relevant patient outcomes. However, health care data come from heterogeneous patient populations and care processes, contain heterogeneous data types, and are often not missing at random. Thus, it can be challenging to directly apply out-of-the-box machine learning methods to health care data.

Clinical data exhibit heterogeneity in *patient populations* and *care processes*. For example, a patient who enters the hospital with renal failure has a different physiology from a patient entering the hospital for elective cardiac surgery. These patients will have different factors that are predictive of adverse outcomes such as mortality. This heterogeneity makes it difficult to build machine learning models that are generalizable across patient populations. In addition, the process of care for these patients will differ; a patient with renal failure will undergo different procedures and receive different medications compared to a patient receiving cardiac surgery. Thus, different sets of treatment decisions are available for different patient populations. In addition, care processes also differ across institutions.

Clinical data contain *heterogeneous data types*. EHRs contain categorical data elements such as medications ordered, procedures performed, lab tests performed, and diagnosis codes. Continuous values, such as test results, are also stored in EHRs. In critical care settings, continuous vital signs might also be regularly monitored and stored. In addition, EHRs contain free-text clinical narratives that summarize history, symptoms, and the course of care for a particular patient. Integrating discrete structured items, continuous time-series, and text in a machine learning model is not straightforward.

Another issue is the incompleteness of clinical data. Clinical data are often not missing at random. Data availability is not consistent across data types, patient populations, and care processes. For example, vital signs monitoring might be routine in the intensive care environment, but not on the hospital floor. Clinical notes are often recorded at routine intervals (e.g., at the beginning of the day, or during rounds), but are otherwise not updated. These data are *systematically* missing, rather than missing at random.

Additionally, secondary analysis of EHRs necessitates understanding the conditions under which the data were originally collected. While EHRs are now used by the majority of physician offices and hospitals, their utility and structure are still debated [6, 7, 8]. In addition, the usage (and capability) of EHRs is financially incentivized based on guidelines set first by the Meaningful Use EHR incentive program in 2009, and more recently by the Advancing Care Information component of the Merit-based Incentive Payment System (MIPS) [9, 10]. As regulations and financial incentives shift, the type, format, and comprehensiveness of the data will change. EHR data are collected for care, billing, and accountability, rather than for large-scale retrospective analyses. Data in EHRs are therefore not only biased in data presence based on heterogeneous patient populations and care practices; they are also biased based on current regulations and intended use.

Another important consideration is that EHRs are constantly evolving. They are not standard across hospitals and care settings, and also evolve in the same hospital over time. Thus, even if information can quickly be retrieved within a single EHR, records from different institutions or from different time periods may not be interoperable. This greatly limits the ability of clinicians to communicate across institutions [11], and limits how well clinical decision-making aids generalize across data encoding systems.

In this thesis, we address many of these challenges to utilizing machine learning on real EHRs for actionable clinical decision-making tools. Our work seeks to contextualize the utility of accurate clinical decision-making aids with respect to the underlying data collection process. Our goal is to gain an understanding of how data character-

istics and relationships might affect down-stream predictive models. We build risk models for adverse outcomes to demonstrate how these insights about EHR data can be used to improve the generalizability of clinical decision-making tools across different EHR systems, care settings, and patient cohorts. In the following section, we briefly describe the problem statements of the works in this thesis.

1.1 Overview

1.1.1 Enabling transfer of machine learning models across EHR systems.

EHR data interoperability is an important goal in facilitating communication about patients across different care facilities. But, it also has implications for learning generalizable clinical decision-making aids and quality assessment tools using machine learning across institutions and over time. Transferring machine learning models across EHR systems is particularly challenging for the structured data, which may have similar semantic meanings but completely different encodings.

Machine learning can provide useful insights into patient state and provide clinicians with useful, actionable information. However, learning accurate models can require large amounts of data. Each change in variable encoding from one EHR system to another means that a model developed on one system relies on information that may not be available in another. Risk models that rely on small numbers of variables can be manually mapped, but this is infeasible for the thousands of items that exist in modern EHRs.

Structured data encoding systems often come with text descriptions of the content in each field. While the structured encodings themselves may not be transferable, the clinical *concepts* contained in each field can be extracted from these text descriptions. We present a method to leverage these text descriptions of the structured data items to enable model transfer across EHR versions. We present a case study of our approach on a transition from one EHR system to another at a single institution.

Using an existing domain ontology of medical concepts, we translate EHR-specific item encodings to a shared semantic space. We demonstrate that models learned in this shared semantic space have significantly improved performance when applied to a new EHR version, compared to models learned in a version-specific encoding space.

1.1.2 Summarizing structured health record data for reducing information overload and improving communication across care teams.

Information overload for health care providers is a well-documented problem [12], and can result in errors in care. Structured health record data can capture thousands of variables during the course of the stay, and clinicians cannot be expected to look at all of them. Effective methods of summarizing health record data to identify the most salient information about a patient can help alleviate this problem. Clinical notes serve this purpose in the course of care. Care staff summarize important aspects of patient history, state, and treatment activity in the clinical notes. These clinical notes serve as important documents in the transition of patient care from one team to the next.

Clinical notes are written intermittently during the care process. In contrast, structured health record data (e.g., lab test results, vital signs monitoring, charted observations, medications, treatments, etc.) are recorded more frequently. We present a learning-based approach to generating potential topics that should appear in summaries of patient state and care at any given time. We learn cross-modality relationships between structured health record data and topic distributions of existing clinical note summaries written by care team members. By using existing summaries as a supervised target, our model is able to learn how to summarize high-dimensional structured health record information into latent *topics*, a text-based feature representation.

Clinical notes in electronic systems are often copy-pasted forward [13]. Because of this, notes frequently contain redundant and/or outdated information. In addition,

notes may be missing important information. This work enables generating topics at any point that would be pertinent in summarizing the care process and patient state. Suggested topics can help clinicians recognize when relevant information about the patient might be missing, and when earlier information might no longer be relevant. In addition, this work is a first step towards generating an actual note from the structured data.

1.1.3 Characterizing care processes

Machine learning for clinical decision-making aids has focused primarily on patient risk assessment tools and patient subtype characterization. In these scenarios, it is important to handle characteristics of the data (e.g., missing values) that are a result of the *process* of care. For example, a test for bilirubin might not be performed if the care provider is not concerned with the healthiness of a patient's liver.

Care processes can be challenging to characterize. In this work, we seek to build an understanding of processes by learning how *observations* about patient state correspond to *actions* taken by care team members. We use a learning-based approach to learn correspondences between observations of patient state and actions that are taken by care team members.

We focus on three categories of actions: 1) laboratory tests, 2) imaging tests, and 3) antibiotic administration. These categories capture both actions that are routinely done during the course of care (e.g., a lab test panel at the beginning of a patient stay), as well as indicators of care process driven by patient observations (e.g., a lab test that is ordered further along in the patient stay because the physician is interested in a particular value).

We compare actions predicted using our model, which takes in observations of patient state, to actions predicted using incidence of actions over time. We show that incidence alone can capture interesting characteristics in care patterns (particularly routine care actions).

1.2 Contributions

The contributions of this thesis are as follows:

1. **We present and evaluate a novel approach to reconcile structured EHR data elements across two EHR systems.** We show how text meta-descriptions of structured EHR data elements can be leveraged to map items that are encoded differently in two EHR systems to a domain-specific ontology. By mapping distinctly encoded items from two EHR systems to the same vocabulary, we enable clinical risk model transfer across the systems. We demonstrate that changing the feature space in this way can significantly improve model generalizability across systems.

2. **We present and evaluate a method that uses structured health record data to predict the topic distributions of clinical notes.** Structured health record data and clinical narratives are very different data modalities. We show that the structured health record data is able to predict the topics in the next clinical note comparably well to using prior notes. We also demonstrate that the structured data can accurately predict the first clinical note in a stay. Finally, we demonstrate that our learned correspondences capture meaningful aspects of patient state using downstream outcome prediction tasks.

3. **We present and evaluate a method that learns correspondences between observations of patient state and care actions.** This is a first step towards characterizing care processes based on observations of patient state using a learning-based method. We compare our model to using baseline incidence in the population at different times during the patient stay. We demonstrate that we are able to capture meaningful correspondences between observations of patient state and provider decisions to administer antibiotics, perform a lab test, or perform an imaging test.

All of the work we present is evaluated on the MIMIC-III dataset, which contains data from a critical care setting [14]. Although our experiments are limited to the

MIMIC-III dataset and therefore to patient populations in a critical care setting, the questions pertaining to diversity of underlying patient conditions, care patterns, and relationships between different modalities of data are generalizable to other health care data and other applications of machine learning.

1.3 Outline

In Chapter 2, we describe the MIMIC-III data set and important attributes of the different data modalities we considered. In Chapter 3, we describe our work on enabling risk model transfer across distinct EHR systems using a shared vocabulary from a domain-specific ontology. In Chapter 4, we describe our work on learning correspondences across different modalities of clinical data. In Chapter 5, we describe our method for characterizing care processes. Related work is presented in each chapter. Finally, in Chapter 6, we summarize our contributions and expand on the implications of these works for the development and application of machine learning models in real-world clinical settings.

Chapter 2

Data

In all of the work described in this thesis, we used the MIMIC-III dataset, an openly accessible critical care dataset [14]. MIMIC-III contains data from the Beth Israel Deaconess Medical Center, an academic hospital in Boston, collected over the years 2001-2012. It provides detailed static patient information, such as demographics upon admission, as well as temporally-varying data, such as regularly sampled vital signs, irregularly sampled lab test results, time-stamped treatments and interventions, and periodic clinical notes.

In this chapter, we first give some background on critical care environments and how they are distinct from other health care settings. Next, we describe the different types of data available in MIMIC. Finally, we give an overview of the data used in each of the subsequent chapters.

2.1 Critical Care Environments

Critical care is a distinct environment from other care settings—patients in the intensive care unit (ICU) tend to be severely ill and have complex combinations of chronic and acute conditions. Despite a decreasing number of hospital beds in the United States, the number of beds in critical care units increased between 2000 and 2010 [15].

Critical care environments are also distinct from other care settings in the volume of care decisions that need to be made in short periods of time. This time urgency

makes tools that accurately stratify patients into risk categories and help clinicians make decisions even more important than in other care settings. In addition, care in the ICU is provided by a team of nurses, residents, and attending physicians that must coordinate actions and observations in the care process for each patient [16]. Finally, the critical care environment is equipped with monitoring systems that capture vital sign measurements and other values that are not routinely available in other hospital services or in outpatient care.

In summary, the data from ICUs capture complex patient conditions and physiological characteristics, decisions and actions taken by a care team, and modalities of data that may not be available in other care settings.

2.2 MIMIC-III

MIMIC-III contains data sourced from the hospital database, as well as data from the ICU databases, over the years 2001-2012. During this period, the hospital database system stayed the same, but the EHR version used in the ICU changed from CareVue (2001-2008) to MetaVision (2008-2012). Thus, tables in MIMIC that are derived from the hospital database (e.g., lab tests) are shared across all admissions, but tables that are specific to the intensive care unit (e.g., charted events) contain distinct item encodings for semantically similar items.

Table 2.1 details the tables in MIMIC that are specific to each EHR version, as well as the ones that are shared.

2.3 Data Modalities

In the following sections, we describe each of the modalities of data we used—*clinical events* and *physiological time-series*, which are from the *structured* portion of the health record, and *clinical notes*, which are composed of text.

In the care setting, clinical narrative notes facilitate communication and help clinicians summarize and identify the most relevant aspects of the deluge of available

Table 2.1: MIMIC data tables used in this thesis. Tables that were common and distinct across the change in EHR systems (Carevue, 2001-2008 to Metavision, 2008-2012) are shown.

	Table name	Contents
Shared	`labevents`*† ‡	Time-stamped lab test results.
	`microbiologyevents`*†	Microbiology tests. Contains time-stamped and only date-stamped** events.
	`noteevents`†	Clinical notes. Certain categories were specific to the CareVue portion (`nursing/other`), and certain categories were specific to the MetaVision EHR system (`physician`).
	`services`†	Time-stamped changes in service during the hospital admission.
	`prescriptions`*	Date-stamped** prescriptions.
Not Shared	`inputevents_cv`*	Time-stamped input events. Includes nutritional items, medications, and IV fluids. Specific to CareVue.
	`inputevents_mv`*† ‡	Time-stamped input events. Includes nutritional items, medications, and IV fluids. Specific to MetaVision.
	`procedureevents_mv`† ‡	Time–stamped procedures. Specific to MetaVision.
	`outputevents`*† ‡	Time-stamped output events.
	`chartevents`*† ‡	Time-stamped charted events and observations by the care staff.
	`datetimeevents`† ‡	Past events with date, time values.

*Event categories used in Chapter 3.

†Event categories used in Chapter 4.

‡Event categories used in Chapter 5.

**Date-stamped (but not time-stamped) events were only used in Chapter 3.

data about each patient [7, 8, 17]. In contrast, the structured data elements are more difficult to extract information from quickly. However, structured EHR data facilitate clinical studies and data analyses [7, 8, 18]. Thus, what is useful in clinical studies and developing decision-making aids can be at odds with what is useful in the practice of care [17, 18, 19].

Even within the scope of clinical decision-making tools, these data modalities provide different types of information. Structured data, such as the physiological time-series and clinical events, are more often available than clinical narratives, which are updated more intermittently. In addition, structured data elements capture fine-grained observations, whereas clinical notes summarize patient state. Thus, machine learning models that leverage these structured data may present more immediately actionable output compared to models that leverage clinical narratives.

In addition, while care events and observations can more easily be extracted from the structured data than from clinical notes to build clinical decision-making tools, the structured portion of the EHR varies across care settings and hospitals. Thus, tools specific to a given EHR can be difficult to adapt to other institutions. In contrast, clinical notes are consistent across EHR systems, in the sense that they always consist of text. Tools developed for processing natural language within clinical notes can therefore readily be applied to new EHR systems.

Finally, these data modalities are recorded through different lenses. For example, physiological time-series directly measure changes in patient state. In contrast, the clinical events and clinical narratives capture observations and opinions through the lens of care team members.

2.3.1 Events

The clinical events are extracted from the structured data and exclude information contained in the clinical notes. They include medications, procedures, lab tests and results, input/output (IO) fluid events, microbiology tests, and observations noted in the chart. These events capture most of the interactions the patient had with the health care delivery system. Different subsets of events were used in each Chapter.

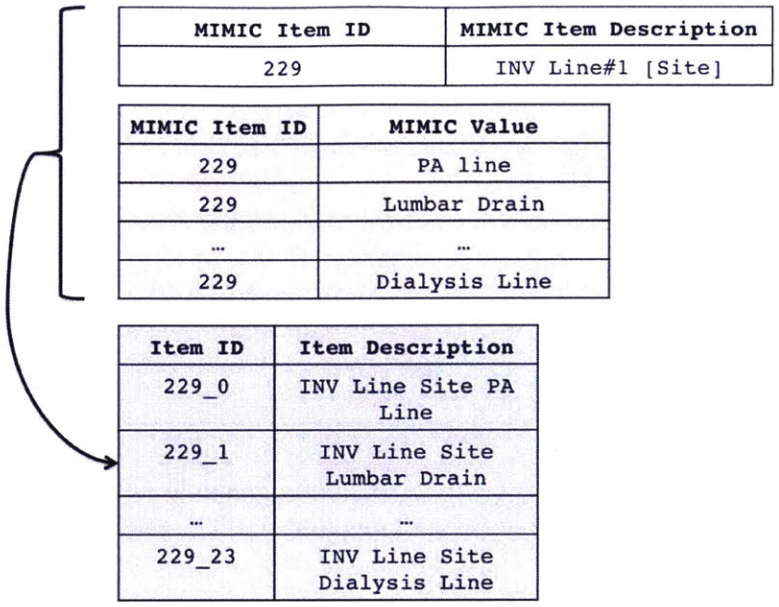


Figure 2-1: Text values often modify the semantic meaning of the corresponding items. We assign new unique item IDs with item descriptions that append these values to the initial item description. In this example, ID 229 in MIMIC is associated with a number of distinct text values in patients' charts that modify its semantic meaning.

These are identified in Table 2.1. Each event is based on a unique numerical identifier from the database. A dictionary is provided with MIMIC containing a short text description for each identifier. For example, the ID 229 in MIMIC-III is associated with the text description "INV Line #1 [Site];" in other words, information about an invasive line that has been placed in the patient.

In each table, events are encoded by the item identifier. Each event is associated with a value. These values are sometimes text-based. We consider distinct (MIMIC event identifier, value) pairs as distinct items when the value is text-based, and assign new unique identifiers to each one. This is because a number of events are semantically modified by the associated text values. For example, the ID 229 is associated with values like "PA Line," indicating a pulmonary arterial line, and "peripherally inserted central catheter." These values modify the original item description in semantically distinct ways, and should be considered separate events. In contrast, while numerical values capture measurements of patient state, they do not alter the meaning of the event itself. This example is depicted in Figure 2-1.

2.3.2 Patient Physiology

In the events data, we excluded numerical values. However, these values are important in ascertaining the state of the patient. We extracted 31 vital signs and lab values from the database for each patient. These time-series capture information about the patient's underlying physiology, and have been shown to be predictive of patient outcomes (e.g., [20, 21, 22, 23, 24]). These time-series differ in the frequency and regularity at which they are sampled. For example, whereas many lab tests are done in the first 24 hours of the stay (routine lab panel), they are sampled intermittently and irregularly (only when needed) afterwards. On the other hand, vital signs are typically regularly monitored through the entire ICU stay, but are not recorded when the patient is on the hospital floor.

The physiological time-series we considered included: diastolic blood pressure, systolic blood pressure, mean blood pressure, heart rate, respiratory rate, temperature, height, weight, white blood cell count, pH, albumin, anion gap, bicarbonate, bilirubin, blood urea nitrogen, chloride, creatinine, fraction inspired oxygen, glucose, hematocrit, hemoglobin, INR, lactate, magnesium, oxygen saturation, partial thromboplastin time, phosphate, platelets, potassium, prothrombin time, and sodium.

2.3.3 Notes

MIMIC contains clinical notes for many patient admissions. Some notes only have date stamps, while others are both date- and time-stamped. Notes come from a number of different genres, including test reports (e.g., radiology reports, ECG reports, etc.), nursing notes, physician notes, and discharge summaries. These notes serve different purposes; test reports contain observations from the performed tests, physician notes and nursing notes contain updates on patient care through the course of the stay, and discharge summaries describe the care process and patient state through the entire hospital stay.

In MIMIC, different note categories are available in the different EHR versions. For example, while "Nursing/other" is a category in the CareVue portion of the dataset,

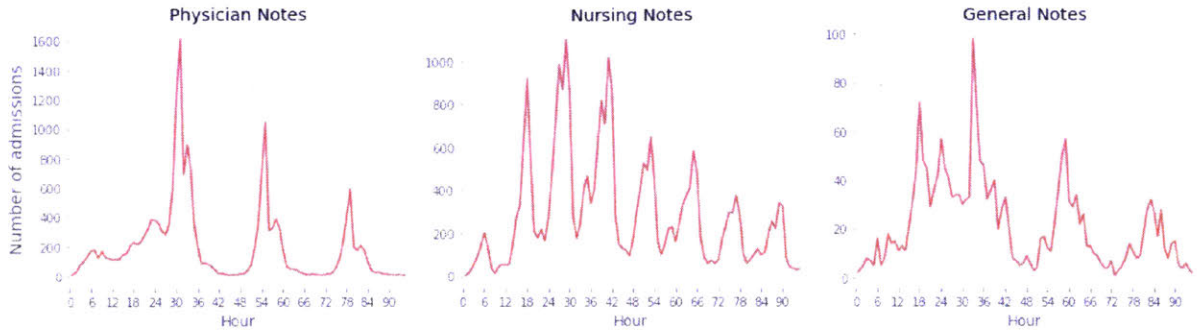


Figure 2-2: Time-stamped physician, nursing, and general notes from the MetaVision portion of MIMIC. Timing of physician notes peaks at 6 a.m. in the morning. Timing of nursing notes is more irregular than physician notes, but exhibits regular inter-event intervals of approximately 6 hours.

it does not exist in MetaVision. Similarly, there are very few physician notes in the CareVue portion of the dataset.

Figure 2-2 shows the number of admissions in the MetaVision data with notes at each hour of the ICU stay, aligned on midnight of the day of ICU admission. The timing of physician notes (Figure 2-2 (left)) has a peak at 6 a.m. every morning. The timing of nursing notes (Figure 2-2, right), also demonstrates regular structure, but more frequently through the course of each day.

2.4 Care Units

MIMIC contains data from 5 distinct critical care units: 1) the Coronary Care Unit (CCU), 2) the Cardiac Surgery Recovery Unit (CSRU), 3) the Medical ICU (MICU), 4) the Surgical ICU (SICU), and 5) the Trauma Surgical ICU (TSICU). While these are all critical care environments, patients in these care units are distinct populations. Figure 2-3 shows the CCS category breakdown of the distribution of primary diagnoses in each unit for the top 20 CCS categories. Certain diagnosis categories appear primarily in a single ICU. For example, most patients with a diagnosis of "fractures" or "acute cerebrovascular disease" were in the TSICU or SICU. All of the patients with a primary diagnosis of "heart valve disorders" were either in the CCU or the CSRU.

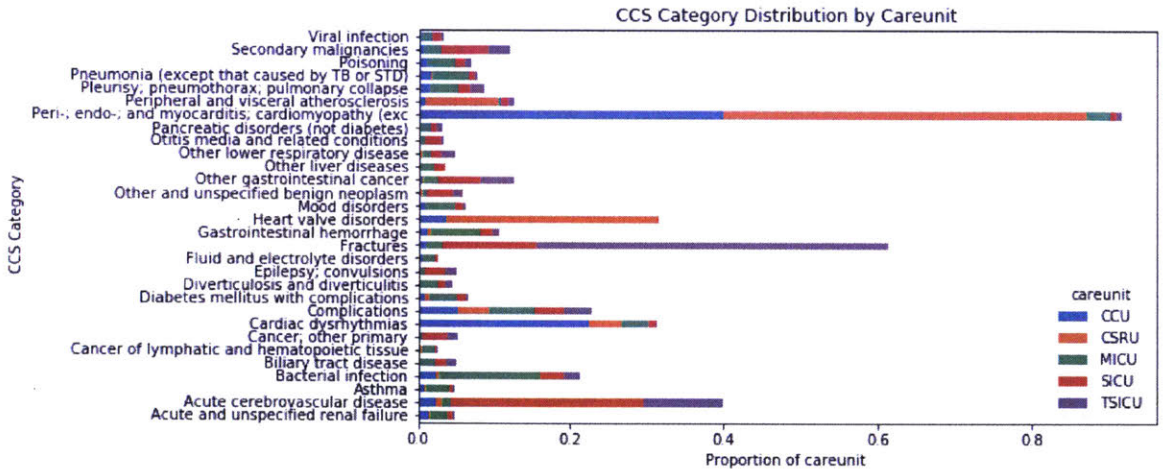


Figure 2-3: Distribution of Clinical Classifications Software (CCS) categories of ICD-9 diagnosis codes across care units.

Thus, the care unit a patient is admitted to can contain important information about underlying disease or condition. The type of care patients undergo in these different care units is distinct; patients in the CSRU are primarily in the intensive care unit to recover from elective cardiac surgery, whereas patients in the MICU may have a diverse set of conditions, from a bacterial infection to gastrointestinal hemorrhage.

2.5 Admission Alignment

Data alignment is a challenging problem, and it is not immediately clear how data from patients in the intensive care unit should be aligned. In MIMIC, many forms of data are not available until ICU admission (e.g., vital signs monitoring). Thus, for models leveraging the physiological time-series as the primary features, it makes sense to align patients on ICU admission. However, care actions are not aligned on time of ICU admission. For example, some care processes (e.g., routine observations during rounds, clinical note entry) are aligned to particular times of the day. Figure 2-2 shows how notes are entered at routine times of day (physician notes around 6-7 a.m. each day, nursing notes at 6 a.m. and 6 p.m. each day, etc.).

Machine learning models for predicting outcomes in the intensive care unit typically align patient admissions on the time of admission. In Chapter 3, we follow this

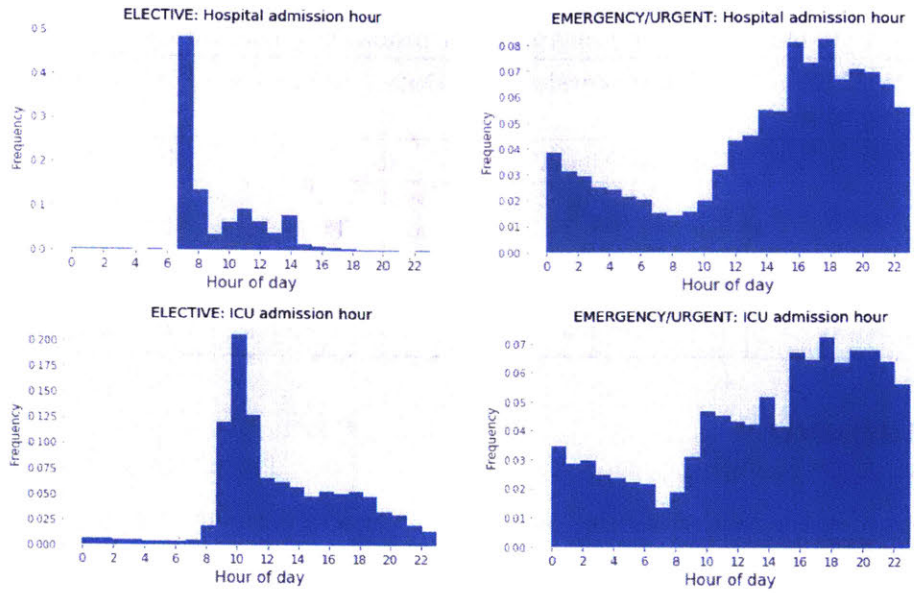


Figure 2-4: Distribution of hour of day of hospital admission (top row) and ICU admission (bottom row) by admission type (elective, left; emergency/urgent, right). Timing of admissions is distinctive across admission type. Elective hospital admissions peak at 7 a.m., and correspond to ICU admissions around 10 a.m. On the other hand, emergency/urgent admissions are spread out throughout the day. A higher proportion of emergency/urgent admissions occur late in the evening compared to elective admissions.

convention and align examples on ICU admission. We consider a feature representation that uses the first 24 hours of the ICU stay. However, for Chapters 4 and 5, we consider time-series representations of patient data during the course of the stay. In these chapters, we align examples on midnight of the day of admission. This method of alignment captures time-of-day characteristics.

The time of day of hospital admission and ICU admission can be an important indicator of patient state. Figure 2-4 illustrates the correlation between hospital admission and ICU admission times with patient state. Patients who are admitted with elective status are mostly admitted to the hospital around 7 a.m. and to the ICU around 10 a.m. Patients who are admitted with emergency/urgent status are more likely than elective admissions to be admitted at early morning hours or later in the evening.

Table 2.2: Database version and data modalities used in each chapter

Chapter	MIMIC Version	EHR Version	Data Modalities	Alignment
3	1.3	CareVue, MetaVision	Events	ICU admission
4	1.4	MetaVision	Events, Physiological Time-Series, Notes	Midnight on day of ICU admission
5	1.4	MetaVision	Events	Midnight on day of ICU admission

2.6 Summary

In this chapter, we provided a brief overview of critical care environments, the structure of data in MIMIC-III, the different data modalities we used in this thesis, and some additional considerations when building machine learning models using health record data. Table 2.2 describes the version of MIMIC and the subsets of data used in each chapter.

Chapter 3

Predicting Clinical Outcomes Across Changing Electronic Health Record Systems

3.1 Introduction

Existing machine learning methods typically assume consistency in how information is encoded. However, the way information is recorded in databases differs across institutions and over time, rendering potentially useful data obsolescent. This problem is particularly apparent in hospitals because of the introduction of new electronic health record (EHR) systems. During a transition in data encoding, there may be too little data available in the new schema to develop effective models, and existing models cannot easily be adapted to the new schema since required elements might be lacking or defined differently.

In this chapter, we explore the effect of data encoding differences on machine learning models developed using EHRs. EHRs are constantly changing, utilizing new variables, definitions, and methods of data entry. In addition, EHR systems across institutions, and even in different departments within the same institution, often differ. While changes can appear minor, each difference means that a risk model

developed on one version may depend on variables that do not exist or are defined differently in another version. For example, the Society for Thoracic Surgeons' Adult Cardiac Surgery Database has undergone many transitions since its introduction in 1989 [25]. During one transition, two variables indicating whether a patient has a history of smoking or whether the patient is a current smoker were remapped to a single variable capturing whether the patient is a current or recent smoker [26].

Remapping variables manually is feasible for small changes, but modern EHRs may contain over 100,000 distinct items, and this number continues to grow over time [27, 28]. Consequently, risk models typically rely on only a small number of variables so that they can be easily adapted. It has been shown, however, that models based on a large number of variables typically out-perform models based on a small number of variables [29]. The alternative, building version-specific models, is prohibitively labor intensive and creates a problem during transitions from one system to another, when there are insufficient data from the new version to build a high-quality risk model.

We enable the application of machine learning models developed using one database on data from another version. We apply natural language processing (NLP) techniques to meta-data associated with structured data elements and map semantically similar elements to a shared feature representation. This approach facilitates building models that can leverage data from another database without restricting the data to a small subset or requiring database integration, a difficult problem [30, 31].

In this chapter, we relate EHR-specific data to clinical concepts from the Unified Medical Language System (UMLS) [32], a collection of medical ontologies. An ontology consists of a set of concepts (*entities*), and *relations* between entities. Although general domain ontologies (e.g., [33]) and tools for identifying equivalent semantic concepts (e.g., [34]) exist, these tools do not work well with the highly domain-specific vocabulary present in clinical text.

We demonstrate that using a shared set of semantic concepts improves portability of risk models across databases compared to using EHR-specific items. We do this by evaluating the performance of clinical risk models trained on one database and tested on another for predicting in-hospital mortality and prolonged length of stay (LOS).

Our work makes the following contributions:

1. We present a novel approach to facilitating the construction and use of predictive models that work across multiple EHR systems.

2. We demonstrate the effectiveness of our approach on two commonly used predictive models and on data from the two epochs of EHR systems in MIMIC-III.

3.2 Related Work

Several solutions to resolving structured data in different EHR versions have been proposed in the literature. Much previous work has developed methods to reconcile health care information with different encodings of variable names by mapping databases to existing clinical vocabularies and ontologies [35, 36, 37].

In [37], the author proposes a method to leverage UMLS to merge two databases. He demonstrates his approach by producing a shared representation for lab items at two different hospitals. This work builds a semantic network for each database structure on its own, and then seeks to merge the two structures by leveraging context and outside sources such as UMLS. In contrast, our work does not seek to relate individual concepts within an EHR as a semantic network. Instead, we map each element directly to concepts in the UMLS ontologies and use this representation for greater generalizability of predictive models.

In the area of clinical risk-stratification, [38] demonstrated that a model for identifying patients with rheumatoid arthritis generalized well at other institutions, despite differences in the natural language processing pipelines used and the differences in structured variable coding across EHR systems. While promising, the logistic regression model they tested used only 21 characteristics (from clinical notes and structured data) drawn from the patient's record. A similar method would not be appropriate for our task, which draws upon thousands of characteristics.

Changing encodings of databases is an opportunity for transfer learning methods, where information from a task that is related (source task) but not directly relevant to

the task of interest (target task) is leveraged to improve performance. For example, [39] transferred information from other hospitals in the same hospital network to improve risk predictions for a hospital-acquired infection at the hospital of interest. In [39], the hospitals had a shared set of features, but also hospital-specific features. Similarly, our EHRs intersect (capturing similarly coded lab tests, microbiology tests, and prescriptions), but each also contains a large set of features that does not appear in the other. Rather than utilizing the EHR-specific features directly in our models, we present an approach to first map the features to semantically equivalent concepts. Unlike most feature-representation transfer methods, which explicitly use the data to learn a feature representation where the source and target data distributions lie closer together [40], we utilize an existing domain-specific vocabulary encoded through expert knowledge.

3.3 Method

In this work, we use the clinical events data described in the previous chapter. We consider this feature space because it relies on the encoding of items in the EHR. Events are represented by the number of times they occurred. Each patient is represented as a bag-of-events (BOE) gathered from the first 24 hours of their stay. The BOE representation omits information about the ordering of events and any associated numerical values (e.g., the result of a blood pressure measurement). This type of BOE representation has been used previously to construct clinical risk models from structured data [41, 42, 43].

We construct our feature representation to demonstrate that mapping to a shared encoding enables building effective risk models *across* EHR versions. The goal of using this representation is not to learn the best possible risk models; instead, it is to elucidate the impact of transferring models from one database to another. While using the values of lab tests or vital signs would certainly lead to improved predictive performance [44, 45, 46], it would obscure information about how the *encodings* affect model performance.

Bag-of-events is analogous to the bag-of-words representation for text. We therefore apply the common normalization technique *term-frequency, inverse-document frequency* (tf-idf). Tf-idf favors terms—or, in our case, events—that occur with high frequency within an individual but infrequently across individuals. These weights tend to filter out features that occur so broadly that they are ineffective in differentiating individuals. Finally, we apply a maximum absolute value normalizer to all features after tf-idf transformation to make the ranges of tf-idf transformed features comparable.

The events we consider are represented in 1) EHR-specific domains, and 2) UMLS concept unique identifiers (*CUIs*). These feature spaces are presented in the following sections. After constructing the BOE representation in the Item ID feature space, we apply a filter to remove events that occurred in fewer than 5 patients to alleviate sparsity in the high-dimensional feature space (15,909 items in CareVue, 5,190 events in MetaVision). After applying the filter, CareVue had 5,875 features and MetaVision had 2,438 features.

3.3.1 Mapping EHR Item ID to UMLS Concept Unique Identifiers

In order to identify the shared semantic concepts represented by the EHR-specific Item IDs, we annotate clinical concepts from the UMLS ontologies in the human-readable item descriptions. Although concepts could be identified using simpler string matching methods such as edit distance, these methods do not handle acronyms and

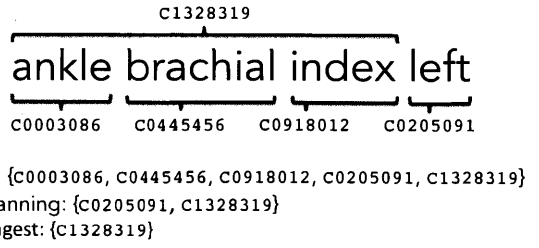


Figure 3-1: **All**, **Spanning**, and **Longest** methods for annotating "ankle brachial index left." These approaches relate the item descriptions to different sets of CUIs.

abbreviations (common in clinical text) well.

Using the Clinical Text Analysis Knowledge Extraction System (cTAKES), a frequently used tool for identifying UMLS concepts, we annotate the human-readable item descriptions from both EHR versions in our data [47]. cTAKES was primarily developed for annotating clinical notes, which contain more context than the EHR item descriptions. This makes identified entities in the item descriptions difficult to disambiguate, and cTAKES often identifies many concepts for each item description. The entity resolution process is further complicated by the differing methods of EHR event entry between CareVue and MetaVision. CareVue allowed for free-text entry of item descriptions, resulting in typos and inconsistent abbreviation and acronym usage. These characteristics result in less context to leverage during the entity resolution process, and lead to some ambiguous annotations. Thus, the relation of Item IDs to CUIs often identifies several relevant concepts, rather than a single one. In contrast, MetaVision had fewer free-text item descriptions, and more consistent text values.

To address this, we consider three methods for defining the set of CUIs corresponding to each item ID: 1) all CUIs found (*all*), 2) only the longest spanning matches (*spanning*) and 3) only the longest match (*longest*). The *spanning* method is also utilized by [48]. The authors suggest that this method identifies the most specific concepts corresponding to a given segment of text, without eliminating useful text auxiliary to the longest concept mention.

Consider, for example, the text "ankle brachial index left" (Figure 3-1). Initially, five CUIs are associated with this text. For this example, *longest* would choose only the CUI for "ankle brachial index," and ignore "left." This is because "ankle brachial index" corresponds to the CUI with the longest span in the text description. This method will likely drop informative CUIs. This is evidenced by the large drop in the average number of CUIs identified compared to *all* (see Figure 3-2). On the other hand, *all* does not remove any CUIs. This may capture concepts that are only marginally relevant to the item description. For example, the *all* annotation of "ankle brachial index" identifies "ankle," "brachial," and "index" as separate CUIs, in

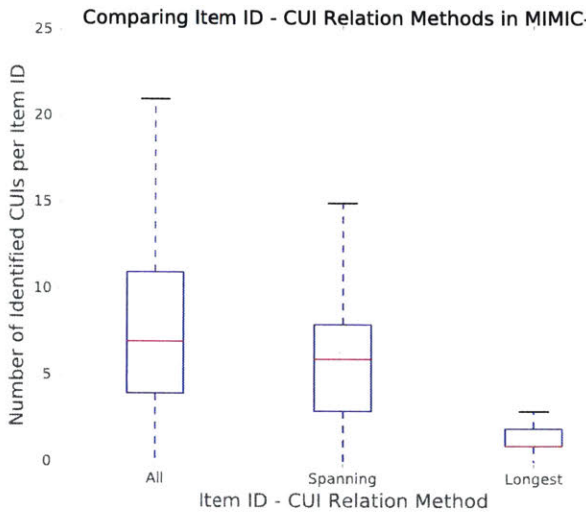


Figure 3-2: Distribution of number of identified CUIs per Item ID: Comparing *All*, *Spanning*, and *Longest* relation methods.

addition to the full concept of "ankle brachial index." Capturing these constituent words—"ankle," "brachial," and "index"—as relevant to the concept of "ankle brachial index" could be misleading rather than informative. Finally, *spanning* presents a medium between *longest* and *all*. For this example, it would identify "ankle brachial index" and "left" as the corresponding CUIs. This captures all of the concepts with the longest spans across the text without dropping text or including concepts with mentions contained within a longer, more specific mention.

Figure 3-2 shows the distribution of number of CUIs per Item ID for the different mapping methods. *Spanning* maintains approximately the same mean number of CUIs per Item ID compared to *all*, while reducing the tail from over 20 to 15 CUIs. In Section 3.5.2, we evaluate these different methods for mapping Item IDs to CUIs.

With the resulting set of CUIs corresponding to each Item ID, we mapped the Item ID BOE feature vectors to CUI feature vectors. For each CUI, we found the set of Item IDs that contained that concept. We then summed the counts from that set of Item IDs to get the count for the CUI. This transformation was done before applying tf-idf normalization. Figure 3-3 depicts an example of this conversion using the "all" method.

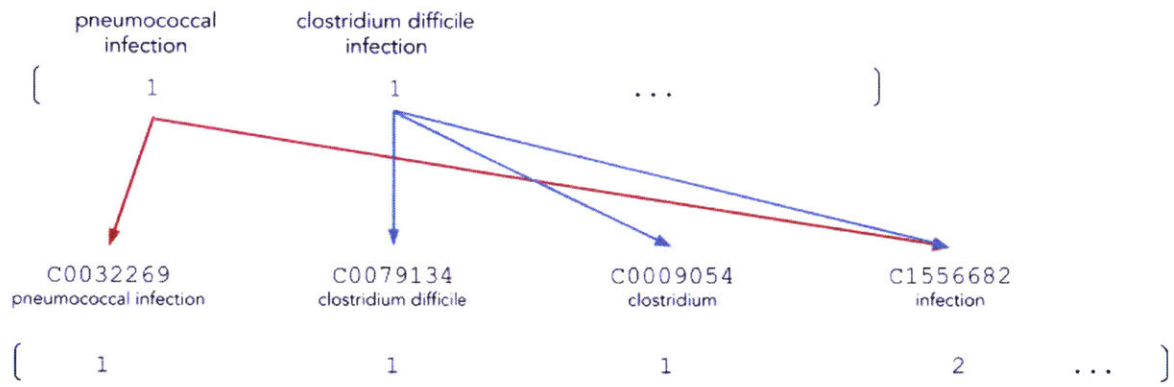


Figure 3-3: Transformation of Item IDs BOE representation to CUIs BOE representation using the *all* method.

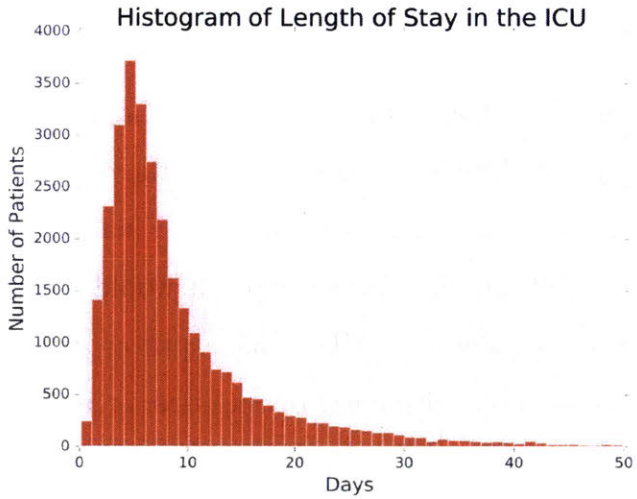


Figure 3-4: Length of stay in the ICU in MIMIC-III. Outliers (LOS > 50 days) truncated for clarity of visualization.

3.4 Experimental Setup

In these experiments, our goal is to demonstrate the utility of our method in building models across related databases. We chose not to combine the databases to build a single risk model in order to clearly demonstrate the utility of our approach for transferring models across databases.

Table 3.1: Number of patients N and clinical outcomes n (in-hospital mortality and prolonged length of stay, i.e. LOS > 11.3 days) in CareVue (2001-2008) and MetaVision (2008-2012) portions of MIMIC III.

EHR	**In-Hospital Mortality**		**Prolonged Length of Stay**	
	N	n	N	n
CareVue	18,244	1,954 (10.7%)	16,735	4,893 (29.2%)
MetaVision	12,701	1,125 (8.9%)	11,758	2,798 (23.8%)
Total	30,945	3,079 (9.9%)	28,493	7,691 (27.0%)

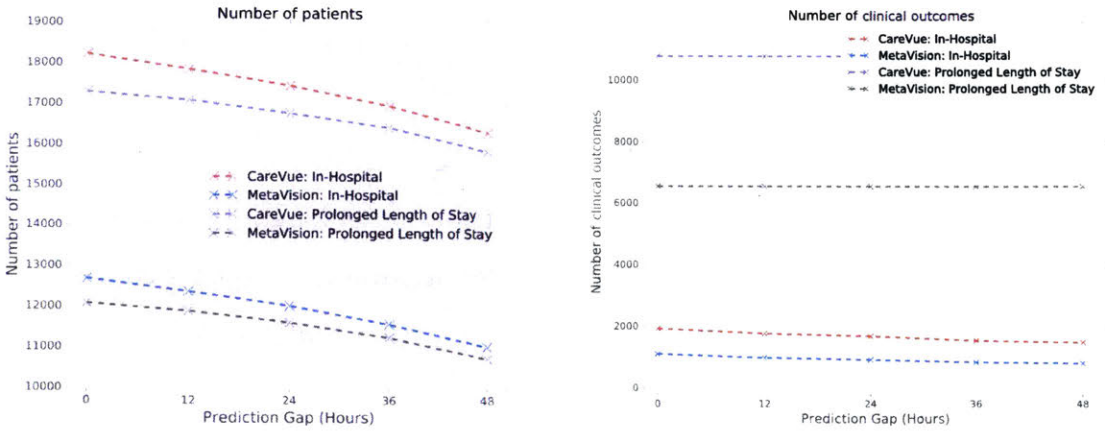


Figure 3-5: Number of patients remaining in the ICU (left) and clinical outcomes (right) with prediction gap 0–48 hours.

3.4.1 Task Definition

We considered patients of at least 18 years of age. We included only these patients' first ICU stay so as to avoid multiple entries for a single patient. This filtering is important because it removes the possibility of training and testing on the same patient (even if they are different ICU stays). We also removed the set of 120 patients whose stays overlapped with the EHR transition and consequently had data in both CareVue and MetaVision.

In the resulting cohort, we extracted data from the first 24 hours of each patient's stay. This provides a fair comparison against baseline acuity scores, which commonly use only information from this time period [44].

We considered the two tasks of predicting in-hospital mortality and prolonged length of stay (LOS). In-hospital mortality is defined as death prior to discharge from

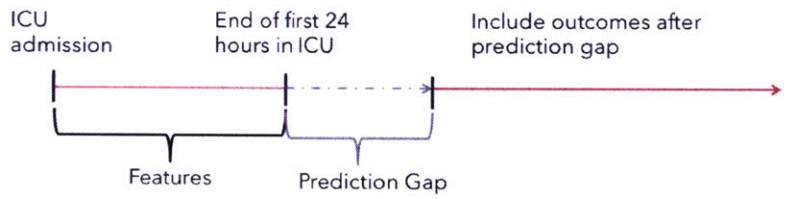


Figure 3-6: Diagram of relationship between information used to construct feature vector (first 24 hours in the ICU) and prediction gap between information used and outcomes.

the hospital. We define prolonged LOS in the ICU as a stay exceeding the upper quartile (> 11.3 days). Figure 3-4 shows the distribution of length of stay across the patients in the ICU. Table 3.1 shows the number of patients in each EHR and the number of cases of the two outcomes. For prolonged LOS, we filtered out patients who died before the 11.3 day cutoff. This was to avoid considering patients who died and patients who were discharged before the prolonged LOS cutoff as equivalent classes. Because of this, the number of patients (N) considered for the outcome of prolonged LOS was lower than the number considered for the outcome of in-hospital mortality.

We considered several prediction gaps ranging from 0 hours (immediately following observation) to 48 hours in 12 hour increments. The prediction gap is the time from the end of the first 24 hours of the ICU stay to when we start counting outcomes. Any patient who experienced the outcome of interest or was discharged during the prediction gap was removed from the data before modeling. This impacts performance by removing the easier cases. For example, a patient who has an item such as "comfort measures only" in the first 24 hours would have an easily predicted outcome. Increasing the prediction gap removes such patients from consideration. Figure 3-5 shows both the number of patients remaining in the ICU and the number of clinical outcomes as we increase the prediction gap (diagrammed in Figure 3-6) for both CareVue and MetaVision.

3.4.2 Model Definition

For all of the experiments, we learned L2-regularized logistic regression models with an asymmetric cost parameter:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_+ \sum_{i:y_i=+1} \log\left(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}\right) + C_- \sum_{i:y_i=-1} \log\left(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}\right) \tag{3.1}$$

We used the scikit-learn LIBLINEAR implementation to train and test all models [49, 50]. We used logistic regression because the model is linear in the features. Therefore the model weights are clinically interpretable, facilitating assessment of the relative importance of features. We employed L2-regularization to reduce the risk of overfitting, since our data are small relative to the data dimensionality (see Table 3.1).

We used 5-fold stratified cross-validation on the training set to select the best value for C_-. We searched for the value in the range 10^{-7} to 10^0 in powers of 10. We set the asymmetric cost parameter ($\frac{C_+}{C_-}$) to the class imbalance (i.e., the ratio of the number patients who did not experience the outcome to the number of those who did). We evaluated our method using the area under the receiver operating characteristic curve (AUC). The AUC captures the trade-off between the false positive rate and the true positive rate of a classifier when sweeping a threshold.

3.5 Experimental Results

3.5.1 EHR-specific Item IDs: Bag-of-Events Feature Representation

We first demonstrate that the simple BOE representation with EHR-specific Item IDs is able to predict clinical outcomes such as mortality and prolonged length of stay. We show the performance against the Simplified Acute Physiology Score II (SAPS-II) [44], a well-established acuity score that is commonly used as a baseline when developing risk models for mortality in the ICU [46, 51, 52] and also uses information

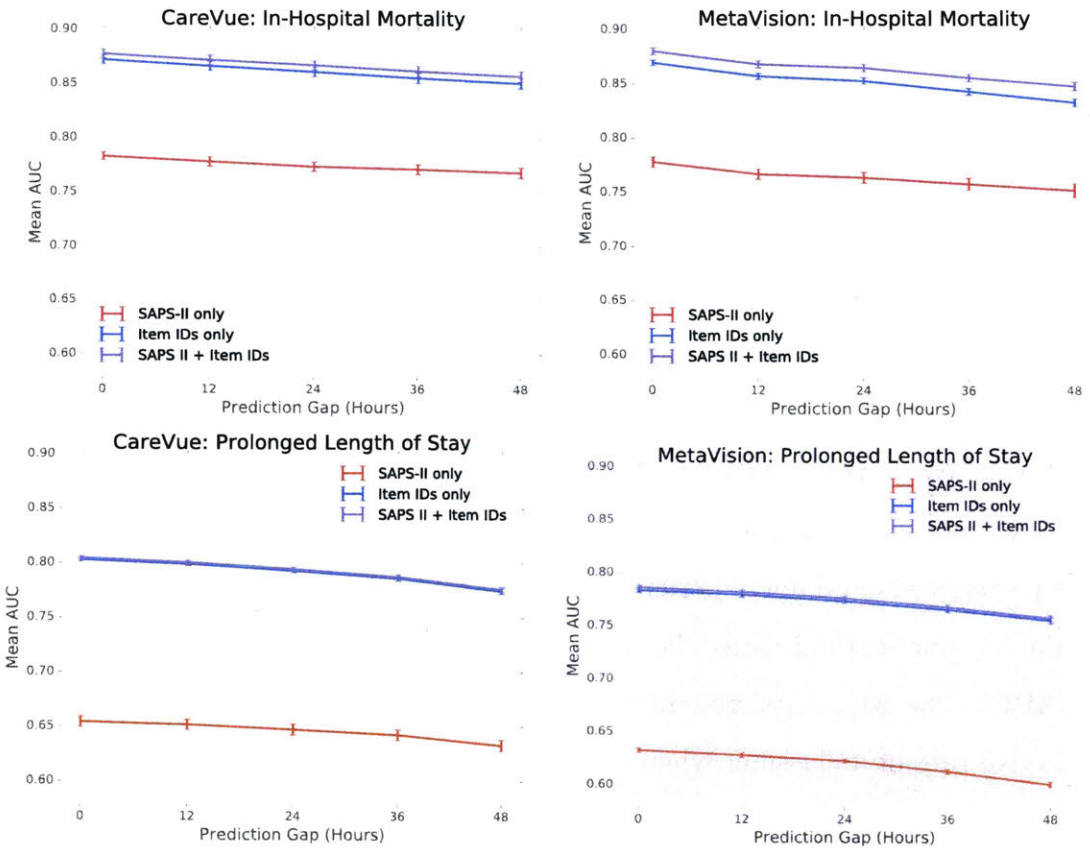


Figure 3-7: Mean AUC across 10 2:1 stratified holdout sets and 95% confidence interval shown for each database and outcome considered. Item IDs + SAPS-II (purple) significantly outperforms Item IDs-only (blue) or SAPS-II only (red) in predicting in-hospital mortality (top) and prolonged LOS (bottom) in CareVue (left) and MetaVision (right).

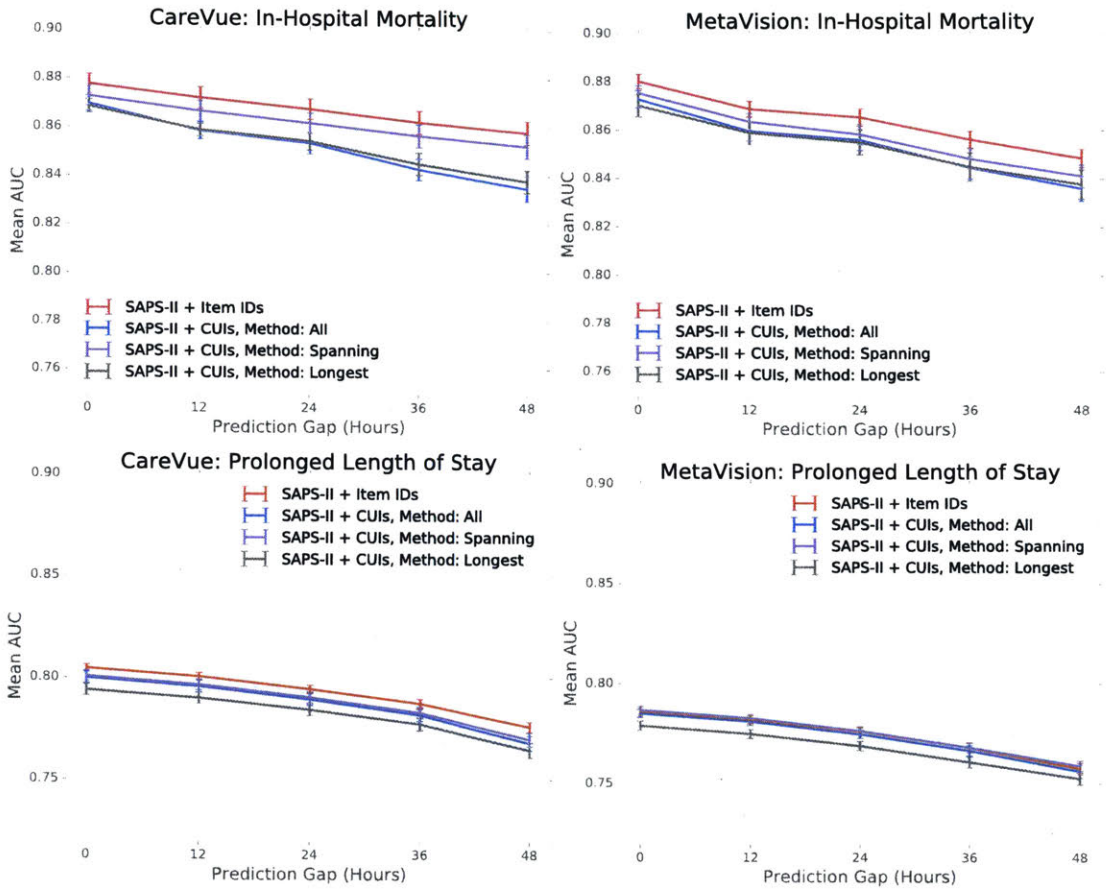


Figure 3-8: Mean AUC across 10 2:1 stratified holdout sets and 95% confidence interval shown for each database and outcome considered. Converting to CUIs from Item IDs results in small, but statistically significant differences in performance in 3 out of the 4 tasks considered. Mean AUC across prediction gaps shown for the outcomes of in-hospital mortality (top) and prolonged LOS (bottom) in CareVue (left) and MetaVision (right).

Table 3.2: Outcome: In-Hospital Mortality. Difference in AUC between SAPS II + Item IDs and SAPS II + CUIs (Spanning) shown. Statistical Significance evaluated using the Wilcoxon Signed-Rank Test.

Prediction	**CareVue**		**MetaVision**	
Gap (Hrs)	Mean Difference in AUC	p-value	Mean Difference in AUC	p-value
0	0.0050	**0.0051**	0.0048	**0.0051**
12	0.0055	**0.0051**	0.0052	**0.0051**
24	0.0058	**0.0051**	0.0071	**0.0051**
36	0.0056	**0.0051**	0.0080	**0.0051**
48	0.0056	**0.0051**	0.0074	**0.0051**

from the first 24 hours in the the ICU.

We evaluate performance on CareVue and MetaVision separately. We computed the AUC on 10 2:1 stratified training:holdout splits. We show that the Item ID BOE features add auxiliary information to the physiological variables captured by SAPS on its own (Figure 3-7). We used the Wilcoxon signed-rank test [53] to evaluate significance of the differences between the Item IDs-only results and the SAPS-II + Item IDs results. The Wilcoxon signed-rank test is nonparametric, and therefore makes no assumptions of normality. All differences for both outcomes and both databases were statistically significant (p-value = 0.0051). Although the magnitudes of the differences are not large (between 0.005 and 0.015 across all prediction gaps for all tasks), they are consistent. In the following experiments, we used the SAPS-II + BOE (Item IDs or CUIs) feature space.

3.5.2 Mapping Item IDs to CUIs Does Not Dramatically Change Predictive Performance

We evaluate the predictive performance of the BOE features when the events counted are represented by UMLS concept unique identifiers (CUIs) rather than EHR-specific Item IDs. We compare the performance of a model trained using SAPS-II + CUIs vs. SAPS-II + Item IDs for each of the tasks of interest. We evaluate the three methods of translating item descriptions to CUIs described in Section 3.3.1.

The mean AUCs across 10 2:1 stratified training:holdout splits are shown in Figure 3-8, and the Wilcoxon signed-rank test p-values for in-hospital mortality and

prolonged length of stay are shown in Table 3.2 and Table 3.3, respectively. The mean differences in AUCs across all of the prediction gaps were statistically significant for both outcomes in Carevue, but only for the outcome of in-hospital mortality in MetaVision ($p = 0.0051$). However, these differences are small in magnitude (Δ AUC ≤ 0.008). For the outcome of prolonged LOS, the differences in MetaVision between SAPS II + Item IDs and SAPS II + CUIs were not statistically significant. Thus, although some statistically significant decreases in AUC occur when CUIs are used, they are small in magnitude. These small differences demonstrate that representing clinical events using CUIs can achieve high predictive performance on predicting mortality in the ICU within a *single* EHR system.

As Figure 3-8 shows, the *spanning* method appears to have better or comparable performance to the other approaches across the four tasks. We therefore use the *spanning* method going forward to map to the CUI BOE representation. Table 3.4 shows the number of item IDs in each EHR version and the resulting number of CUIs from the cTAKES mapping using the *spanning* approach.

3.5.3 CUIs Enable Better Transfer Across EHR Versions

We evaluate performance on predicting in-hospital mortality and prolonged length of stay *across* EHRs. To do this, we train a model on data from one EHR system (Train DB) and evaluate on data from the other EHR system (Test DB). We hypothesize that models trained on CUIs will better generalize across EHRs compared to Item IDs because 1) mapping to CUIs removes redundancy within each EHR, particularly CareVue, and 2) the intersecting set of CUIs between EHRs is larger than the intersecting set of Item IDs relative to the number of features in each EHR.

We compare our approach of training a model on CUIs to two baselines: 1) training on *all* Item IDs from Train DB (Figure 3-9(a)), and 2) training on the *shared* set of Item IDs between Train DB and Test DB (Figure 3-9(b)). Training on *all* Item IDs from Train DB and testing on Test DB effectively means excluding most of the charted events from consideration during prediction. While this obviously will not result in the best prediction performance, it is a realistic simulation of how a model

Table 3.3: Outcome: Prolonged Length of Stay. Difference in AUC between SAPS II + Item IDs and SAPS II + CUIs (Spanning) shown. Statistical Significance evaluated using the Wilcoxon Signed-Rank Test.

Prediction	**CareVue**		**MetaVision**	
Gap (Hrs)	Mean Difference in AUC	p-value	Mean Difference in AUC	p-value
0	0.0048	**0.0051**	0.0001	0.7989
12	0.0053	**0.0051**	0.0015	0.5076
24	0.0071	**0.0051**	0.0017	0.3863
36	0.0080	**0.0051**	0.0017	0.2845
48	0.0074	**0.0051**	0.0018	0.2845

that has been developed on one database version might directly be applied to data from a new schema early on in a transition.

These results are shown in Figure 3-10. 95% confidence intervals are shown on the test AUC, generated by bootstrapping the test set 1000 times to have the same size and class imbalance as the original test set. The difference between the training AUC and test AUC provides a sense of how well the model is able to generalize from Train DB to Test DB, and to what extent it is overfitting to the training data.

These results demonstrate that the models trained on CUIs outperform those trained on both *all* and *shared* Item IDs for both outcomes. In addition, the difference between the training and test AUC when *all* Item IDs are used (red lines) is much larger than the same difference when CUIs are used, or when *shared* Item IDs are used. This demonstrates that using CUIs is less prone to overfitting and results in more generalizable models.

Table 3.4: Number of Item IDs and CUIs in CareVue, MetaVision, and intersection for in-hospital mortality after filtering ($\geq$ 5 occurrences in data). For MetaVision, the filter selects 2,438 of the 5,190 features. For CareVue, the filter selects 5,875 of the 15,909 features.

Prediction Gap	**CareVue**		**MetaVision**		**Intersection**	
(Hrs)	Item IDs	CUIs	Item IDs	CUIs	Item IDs	CUIs
0	5875	3660	2438	2192	2118	2052
12	5843	3645	2421	2182	2102	2046
24	5795	3619	2405	2175	2094	2041
36	5746	3595	2384	2161	2076	2035
48	5703	3573	2351	2151	2048	2017

Using the UMLS CUIs, we increase the AUC on in-hospital mortality by at least 0.01 across all tasks. Similarly, we improve the AUC on prolonged LOS by at least 0.009 when training on MetaVision and testing on CareVue. When we train on CareVue and test on MetaVision, we achieve larger improvements compared to *shared* Item IDs (Δ AUC > 0.03) and *all* Item IDs (Δ AUC > 0.07).

For predicting prolonged LOS with a gap of 24 hours when training on CareVue and testing on MetaVision, these differences translate to an AUC of 0.77 (0.76, 0.78) when using CUIs, compared to an AUC of 0.70 (0.69, 0.71) when *all* Item IDs are used and 0.74 (0.73, 0.75) when *shared* Item IDs are used. Converting our EHR-specific Item ID features to a shared CUI representation results in significantly better performance when applying a model learned on data from one EHR version to data from another.

3.6 Summary and Discussion

We introduced an approach to constructing machine learning models that are portable across different representations of semantically similar information. When a database is replaced or a schema changed, there is inevitably a period of time during which there are insufficient data to learn useful predictive models. Our method facilitates the use of models built using the previous database or data schema during such periods.

We demonstrated the utility of our approach for constructing risk models for

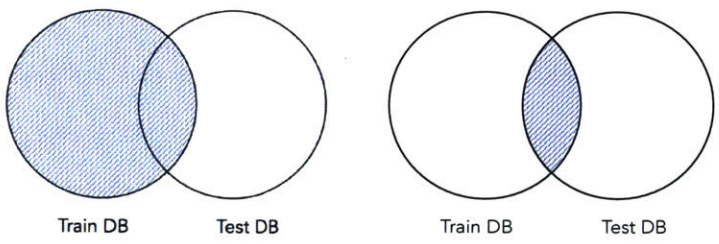


Figure 3-9: Baseline approaches: (a) Train a model on *all* items in the training database (Train DB) (left), and (b) Train a model only on *shared* items that appear in both the training and test databases (right).

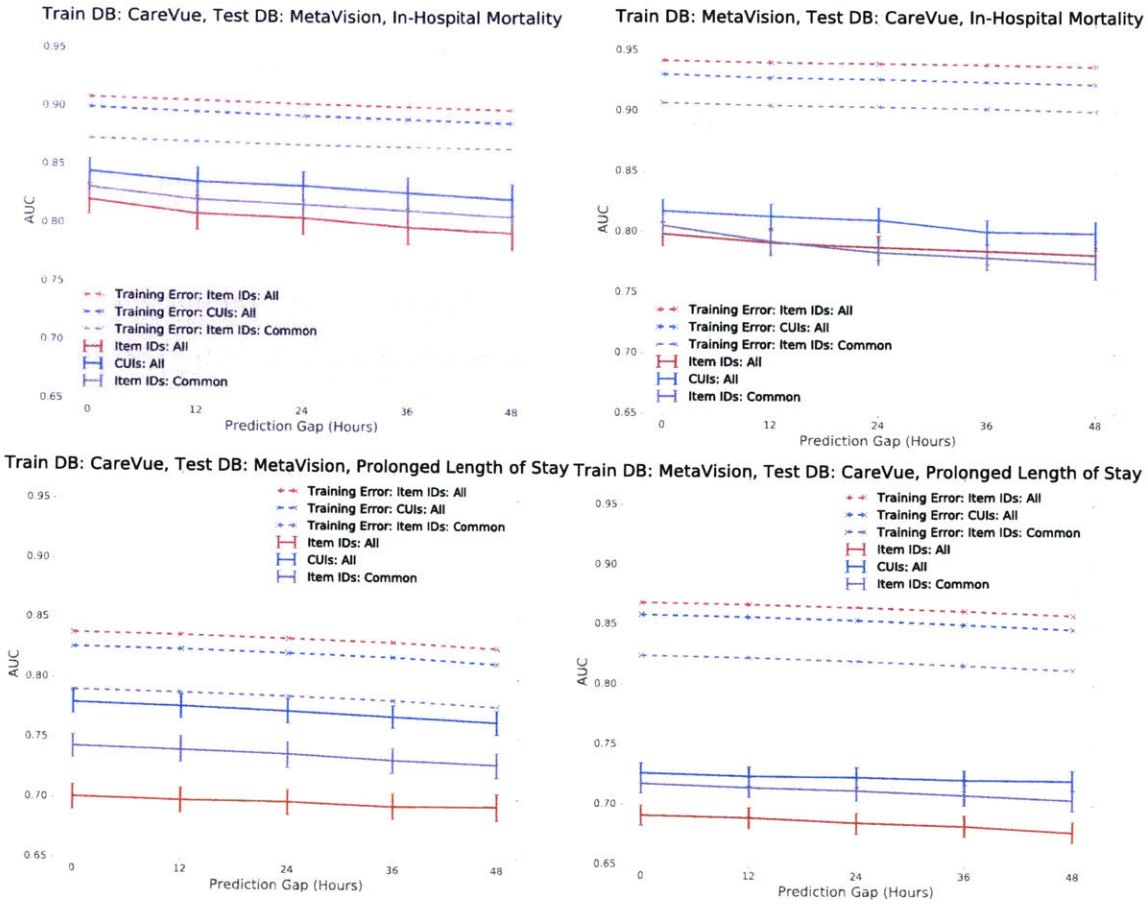


Figure 3-10: AUC when training on TrainDB and testing on TestDB using EHR-specific Item IDs (all), Item IDs (shared), and CUIs. 95% confidence intervals are shown for each database and outcome considered. The dashed lines show the training AUC of each model on Train DB, while the solid lines show the AUC on Test DB. Training using the CUIs representation results in the best training and test AUCs across all prediction gaps compared to Item IDs (all) or Item IDs (shared) representations. These improvements are more pronounced for the outcome of Prolonged Length of Stay when training on CareVue and testing on MetaVision (bottom left).

patients in the intensive care unit. We leveraged the UMLS medical ontology to construct clinical risk models that perform well across two different EHRs on two different tasks: in-hospital mortality and prolonged length of stay. Our method of mapping to CUIs results in increased AUC over EHR-specific item encodings for all prediction gaps, both outcomes, and both directions of training on one EHR and testing on the other.

While our method generalized well across the two EHR versions in our data, our

use of MIMIC-III limits our experiments to data from the same institution. We chose to work with MIMIC because it is an open, freely-accessible database, and it allowed us to conduct a reproducible case study that highlights many of the challenges associated with the portability of models in a more general setting. Applying our method to other institutions could lend insight to how well our approach performs in the presence of different care staff, practices, and patient population characteristics, as well as differences in EHR systems. It would also allow us to investigate how our method performs in transferring models across institutions.

Our method is intended to be a general approach to reconciling semantically equivalent concepts to enable model portability across EHRs. Because of this, our approach is *task-agnostic*. Therefore, the method we propose is generalizable beyond these two outcomes. However, it may be desirable to explore portability of models in a *task-specific* manner. Our approach maps a large set of EHR-specific encodings to a shared set of concepts, but only some of these features may actually be needed to predict an outcome of interest. A task-specific mapping could use information from a model trained on one EHR system to identify only the features that are relevant to the outcome, and then find the corresponding concepts in a new EHR system.

Despite improving performance, our method suffers from several limitations. First, although using the CUI BOE representation leads to significantly higher overlap in feature spaces between the two EHRs (CareVue and MetaVision), a significant number of CUIs is lost when the intersection is taken. We believe that this is the result of insufficient disambiguation of entities from the free-text item descriptions utilized in CareVue. Identifying all relevant concepts from short item descriptions is challenging for existing natural language processing tools that depend on context for term disambiguation. Leveraging other sources of text with sufficient context to disambiguate these terms (e.g., clinical notes) is a plausible way to address this problem.

In this chapter, we addressed the notion of *semantic equivalence* of clinical concepts that are encoded differently in different EHR systems. Semantic equivalence, defined using clinical concepts contained in the text descriptions, is different from *feature* equivalence. Even if the feature spaces of two systems $\mathcal{X}_1$ and $\mathcal{X}_2$ are equiva-

lent, the marginal distributions $P(\mathcal{X}_1)$ and $P(\mathcal{X}_2)$ may not be, because of differences in care practice, monitoring systems, and more. Thus, even after semantic concepts have been reconciled, there may still be differences in the feature spaces that should be adjusted for before effective model transfer is possible. Future work could investigate other methods in the *domain adaptation* area of transfer learning to tackle these problems. Importantly, reconciling semantically equivalent concepts as we do in this work is a necessary step to identifying distributional shifts in values associated with the same concept.

Chapter 4

Learning Cross-Modality Correspondences in Clinical Data

4.1 Introduction

In the previous chapter, we considered a single modality of data, *clinical events*, and demonstrated how the encoding of structured data elements can affect the portability of machine learning models. In this chapter, we leverage the relationship between data modalities from the *structured health record* (i.e., clinical event sequences and physiological time-series) and the *clinical notes* to learn how to summarize the high-dimensional structured health record data using a text-based representation. We use data from the MetaVision portion of MIMIC-III.

Electronic Health Records (EHRs) contain an overwhelming amount of information about each patient, making it difficult for clinicians to quickly find the most salient information at various points during an admission. Information overload can also result in health care providers missing important information during the course of care [12]. Accurate, concise summarization of relevant data can help alleviate this cognitive burden.

Clinical narrative notes serve this purpose during the course of care. They help clinicians summarize and identify the most relevant aspects of the deluge of available data about each patient, and facilitate communication among care teams [17].

However, clinical notes are written at infrequent intervals. Information from the most recent note can quickly become outdated, particularly in critical care settings, where patient state can suddenly change and interventions are frequently administered. Missing information during communication between care team members can lead to adverse events [54]. Methods for assisting care team members in writing summaries of patient state and the course of care could help address potential errors of omission.

In this chapter, we propose a system that generates relevant patient- and time-specific *topics* from structured health record data. We utilize a supervised modeling approach to learn correspondences between detailed, high-dimensional structured data and existing clinical notes. We model each note as a distribution over topics using latent Dirichlet allocation (LDA) [55]. These topics have been shown to capture relevant patient subtypes, and are predictive of adverse outcomes such as mortality [46] and interventions [22]. We then use our model to generate topic-based summaries of structured health record data. Our approach uses a recurrent neural network to learn correspondences between the structured data and clinical note topics over the course of a patient stay.

Our contributions are as follows:

1. We present a supervised framework to learn correspondences between high-dimensional structured EHR data elements and low-dimensional topic representations of clinical notes over the course of a patient stay. This model can be used to summarize patient care and physiology – even when a note was never written.

2. We evaluate the generated topic distributions. We show that the generated topic distributions reflect changes in patient state earlier than recorded clinical notes, and reflect meaningful correspondences between topics and relevant structured items.

3. We show that using structured data alone to predict the next note performs similarly to using all prior notes when they exist. In addition, structured data

can accurately predict the first note in a patient stay, when a model using only the notes data has no information. We show that combining structured data and notes can improve predictions over either one alone when prior notes exist.

4. We evaluate topics generated from the structured data alone by evaluating performance on two downstream prediction tasks: in-hospital mortality and first intubation. We demonstrate that using only our predicted notes leads to comparable performance to using the actual notes.

We first discuss related work in Section 4.2. Next, we describe our data processing methods in Section 4.4. We describe our methods in Section 4.5, and our experimental results in Section 4.6. Finally, we summarize our findings in Section 4.7.

4.2 Related Work

4.2.1 Summarizing Health Record Data

A great deal of work has investigated how to summarize structured health record data in a more accessible manner. Some works have utilized visualization interfaces [56, 57, 58]. Others have used natural language to generate descriptions of structured time-series [59, 60, 61]. [62] contains a comprehensive summary of techniques for summarizing health record data. In contrast to these works, our goal is to automatically learn correspondences between structured data and existing summaries written during the course of care.

4.2.2 Modality Translation

Our task is a translation from one modality of data (structured health record) to another (clinical text). Recent work utilizing deep learning approaches has demonstrated success in translating images to text descriptions. Approaches such as those proposed by [63] seek to generate an entire text caption when given an image. As in our approach, these methods utilize recurrent neural networks and a supervised framework.

In the medical domain, [64] presents an approach utilizing convolutional and recurrent neural networks to generate text annotations of chest X-rays by leveraging both the image itself and the associated radiology report. [65] presents a multi-component approach to automatically generate descriptions of echocardiograms. This approach utilizes three neural networks that are first separately trained to 1) use *doc2vec* to map text reports to a fixed length, dense vector, 2) extract useful image features from each available image, and 3) transform from a fixed length vector representing the corresponding image to a fixed length vector representing a text document.

All of these works generate text corresponding to static images. Our problem differs in two ways. First, we translate temporal structured health record data (rather than static images) to a text-based representation. Second, rather than predicting a sequence of text, we seek to predict an intermediate representation. This is a first step to understanding the relationship between complex structured health record data and the information content in clinical notes.

4.2.3 Clinical Note Time-Series

Our work leverages cross-modal data relationships to predict notes at times when they are not usually written. [51] handles the problem of missing notes by learning a multi-task Gaussian Process (MTGP) over the time-series of clinical notes. The authors do not evaluate the ability of the MTGP to forecast notes. They instead demonstrate the utility of the MTGP parameters for downstream prediction tasks (e.g., in-hospital mortality). In contrast, we are interested in the task of forecasting topic membership of missing clinical notes, to generate summaries of care even when they are not present.

[66] models evolving patient state from nursing notes using a model that integrates a hidden Markov model (HMM) and latent Dirichlet allocation (LDA). This model captures changing patient dynamics (and therefore changing topic memberships) over time, but does not consider the additional value of structured health record data for generating clinical note topics.

4.2.4 Integrating Clinical Data Modalities

In this work, we consider physiological time-series, clinical events, and clinical notes. Each modality of data has been shown to be successful in predicting clinical outcomes such as mortality (e.g., [24, 46, 67]) and intervention administration [21, 20]. In addition, multi-modal EHR data have been integrated, primarily for the tasks of 1) patient phenotyping (e.g., [68, 69, 70]), and 2) clinical outcome prediction (e.g., [22, 71, 72]). While we demonstrate the utility of our learned correspondences in downstream predictive tasks, we are primarily focused on the task of learning a correspondence between the structured data time-series and a note summarizing patient status and the care process.

4.3 Cohort Selection

We considered patients $\geq$ 15 years of age. Because the encodings of clinical events differed significantly between the two EHR systems in MIMIC-III, we considered only data from the latter version (MetaVision, 2008-2012). We used each patient's first ICU stay, to avoid multiple admissions from the same patient. Patients who died, were discharged, or had a note of "comfort-measures only" within 12 hours of ICU admission were removed from the study. Patients missing any of the three modalities of data were also removed, reducing our patient population from $> 15,000$ patients to 6,360 patients. This difference was primarily a result of dropping patients without regular physician and nursing notes. The differences between patients with and without notes are detailed in Table 4.1. Patients with missing notes are not noticeably different from patients with notes in length of stay in the ICU, presence in different care units, or admission status. However, mortality rate was elevated in patients with missing notes. We divided the remaining patients into a 60/20/20 training/validation/test split. These divisions are described in Table 4.2.

Table 4.1: Differences in length of stay, care units, admission type, and adverse outcome incidence between patients with and without physician, nursing, and general notes.

	Notes Missing	Notes Present
Number of patients	9171	6360
Mean LOS in ICU (days)	2.6	2.5
In-Hospital Mortality (%)	8.8	7.2
Intubation (%)	39.5	36.5
CCU (%)	12.4	13.0
CSRU (%)	17.1	16.7
MICU (%)	38.7	38.6
SICU (%)	18.8	18.1
TSICU (%)	13.0	13.7
Elective admission (%)	16.7	15.4
Emergency admission (%)	82.3	83.1
Urgent admission (%)	1.0	1.5

Table 4.2: Cohort and training/validation/test data split descriptions.

	Train	**Validation**	**Test**
Number of Patients	3816	1272	1272
Number of Notes	111,938	34,553	38,747
In-Hospital Mortality	7.0%	7.5%	7.2%
Mean (std) LOS in ICU (days)	2.5 (1.9)	2.4 (1.8)	2.5 (2.0)

4.4 Data Processing

All data were aligned to midnight on the day of ICU admission, to preserve time-of-day characteristics, and discretized to the hour. All admissions were padded or truncated to 96 hours from midnight of the first day of ICU admission. The following sections describe processing details for each data modality.

4.4.1 Events

We discretized the times of events to the hour, from midnight on the day of ICU admission. Events that occurred in the same hour were represented with a binary bag-of-events (BOE) vector, indicating whether or not each event occurred in that hour. We considered two types of events: 1) point events, that occurred at a single

point in time, and 2) duration events, which were specified with a start and stop time. Events that spanned a duration of time were 1 between the start and stop times, and 0 otherwise. Point events were 1 if the event was present and 0 otherwise. The events tensor was then constructed by building this BOE vector over time. Events that occurred in fewer than three unique admissions in the training data were filtered out. In total, we considered 6,556 kinds of events.

4.4.2 Physiological Time-Series

Continuous-valued vital signs and lab test measurements were binned to the hour by taking the median of the values in each hour. The hourly values were then discretized by taking the z-score, rounding to the nearest integer, and mapping outliers ($|z| > 4$) to -4 and 4, following the procedure used in [22] and [20]. The means and standard deviations of all of the features were determined across all admissions in the training and validation data. These features were then binarized. An additional bin was added for each variable to indicate a missing value.

4.4.3 Notes

We filtered out a set of pre-defined clinical stop words (e.g., patient, report, pt, admission, discharge, etc.), as well as tokens that occurred in fewer than 3 documents or more than 95% of documents. Additionally, punctuation and numerical values were filtered out. We used latent Dirichlet allocation [55] to reduce the dimensionality of the clinical notes from a $> 47K$ vocabulary to a distribution over 50 topics. Topic models were trained using gensim [73]. For each patient, the topic distribution at each hour was computed by taking the average of the topic distributions for all notes in that hour.

Table 4.3 describes the top five and bottom five topics by enrichment for in-hospital mortality. Enrichment was computed using the training data by taking the average topic probability for each topic across all notes, weighted by the outcome of the patient the note was written about, as in [74]. The full set of topics is described

Table 4.3: Top 5 and bottom 5 topics by enrichment for in-hospital mortality.

Topic	**Top 5**
14	family, care, dnr, support, daughter, dni, son, comfort, morphine, social
37	hypotension, line, shock, sepsis, levophed, cvp, fluid, bp, pressors, map
16	liver, cirrhosis, lactulose, transplant, encephalopathy, ascites, hepatic, varices, sbp, albumin
25	spontaneous, rr, min, set, vt, tube, ventilator, peep, mode, vc
36	intubated, sedation, vent, propofol, abg, extubation, sedated, fentanyl, wean, respiratory
Topic	**Bottom 5**
15	etoh, abuse, ciwa, withdrawal, alcohol, pancreatitis, valium, scale, thiamine, seizures
43	pain, control, chronic, acute, continue, prn, dilaudid, morphine, po, iv
42	valuables, transferred, rate, pmh, weight, heart, bp, total, sent, money
13	present, pulse, min, extremities, mmhg, current, regular, rhythm, insulin, chest
38	cabg, artery, wires, coronary, bypass, temporary, graft, svg, avr, valve

in Table 4.4.

4.5 Methods

4.5.1 Learning Correspondences

To learn correspondences between the structured clinical data and the clinical notes, we utilize a supervised deep learning approach that leverages the temporal nature of the structured data and the clinical notes.

Network Architecture

struct2note uses all structured data up to and including the hour of the note of interest to predict topics for a clinical note. We compare against two other models which leverage prior notes: 1) *notes2note* uses all prior notes to make a prediction, and 2) *struct-notes2note* integrates prior notes and structured data. Figure 4-1 diagrams our model architecture for *struct2note*.

In *struct2note*, a temporally shared fully-connected embedding layer with a rectified linear activation function maps the structured data at each time step from a sparse, high-dimensional feature space to a low dimensional dense embedding space. This captures relationships between co-occurring events at each time-step. We use a long short-term memory (LSTM) network to capture the temporal patterns in the structured data [75]. LSTMs have been shown to encode temporal patterns that are

Table 4.4: Top 10 tokens describing each topic.

Topic	Tokens
0	post, surgery, op, epidural, bladder, repair, iabp, stent, urology, pain
1	pleural, effusion, chest, tube, ct, effusions, fluid, drain, cxr, placement
2	fluid, na, stool, acidosis, diarrhea, diff, sodium, free, hyponatremia, cont
3	cancer, mass, ca, metastatic, lung, malignant, tumor, neoplasm, chemo, cell
4	skin, left, right, site, wound, groin, area, leg, impaired, intact
5	pain, abdominal, nausea, ct, vomiting, abd, ercp, zofran, iv, abdomen
6	lithium, morbid, myasthenia, suprapubic, mtx, girlfriend, atropine, cystitis, aureus, shocks
7	respiratory, pneumonia, pna, copd, aspiration, cxr, distress, bipap, sputum, nebs
8	code, continue, total, balance, rhythm, review, systems, labs, comments, prophylaxis
9	mental, status, altered, airway, delirium, cont, aspiration, agitation, agitated, risk
10	heparin, pe, ptt, started, dvt, gtt, pulmonary, transferred, cta, filter
11	impaired, problem, description, skin, enter, abscess, comments, integrity, tooth, clindamycin
12	right, left, ct, fractures, hematoma, injury, lobe, chest, posterior, thoracic
13	present, pulse, min, extremities, mmhg, current, regular, rhythm, insulin, chest
14	family, care, dnr, support, daughter, dni, son, comfort, morphine, social
15	etoh, abuse, ciwa, withdrawal, alcohol, pancreatitis, valium, scale, thiamine, seizures
16	liver, cirrhosis, lactulose, transplant, encephalopathy, ascites, hepatic, varices, sbp, albumin
17	seizure, sdh, dilantin, subdural, activity, neuro, seizures, brain, head, keppra
18	hct, bleeding, blood, stable, prbc, monitor, bleed, inr, cont, transfusion
19	afib, atrial, fibrillation, coumadin, rate, af, fib, po, metoprolol, amiodarone
20	gi, bleed, hct, bleeding, gib, egd, stable, gastrointestinal, protonix, upper
21	lasix, chf, diuresis, edema, failure, iv, heart, chronic, acute, goal
22	cath, cardiac, cad, heparin, chest, asa, nstemi, plavix, pain, disease
23	fever, temp, cont, wbc, cultures, sent, abx, cx, vanco, culture
24	neuro, commands, exam, extremities, eyes, pupils, checks, continue, noted, monitor
25	spontaneous, rr, min, set, vt, tube, ventilator, peep, mode, vc
26	arrest, cardiac, vt, icd, av, ccu, bradycardia, ep, rhythm, pacer
27	fx, fracture, fall, trauma, rib, collar, multiple, neck, injuries, pain
28	insulin, dm, diabetes, type, blood, gtt, scale, sliding, fs, bs
29	iv, order, total, extremities, rhythm, current, po, prn, fluid, balance
30	bed, oriented, oob, able, swallow, po, speech, chair, today, alert
31	present, normal, sounds, left, right, cardiovascular, respiratory, nose, pulse, absent
32	left, ct, head, hemorrhage, right, neuro, sbp, sah, stroke, sided
33	gtt, monitor, sbp, iv, bp, continue, remains, stable, noted, shift
34	neo, map, hypothermia, wean, pad, bair, hugger, temp, bypass, sfa
35	note, time, agree, section, protected, resident, present, saw, examined, services
36	intubated, sedation, vent, propofol, abg, extubation, sedated, fentanyl, wean, respiratory
37	hypotension, line, shock, sepsis, levophed, cvp, fluid, bp, pressors, map
38	cabg, artery, wires, coronary, bypass, temporary, graft, svg, avr, valve
39	likely, continue, pending, culture, negative, blood, cultures, infection, consider, cx
40	renal, failure, acute, hd, arf, chronic, cr, urine, bun, kidney
41	po, pain, denies, past, ed, prn, home, chest, prior, recent
42	valuables, transferred, rate, pmh, weight, heart, bp, total, sent, money
43	pain, control, chronic, acute, continue, prn, dilaudid, morphine, po, iv
44	abd, bowel, drainage, soft, output, urine, draining, bs, abdomen, ngt
45	ct, head, mri, status, mental, negative, osh, lp, spine, eeg
46	left, aortic, valve, right, normal, ventricular, mitral, systolic, stenosis, wall
47	ed, received, micu, bp, transferred, noted, iv, arrival, started, sent
48	sats, cough, nc, clear, face, mask, diminished, resp, bases, secretions
49	assessed, pulse, total, comments, left, right, balance, review, systems, labs

effective in predicting interventions and identifying patient diagnoses [22, 23]. Finally, a temporally shared fully-connected layer with a softmax activation outputs predicted probabilities for the 50 topics at each time step. *notes2note* uses a similar architecture, but because the topics are already a dense embedding space, we do not need an embedding layer. *struct-notes2note* combines both data modalities by concatenating the topic distribution tensor with the embedded structured data tensor as the input through the LSTM.

Loss Function and Evaluation Metric

To compare predicted topic distributions with the true topic distributions, we utilize *cosine similarity*. Cosine similarity is defined as the normalized dot product between two vectors:

$$C(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}. \tag{4.1}$$

It takes a maximum value of 1 when $\mathbf{u}$ and $\mathbf{v}$ are parallel, a value of 0 when $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, and a minimum value of -1 when $\mathbf{u}$ and $\mathbf{v}$ are anti-parallel. In our application, the minimum value the cosine similarity measure can take is 0, because we are comparing two probability distributions (all elements are non-negative). Cosine similarity is an appropriate loss function because it evaluates how close $\mathbf{u}$ and $\mathbf{v}$ are in directionality, rather than in magnitude. Because our topic distributions always sum to 1, magnitude is not important in assessing the differences between the topic distribution of the actual note and the predicted topic distribution. Cosine similarity has been used in prior work to evaluate differences between dense embeddings of words [76]. We use cosine similarity both as the loss function during training, and as an evaluation metric to determine how close our predicted topic distributions are to the true ones.

Clinical notes are not present at every time step. The cosine similarity loss is only considered at time-steps when notes are present. When prior clinical notes are used as input to the *notes2note* and *struct-notes2note* models, notes are forward-filled with the most recent note up until the latest of time of death, discharge, or the final note.

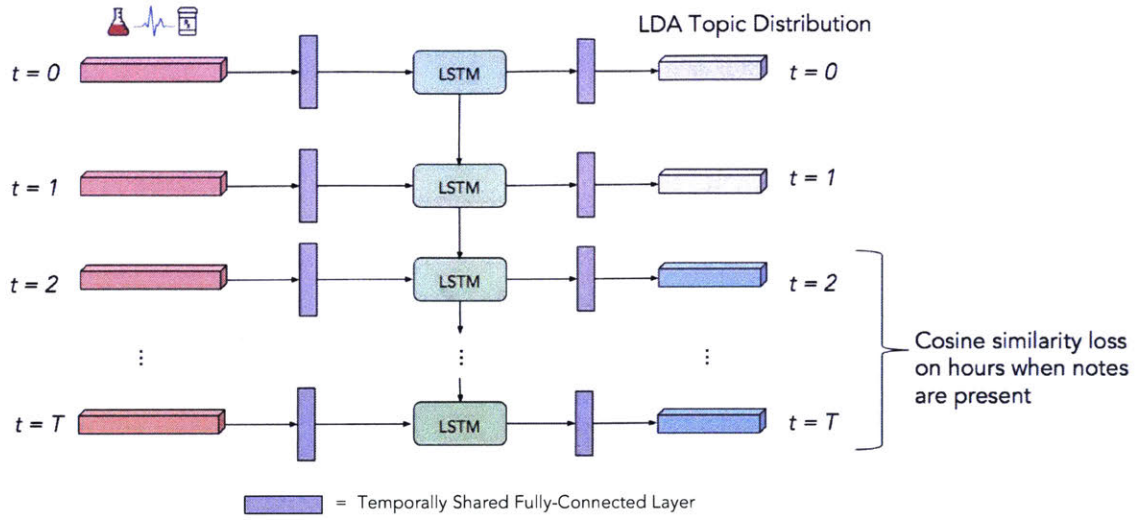


Figure 4-1: Model architecture for structured data (*struct2note*). The network is shown unrolled over time. Sparse, high-dimensional time-series of structured data are first passed through a fully-connected layer shared over time to get a dense embedding. The time-series are then encoded using an LSTM. The topic distribution for the note at each time step is predicted with a fully-connected layer (shared over time) with a softmax activation. During training, the loss was computed on hours when notes were present.

Time-steps where input data are not present (e.g., prior to ICU admission on the first day) are masked out.

Training and Implementation

We implemented our models using Keras 2.1.3 with Tensorflow backend (1.5.0) [77]. The size of the first temporally shared fully-connected layer for embedding the structured data was set to 30 units, and a grid search from 8 to 256 in multiples of 2 was performed to choose the LSTM hidden layer size. All models were chosen based on the validation loss.

4.5.2 Outcome Prediction

To demonstrate that our predicted note topics capture meaningful aspects of patient care and state, we predict in-hospital mortality and first intubation using models trained on 1) existing clinical notes, and 2) the predicted clinical notes. In-hospital

mortality is often used as a proxy for patient severity of illness. First intubation is an important outcome because it indicates the need for a severe intervention. We utilize the network architectures from Section 4.5.1, replacing the topics-over-time tensor with a binary tensor indicating the outcome for that patient at that time.

For the task of in-hospital mortality, we predict whether or not in-hospital mortality occurs at least 24 hours after the hour the prediction is made. We define the outcome using the earliest of the patient's time of death or a note of "comfort-measures only" (CMO). When a patient is declared CMO, few (if any) interventions are made, and the prediction is no longer relevant to the course of care. At each hour, a prediction is made for each patient. Predictions for patients who are discharged or die prior to the hour of prediction or within the 24 gap period are excluded from the loss at that time step. Models trained for in-hospital mortality were further restricted compared to the training, validation, and test sets described in Table 4.2: patients who died, were discharged, or had a note of "CMO" in the first 24 hours of the ICU stay were filtered out from model training and evaluation.

When predicting intubation, we follow the framework of [21] and [22] and predict intubation in the next 4 hour window, following a gap period of 4 hours. We make a prediction at each hour of the patient's stay. We consider only the first intubation event for each patient [21]. A patient is assigned an outcome of $+1$ if she is intubated anytime in the 4 hour window after this gap period. If the patient has been intubated at any point prior to the hour of prediction, or is discharged or dies, she is filtered out. The intubation prediction task also considered a reduced population compared to Table 4.2; patients who died, were discharged, had a note of "CMO" in the first 24 hours of the ICU stay, were intubated in the first 6 hours of the ICU stay, or had an indication of "Do Not Intubate" were filtered out. This follows the filtering procedure used in [21]; intubation events that happen close to ICU admission may be substantively different from those occurring later in the ICU stay.

In contrast to prior works predicting intubation, we utilize the entire duration of the patient's stay up to the point of intubation, death, or discharge to make predictions at each hour, rather than windowing the data and making predictions

Table 4.5: Cosine similarity performance of different models on test set. Mean, standard deviation, and quartiles of performance are shown, broken down by notes where a prior note existed, and notes where no prior note existed.

	Notes with prior notes (9290)				Notes without prior notes (1272)			
	mean (std)	25%	50%	75%	mean (std)	25%	50%	75%
notes2note	0.63 (0.19)	0.50	0.65	0.78	0.41 (0.09)	0.35	0.42	0.48
struct2note	0.63 (0.21)	0.49	0.66	0.80	**0.61** (0.17)	0.49	0.62	0.74
struct-notes2note	**0.66** (0.21)	0.53	0.69	0.82	**0.61** (0.17)	0.49	0.62	0.73
Prior note	0.39 (0.29)	0.15	0.31	0.62	–	–	–	–
Average note	0.40 (0.09)	0.34	0.41	0.46	0.42 (0.09)	0.36	0.42	0.48

using only the immediate past. In addition, we use physiological signals and clinical notes, as in [21] and [22], and events data.

4.6 Results

4.6.1 Predicting the Next Note

To evaluate our model's ability to predict topic vectors for existing clinical notes, we compared against two baselines: 1) *prior note* topic membership, where we used the most recent note topic membership to predict that of the current note, and 2) *average note* topic membership, where we used the average note topic membership from the training data to predict the topic membership of each note in the test data.

Table 4.5 shows the aggregate prediction results of each model on the notes in the test data. Performance is broken down by notes with prior notes (n = 9290), and notes without prior notes (n = 1272). We evaluated statistical significance of the difference between the *average* performance of each model across all notes for *each patient*. We evaluated differences in model performance at the patient level, rather than at the note level, because notes belonging to the same patient do not satisfy the test assumption of independence. We used a paired t-test with a significance level of 0.001.

Using all prior notes (*notes2note*) and using structured data (*struct2note*) performed comparably well in predicting the next note (mean cosine similarity of 0.63, p = 0.006). However, *struct2note* outperformed *notes2note* on predicting the first note

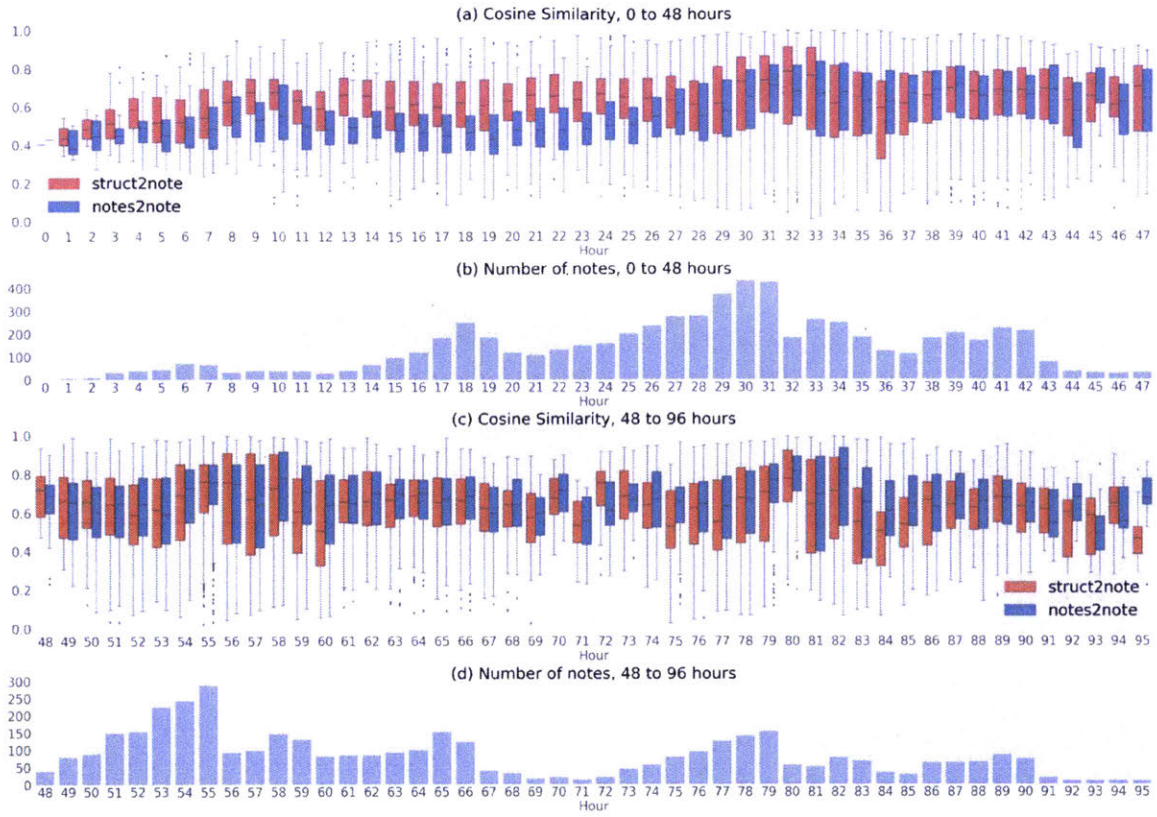


Figure 4-2: *struct2note*, *struct-notes2note* and *note2note* performance is shown in (a) (0 to 48 hours) and (c) (48 to 96 hours). Number of admissions with a note at each hour is shown in (b) (0 to 48 hrs) and (d) (48 to 96 hrs).

in the stay ($p < 1e-200$). In this case, *notes2note* performance was similar to taking the average note from the training data (cosine similarity of 0.41 vs. 0.42).

In addition, integrating the structured data and prior notes to predict the next note outperformed using either modality on its own (mean cosine similarity of 0.66 vs. 0.63, $p < 1e-46$). For notes without a prior note, the average cosine similarities of *struct2note* and *struct-notes2note* were similar (0.61), but *struct2note* significantly outperformed *struct-notes2note* ($p = 0.0003$).

Because notes have differing availability over time, we investigated the performance of these models on notes at different hours during the ICU stay. The differences in performance between the models utilizing structured data (*struct2note* and *struct-notes2note*) and the *notes2note* model are shown in Figure 4-2(a) and (c). The number of notes at each hour is shown in Figure 4-2(b) and (d).

In the early hours of the ICU stay (0 to 30), the structured data outperforms using the notes. Since there are very few notes available at this time, it is challenging for the *notes2note* model to make meaningful predictions. This performance improvement drops off around hour 30, or 6 a.m. on the second day of the patient's stay in the ICU. Recall from Figure 2-2 that physician notes are recorded regularly around 6 a.m. each day. At these times, the availability of notes grows, and predictive accuracy of the note prediction models increase. The improvement of using structured data rather than prior notes becomes marginal at later hours of the stay (48-96), when more notes are available.

4.6.2 Outcome Prediction

We evaluated the note predictions generated from the structured data alone (*struct2note*) by training supervised networks using 1) actual notes and 2) predicted notes for predicting in-hospital mortality and first intubation.

We evaluated performance in terms of the Area Under the Receiver Operating Characteristic Curve (AUC). We evaluated statistical significance by evaluating model performance on 100 bootstrapped sets for each model. A paired t-test was performed between the bootstrapped AUCs for a pair of models, at a significance level of 0.001. Bootstrapped samples were constructed so that the outcomes were represented in the same incidence as in the original test set. We also trained models using 1) static demographic characteristics such as age, gender, admission type, and first care unit and 2) structured data (events and physiological time-series) as performance baselines.

The results are shown in Figure 4-3. We show performance results at the last hour of each day (11 p.m.), when information from the course of the day can be taken into account. Our predicted note topic distributions performed comparably to the actual notes at hours 47 and 95 ($p = 0.78$ at hour 47 and $p = 0.46$ at hour 95). At hour 71, the difference in performance between the predicted note topics (AUC = 0.81) and the actual note topics (AUC = 0.83) was statistically significant ($p < 1e-5$), but not large. In addition, the predicted note topics significantly outperformed the actual ones at hour 23 ($p < 1e-50$).

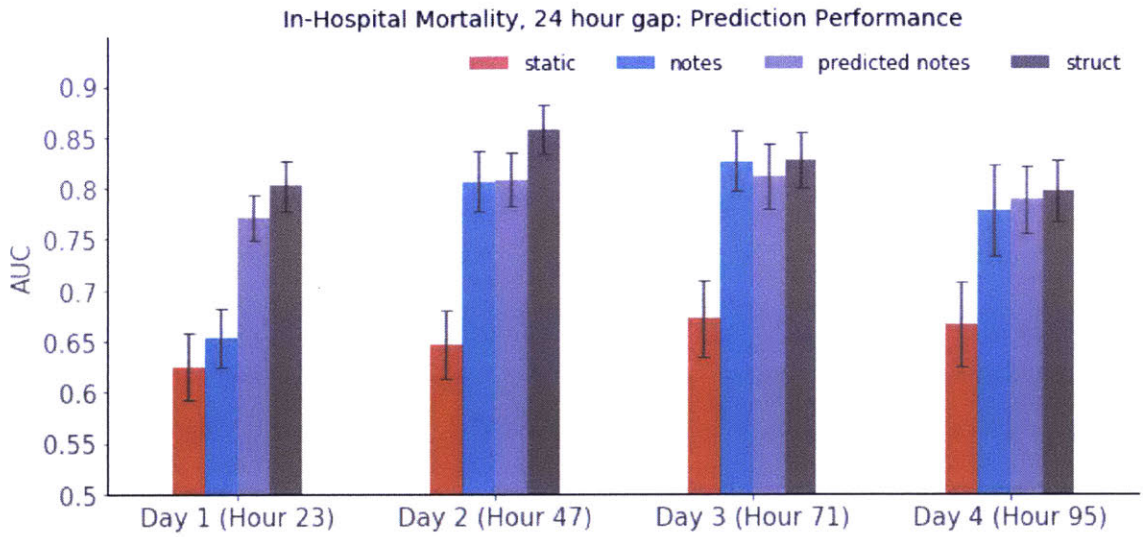


Figure 4-3: AUC using different data modalities to predict in-hospital mortality in the final hour of each day in the ICU (23, 47, 71, 95 hours). Error bars indicate standard deviations computed across 100 bootstrapped samples.

The results for first intubation prediction are shown in Figure 4-4. We evaluated a single AUC over all of the windows, rather than evaluating at specific times during the stay. First intubation is a much harder task than in-hospital mortality, as evidenced by the significantly lower AUCs.

On this task, the topic distributions of the actual clinical notes performed comparably to the static data ($p = 0.04$), and the predicted note topic distributions (AUC $= 0.66$) performed significantly better than the topic distributions of the actual notes (AUC $= 0.61$, $p < 1e-40$). The structured data (AUC $= 0.78$) significantly outperformed the predicted note topic distributions (AUC $= 0.66$, $p < 1e-55$. While the AUCs on this task are lower than the AUCs for in-hospital mortality, they similarly demonstrate that we are able to generate clinically meaningful note topics that result in predictive performance close to that of the true notes.

These performance results indicate that our method of learning correspondences between structured health record data and topic distributions of existing clinical notes allowed us to generate meaningful topics that capture changes in patient state. Importantly, although the predicted notes do not include *any* of the existing notes, they achieve predictive performance comparable to the topics of the actual notes in

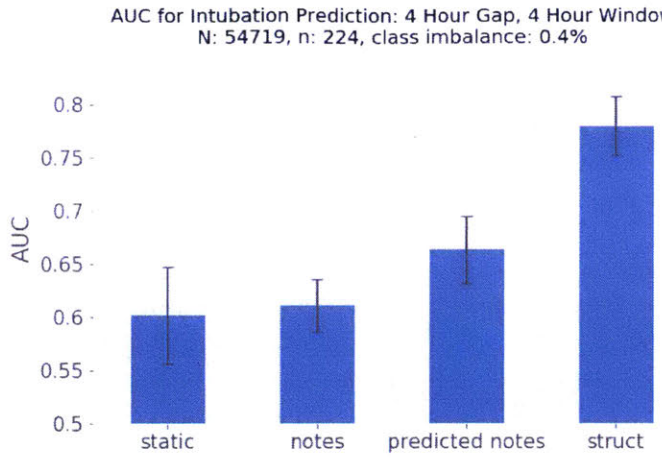


Figure 4-4: AUC using different data modalities to predict first intubation. A prediction is made at each hour to determine if after a gap period of 4 hours, the patient will be intubated in a 4 hour window. Error bars indicate standard deviations computed across 100 bootstrapped samples.

downstream prediction tasks.

Visualizing Correspondences

To qualitatively evaluate the learned correspondences,we identified individuals with high presence of certain topics and visualized structured data elements with meaningful relationships to those topics. Figure 4-5 shows the original topic distributions over time for topics corresponding to intubation or respiratory status (topics 25 and 36). This 88 year-old patient was admitted to the ICU shortly after 11 p.m. (hour 23). Her admission status was "emergency." She died in the hospital, 8 days after admission.

This patient was intubated shortly after ICU admission, at around 4 a.m. (hour 28). Whereas the original note only indicates a rise in corresponding topic membership around hour 31, our predicted note topics show an immediate rise in topic 36. This indicates that our predicted note topics are able to capture changes in patient state before the actual notes are recorded. This occurs again at hour 81, when the patient is extubated and then intubated again shortly after. While the predicted topics show an immediate rise in Topic 36, the note was not written until 8 hours later, at hour 89.

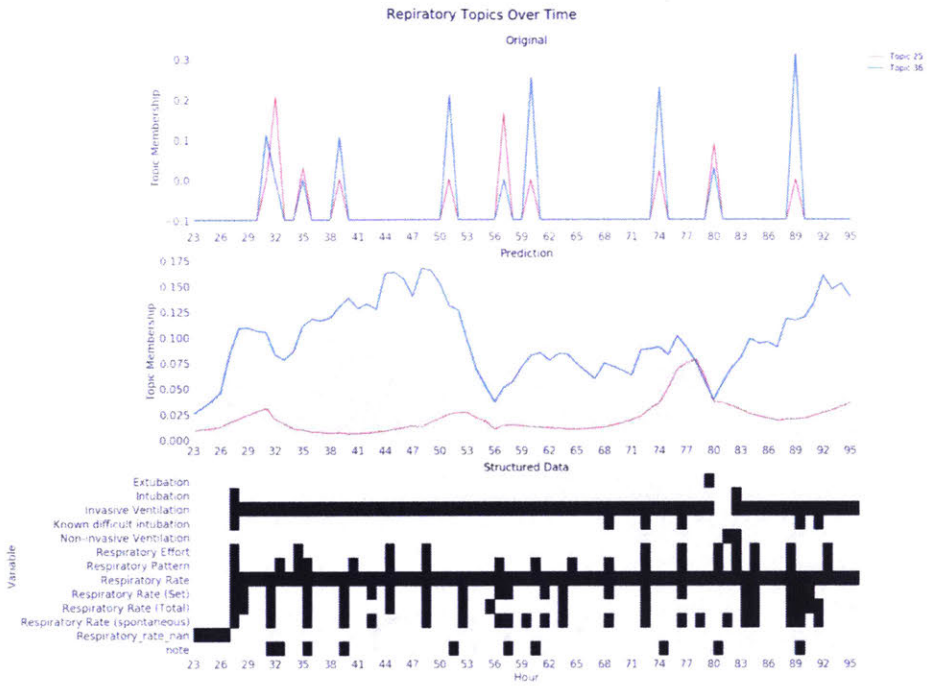


Figure 4-5: Correspondences between topic distributions of ground truth notes (top), predicted topic distributions (middle), and structured health record data (bottom) for a single admission. Topic membership values are shown as negative when no note was present. Topics corresponding to intubation and respiratory status (25 and 36) are shown, along with structured data elements pertaining to respiratory status and ventilation.

This example demonstrates how our method enables learning meaningful correspondences between the high dimensional structured EHR data and clinical summaries written during the course of care. In addition, we note that while our predicted topics did not always accurately represent the true topic distributions of notes (e.g., at hour 56), they still reflect meaningful correspondences with the structured data. This suggests that even if cosine similarity between the predicted note and the true note is low, our predictions might offer useful suggestions regarding topics that might be missing from the recorded notes.

Failure Cases

To better understand which notes our model predicted well and which it predicted poorly, we investigated the topic distributions of the true notes and the predicted topics both for patients we were able to predict well and for patients we were unable

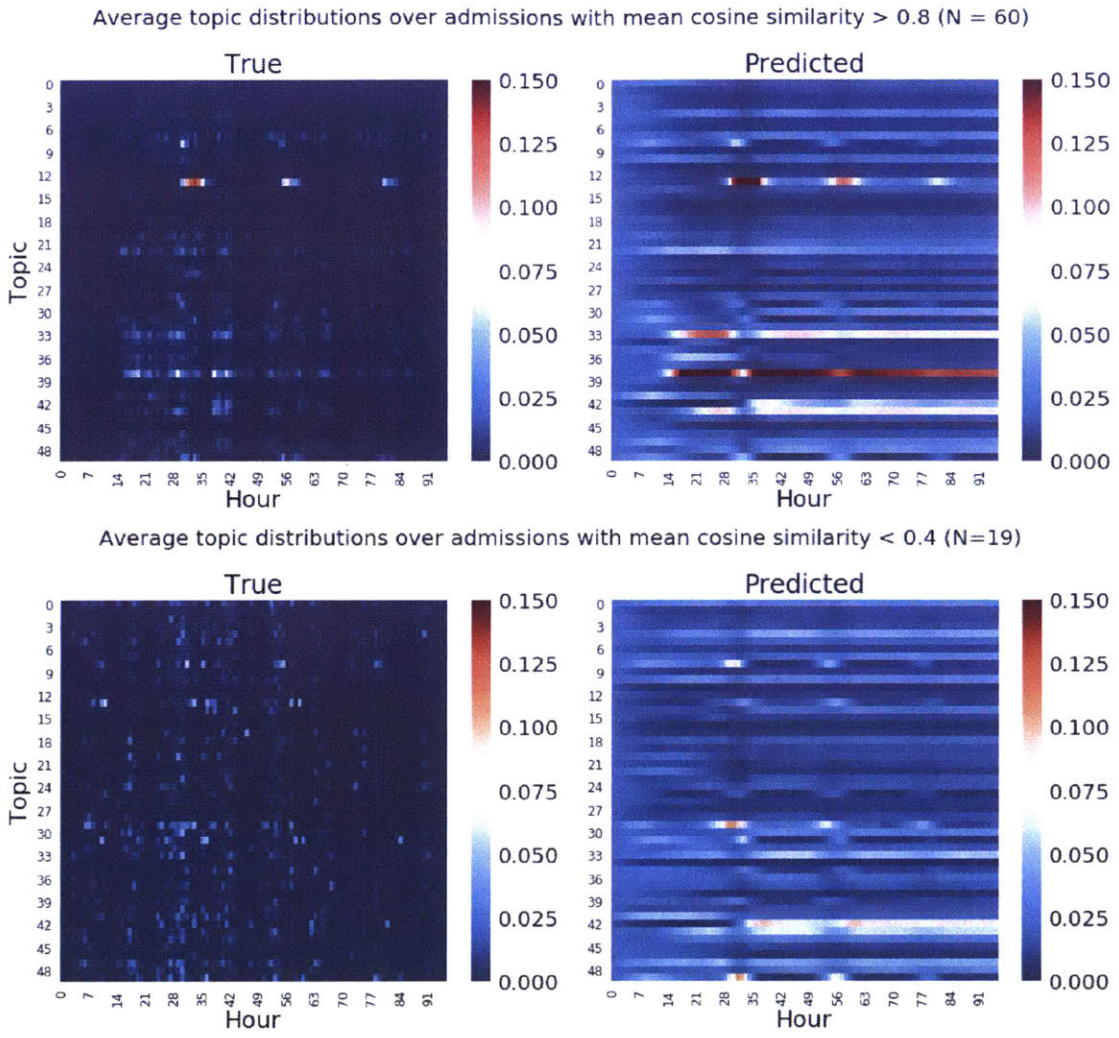


Figure 4-6: Topic distributions of patients with average cosine similarity > 0.8 (top), and patients with average cosine similarity < 0.4 (bottom).

to predict well. We considered performance for each admission by taking the mean performance across all notes corresponding to that patient. Patients with an average cosine similarity > 0.8 were considered good predictions, and patients with average cosine similarity < 0.4 were considered bad ones. Figure 4-6 shows heatmaps of the average topic distribution over time for these subsets of patients.

Patients with accurate predictions (top) had far more distinct topic distributions than patients with inaccurate predictions (bottom). In the top plot, there are fewer unique topics with high intensity. In the bottom plot, on the other hand, many topics are present with low intensity. In addition, the topics that are present propagate more

over time in the top figure than in the bottom, as indicated by the stronger horizontal bands in the top plot. Patients with predicted note topics closest to the topics in the actual notes consistently had high topic membership for topic 38 (cardiac surgery) and the topics 33 and 13 (routine vital signs monitoring). Of the 60 patients in this group, the majority (37) were admitted to the cardiac surgery recovery unit (CSRU).

Patients with inaccurate predictions had fewer clearly prominent topics over time. Topics 26 and 9, both of which indicate routine observation rather than clear patient phenotypes, had the highest topic membership at sporadic times during the stay. In addition, the overall distribution for patients that were difficult to predict was much flatter compared to the distribution for patients where our predictions were accurate. These patients were more evenly distributed among care units; ten patients were admitted to the surgical ICU, two to the trauma surgical ICU, three to the medical ICU, three to the cardiac surgery recovery unit, and one to the coronary care unit.

This analysis demonstrates that while notes with certain topic distributions are easy to predict well (e.g., notes for cardiac surgery patients), patients with more diverse conditions have notes that are more difficult to predict. We confirmed this hypothesis by breaking down cosine similarity performance on notes from patients in the different care units. These results are shown in Figure 4-7. The distribution of cosine similarity for notes from patients in the CSRU is more right-shifted towards 1 compared to the notes from patients in the other care units. On the other hand, patients in the MICU and SICU have less right-shifted distributions. Patients in these units are also more diverse in diagnosis compared to the CSRU (as shown in Figure 2-3).

In addition, different disease subtypes may not be equally well-represented. For example, topic 3 corresponds to cancer, topic 16 corresponds to liver disease, and topic 40 corresponds to renal failure and kidney disease. If small numbers of patients exist in the data with these conditions, it would be difficult to accurately learn correspondences between the structured data and the predicted topics.

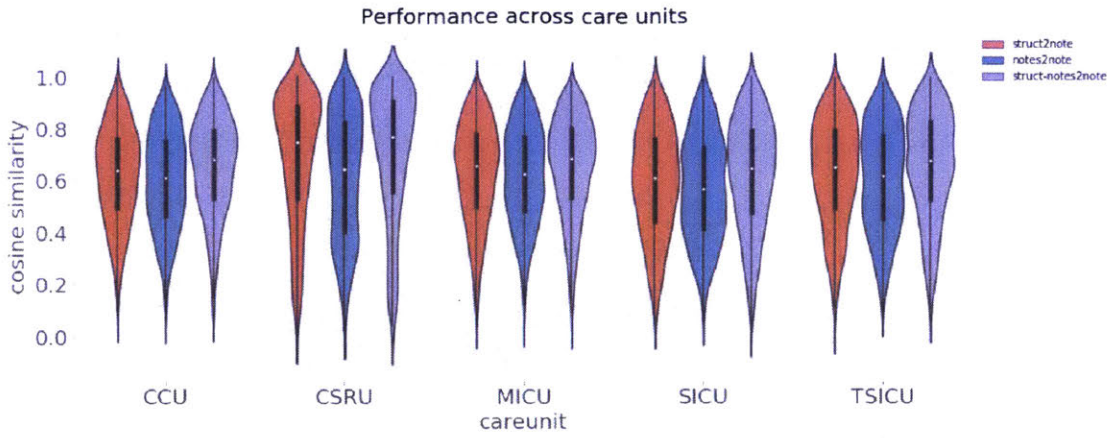


Figure 4-7: Cosine similarity distribution for notes from patients in different care units.

4.7 Summary & Discussion

In this work, we proposed a method to *learn* to generate meaningful topic summaries from structured patient health record data. We used existing summaries written by clinical care team members to learn correspondences between structured health record data and the topics underlying clinical notes. We demonstrated that using structured data alone, we are able to generate note topics with an average cosine similarity to actual notes of 0.63, comparable to the performance of using prior notes alone. Integrating structured data with prior notes results in an average cosine similarity of 0.66. Using the structured data, we are also able to generate the first note in the stay with an average cosine similarity of 0.61.

We also demonstrate that our generated topics are able to predict clinical outcomes such as in-hospital mortality with comparable performance to topic distributions of actual notes written by care team members. We additionally present qualitative evaluations of correspondences between structured data elements and changes in topic distribution.

Inherent to our approach is an assumption that clinical notes are *good* summaries. We believe this is usually a reasonable assumption because notes are used at the point of care for this purpose. However, clinical notes, particularly in electronic systems, have been shown to contain redundancies and incorporate outdated information.

Our approach is a first step towards the generation of clinical text summarizing structured health record data. Generating topic distributions could be useful in proposing potentially missing topics to care staff while they are writing a note. Future work could include generating candidate phrases corresponding to patient history. In addition, while our analysis is limited to the intensive care setting and to the structure and notes in MIMIC, our approach could similarly be used to generate topics summarizing longitudinal health record data in outpatient settings.

Chapter 5

Characterizing Clinical Care Pathways in Critical Care Settings

5.1 Introduction

In this chapter, we return to the question of distinguishing care actions from observations of patient state. Care actions are taken based on established *care processes*; e.g., if a patient's blood pressure is high, blood pressure medication might be administered. If a patient is unable to breathe on her own, she might be intubated. In this chapter, we define care process as *actions* that are associated with *observations* of patient state.

We explore how to characterize typical *care processes* in the intensive care unit. These actions include medications, treatments, procedures, and ordered tests. Accurately characterizing associations between care patterns and observations of patient state is an important and challenging task. These learned associations can provide insight into how physicians typically make decisions. By *learning* what actions are performed conditioned on patient attributes, we can capture differences in how doctors act on similar signs and symptoms.

In the prior two chapters, we demonstrated that care events are important predictors of adverse outcomes, and that they capture a great deal of information about patient state and clinical care actions that can be used to generate meaningful sum-

maries. In these works, we treated events as features, rather than as labels. In addition, we did not distinguish between events that captured observations of patient state and events that captured actions taken by care providers (e.g., medication or treatment administered).

In this chapter, we present a preliminary formulation of separating actions from observations. We utilize *observations* as *features*, and *actions* as *labels*. We are interested in learning correspondences between observations of patient state and actions taken by care team members.

As a case study, we consider lab tests, antibiotics, and imaging tests. These actions are diverse in when they occur during the course of the stay, the patient populations in which they occur, and the type of event. Lab tests are often routinely done during the first portion of a patient ICU stay, and then irregularly thereafter. Antibiotics can be administered prophylactically, but also to treat infections once they are in progress. Imaging tests are diverse (e.g., chest X-ray versus CT scan), and different tests may be done for different patient populations.

These sets of events can occur as a routine part of care, but also capture physician intuition or knowledge about what certain observations of patient state may indicate.

5.2 Related Work

5.2.1 Capturing Patterns in Care

Observational health record data capture complex interactions between patient condition and care patterns. Disentangling these sources is important to accurately assessing either one [78].

A great deal of work has investigated patterns in sequences of physician orders. [79] summarizes event sequences by identifying frequent medical behavior patterns. [80] investigates common treatment pathways in different patient subsets. These works focus on patterns in care actions. However, they do not connect care actions with other patient attributes. In contrast, our goal is to characterize care actions by

learning a correspondence between observations of patient state and the care actions that are taken.

[81] use latent Dirichlet allocation to model clinical orders. The authors use data from the first 24 hours of the patient stay, and ignore the sequential nature of the events. In addition, the authors combine orders, such as medications, lab tests, and imaging tests, with observations (e.g., abnormal lab test results, problem list entries, and diagnosis codes). In contrast, our approach views observations and orders as distinct, and we consider the sequential nature of the data.

[82] proposes a probabilistic topic model that captures relationships between patient features and treatment patterns to learn latent treatment patterns. The authors demonstrate that the learned topics capture meaningful topics describing treatment patterns, and that different treatment patterns can correspond to different patient attributes. However, the authors use only patient attributes available on admission. In contrast, our approach leverages observations through the course of the patient stay to learn correspondences with care actions.

[83] investigates correlations between information about patient condition (e.g., lab tests, concepts in the clinical notes identifying symptoms, medications, etc.) and events that the authors define as part of the healthcare process: inpatient admission, inpatient discharge, outpatient visit, emergency department visit, and ambulatory surgery. Using data from a large clinical data warehouse that includes both outpatient and inpatient visits, the authors show that the variables from the EHR cluster differently for different health care process events. Their study is limited to a small set of health care process events and EHR variables.

[84] characterizes physician behavior using the time between an initial lab test order and another order of the same test for each patient. They demonstrate that analyzing lab tests in the context of *timing* is one way of capturing physician expertise, or "group intelligence." The authors show that high time-to-repeat correlates with lab test values in the reference range for a "normal" result. In addition, the authors demonstrate that physician behavior (in ordering lab tests) differs across different care settings (e.g., inpatient vs. outpatient), and that an understanding of "group

intelligence" can be used to identify situations where a lab test might be unnecessary.

Similar to these works, our goal is to characterize patterns in care. Our definition of care patterns centers around actions taken by care team members in response to observations; in other words, the clinical decision-making process through the care provider's lens. We utilize a supervised learning approach to learn correspondences between observations of patient state and care actions. In addition, we focus on the critical care inpatient setting.

5.2.2 Learning to Predict an Intervention

Prior works have sought to predict intervention administration from patient physiology [20, 21, 22]. These works are pertinent to the work discussed in this chapter because both relate observations of patient state to actions. However, there are key differences in the set of events we consider as actions, and in how we frame the learning problem.

[20] uses unsupervised methods to learn latent physiological states in patients who were administered vasopressors. The authors investigate several intervention onset and weaning prediction tasks and demonstrate that using latent representations of physiological state improved predictive performance. [21] extends this framework by learning unsupervised representations of physiological state for all patients (rather than using only patients who were administered vasopressors). The authors demonstrate the generalizability of these representations to predictive models for the onset of different interventions, including intubation, vasopressor administration, and transfusions [21].

Finally, [22] considers the use of recurrent neural networks and convolutional neural networks to learn latent representations of physiological state in the context of the outcome of interest. The authors consider a multi-class formulation by predicting whether the intervention was started, ended, or maintained. They evaluate their method on a variety of interventions, including ventilation, vasopressors, and colloid boluses.

These works differ from ours in several aspects. First, these works develop models

that seek to make predictions about *future* outcomes. They do this by enforcing forward-facing predictions of interventions, sometimes with a gap period between the information being used to make a prediction and the outcome that is being predicted. In contrast, we seek to learn *correspondences* between observations of patient state and the care process, regardless of whether care actions occurred before or after the observations. While our model captures associations between observations and actions that are *close* in time, we do not enforce any temporally causal relationship. Secondly, we are interested in all care actions, which subsume the interventions considered in these works. Instead of targeting need for an intervention from a *patient* perspective, we characterize *clinician* response to observations of patient state.

5.3 Methods

5.3.1 Data Processing

As in the last chapter, all data were aligned to midnight on the day of ICU admission and discretized to the hour. All admissions were padded or truncated to 96 hours from midnight of the first day of ICU admission.

5.3.2 Cohort

We used data from the first ICU stay corresponding to each patient in the MetaVision portion of the MIMIC-III dataset. We utilized all patients who had both events data and physiological time-series. This is a larger subset of patients than the cohort used in Chapter 4. The data were divided into a 60/20/20 training/validation/test split. We filtered the cohort to consider only patients with at least a 36-hour stay in the ICU. This was to ensure that patients were in the ICU for a sufficient period of time to capture care patterns. The final training, validation, and test cohorts are described in Table 5.1.

Table 5.1: Training, Validation, and Test splits. Distributions of patients are similar in terms of % in-hospital mortality, length of stay, care unit, and admission type.

	Train	Validation	Test
N	5605	1841	1869
In-Hospital Mortality (%)	10.1	9.5	11.4
Mean (std) LOS in ICU (days)	3.5 (2.0)	3.5 (2.0)	3.5 (1.9)
	Care Unit		
CCU (%)	13.0	14.0	13.3
CSRU (%)	14.5	15.7	15.0
MICU (%)	39.1	38.1	39.3
SICU (%)	19.2	18.9	19.3
TSICU (%)	14.1	13.3	13.2
	Admission Type		
Elective (%)	13.3	14.3	12.9
Emergency/Urgent (%)	86.7	85.7	87.1

5.3.3 Categorizing actions and observations

We first separated categories of actions from categories of observations. In MIMIC-III, recorded items are associated with category labels. We used these labels to distinguish *actions* from *observations*. Table 5.2 details the categories identified as actions versus categories identified as observations. These definitions of "observation" versus "action" are based on item category rather than on specific item definitions.

All events that belong in the observations categories were used as input features. In addition, we considered static features on ICU admission, including patient age, gender, ethnicity, admission type (i.e., elective, emergency, or urgent) and first care unit. We considered three categories of events described as "actions:" 1) lab tests, 2) antibiotic administration, and 3) imaging tests.

These three action categories capture events that are dependent on patient condition, and vary in whether they indicate a routine care process or a decision made by a clinician because there was cause for concern. As we will show, while some actions depend on patient attributes (e.g., antibiotics are administered differently in different care units), others are performed regardless of patient attributes (e.g., glucose lab tests are done for all patients).

Table 5.2: Categories of events identified as actions and observations.

Category	Events
Actions	Intubation/Extubation, Ventilation, Cultures, Antibiotics, Pain/Sedation, Medications, Access Lines - Invasive, Access Lines - Peripheral, Blood Products/Colloids, Dialysis, IABP, Impella, Drains, Labs, Imaging, Service Changes, Microbiology Tests
Observations	General, ADT, Adm History/FHPA, Alarms, Nutrition - Enteral, Nutrition - Parenteral, Blood Gas, Chemistry, Hemodynamics, Toxicology, Significant Events, OB-GYN, OT Notes, Cardiovascular, Cardiovascular (pulses), Output, Fluids/Intake, GI/GU, Pulmonary, Respiratory, Restraint/Support Systems, Routine Vital Signs, Skin - Assessment, Skin - Impairment, Skin - Incisions, Neurological

5.3.4 Data Representation

We represent both observations and actions as tensors over time. As in the previous chapter, we binarized the tensors to capture whether or not an event occurred at that hour, rather than the number of times it occurred.

Blurring Actions Over Time

Our goal in this work is to learn correspondences between observations and actions. However, actions recorded in the EHR are noisy records of the true actions taken. For example, if an intervention is noted at a particular point in time, it may be that that intervention could have been administered earlier or later without changing the effect on patient state. Because of this, works such as [20], [51], and [22] consider an intervention over a window of multiple hours, rather than at a single point in time. By considering the presence or absence of actions only at the time they are recorded in the EHR, we would miss correspondences in the learning process that are meaningful. Thus, we blur the action labels over time. We consider the action as having a positive label within 2 hours before or 2 hours after the hour recorded in the data.

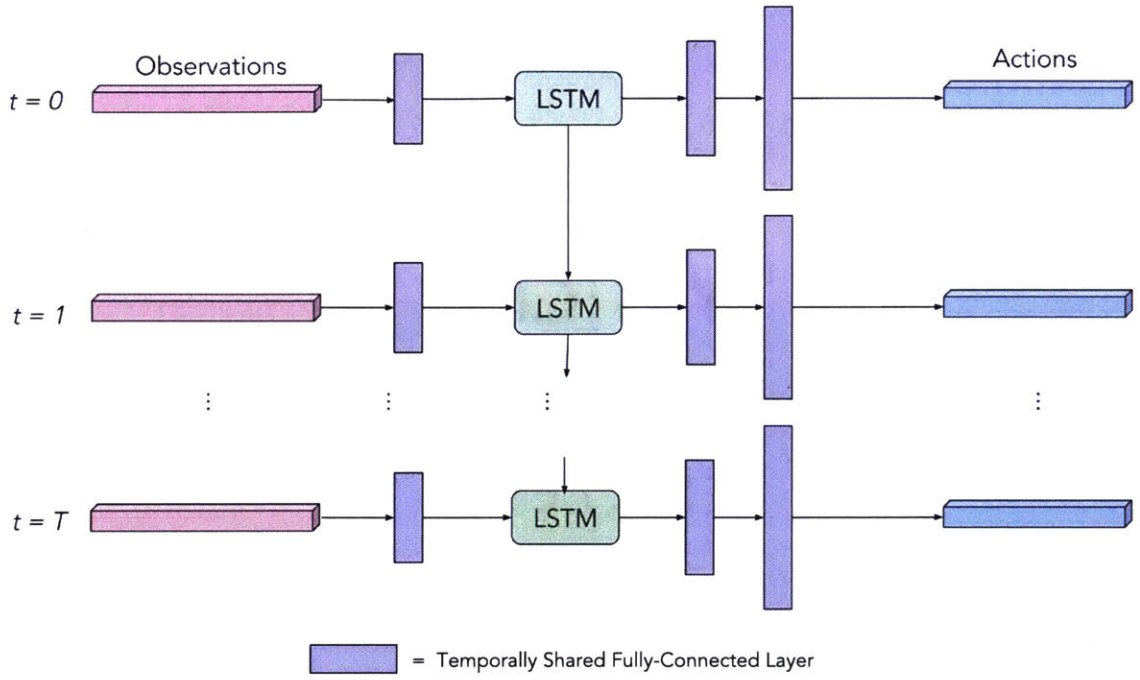


Figure 5-1: Model architecture for learning correspondences between observations and actions. The network is shown unrolled over time. An initial temporally shared fully-connected layer with a rectified linear activation function is used to embed the high-dimensional observations into low-dimensional dense embeddings. An LSTM layer is then used to encode relationships over time. Finally, two more temporally shared fully connected layers are used to relate the encoded observations to the action space. The first has a rectified linear activation function, and the second has a sigmoid activation function.

5.3.5 Learning Associations Between Actions and Observations

Network Architecture

The network architecture we use is shown in Figure 5-1. Similar to our network in Chapter 4, we first use a temporally shared fully-connected layer to embed the input observations at each time step. Next, an LSTM layer is used to capture relationships over time. Finally, two fully-connected layers are used to relate the encoded observations to the space of actions. We trained a separate model for each action class we considered.

Loss Function

We use binary cross-entropy as our loss function. Because of the high class imbalance (most actions are infrequently performed), we use an asymmetric cost parameter to weight labels where actions were present ($y = 1$) as more important.

Evaluation Metric

To evaluate the performance of our model, we use mean-squared error over the predicted actions at all time steps. Because of the high class imbalance, we consider the mean-squared error over actions that were *present* ($y = 1$), the mean-squared error over actions that were *absent* ($y = 0$), as well as the total error. For a given patient i and a given action class,

$$MSE^i_{\text{present}} = \frac{\sum_{t=1}^{T_i}\sum_{k=1}^{K}(\hat{y}^i_{t,k} - y^i_{t,k})^2 y^i_{t,k}}{\sum_{t=1}^{T_i}\sum_{k=1}^{K} y^i_{t,k}}, \tag{5.1}$$

$$MSE^i_{\text{absent}} = \frac{\sum_{t=1}^{T_i}\sum_{k=1}^{K}(\hat{y}^i_{t,k} - y^i_{t,k})^2 (1 - y^i_{t,k})}{\sum_{t=1}^{T_i}\sum_{k=1}^{K}(1 - y^i_{t,k})}, \tag{5.2}$$

$$MSE^i_{\text{total}} = \frac{\sum_{t=1}^{T_i}\sum_{k=1}^{K}(\hat{y}^i_{t,k} - y^i_{t,k})^2}{T_i K}, \tag{5.3}$$

where K is the number of actions in the action class and T_i is the number of time steps in patient i's stay.

This evaluation metric emphasizes how well we are able to predict actions that occurred for each *patient*. In contrast, an evaluation metric such as the Area Under the Receiver Operating Characteristic Curve (AUC), which is typically used to evaluate performance on binary classification tasks, would evaluate performance for each *action*. We use MSE rather than AUC because we are interested in evaluating our ability to predict *patterns* of care for each patient over time and across all items within each action class.

Another consideration is that the AUC evaluates performance on each action independently. However, items within an action class (e.g., antibiotics) may not be independent. Evaluating discriminative performance for each individual action

may not capture relationships across items within the same class. For example, an interesting care pattern such as the administration of a different antibiotic within hours of the administration of a broad-spectrum antibiotic would not be captured when evaluating performance on a single action.

Training and Implementation

We implemented our models using Keras 2.1.3 with Tensorflow backend (1.5.0) [77]. The size of the first temporally shared fully-connected layer for embedding observations was set to 50 units. The LSTM layer and the following fully-connected layer were set to 128 units. Models were trained with a batch size of 32, until the validation loss converged.

We implemented the asymmetric cost by weighting the loss for positively labeled examples by a parameter λ. A grid search over $\lambda = 1$, 5, 10, 20, and 100 was performed. Figure 5-2 shows the results of these grid searches for each action class on the validation set. Performance is broken down by error on the present actions (left), absent actions (middle), and all (right). These figures show that the value of the asymmetric cost parameter dramatically affects the ability of the model to capture present actions in addition to absent ones. In addition, the curves for each action class exhibit an elbow at $\lambda = 10$. Although the error for present actions continues to drop as λ increases, there is a trade-off with the error for absent actions. Because actions are rare, the error on absent actions (middle) dominates the overall error (right).

Baseline Comparisons

As a baseline comparison, we use *incidence* of care events in the population over the course of the patient stay. Incidence is a naive way to capture care patterns; it characterizes regularity in care, agnostic to observations of patient state.

We compute incidence as the proportion of patients for whom an action occurred at each hour. We consider incidence in 1) the entire ICU population, and 2) in each care unit. We compute incidence values using only the training data.

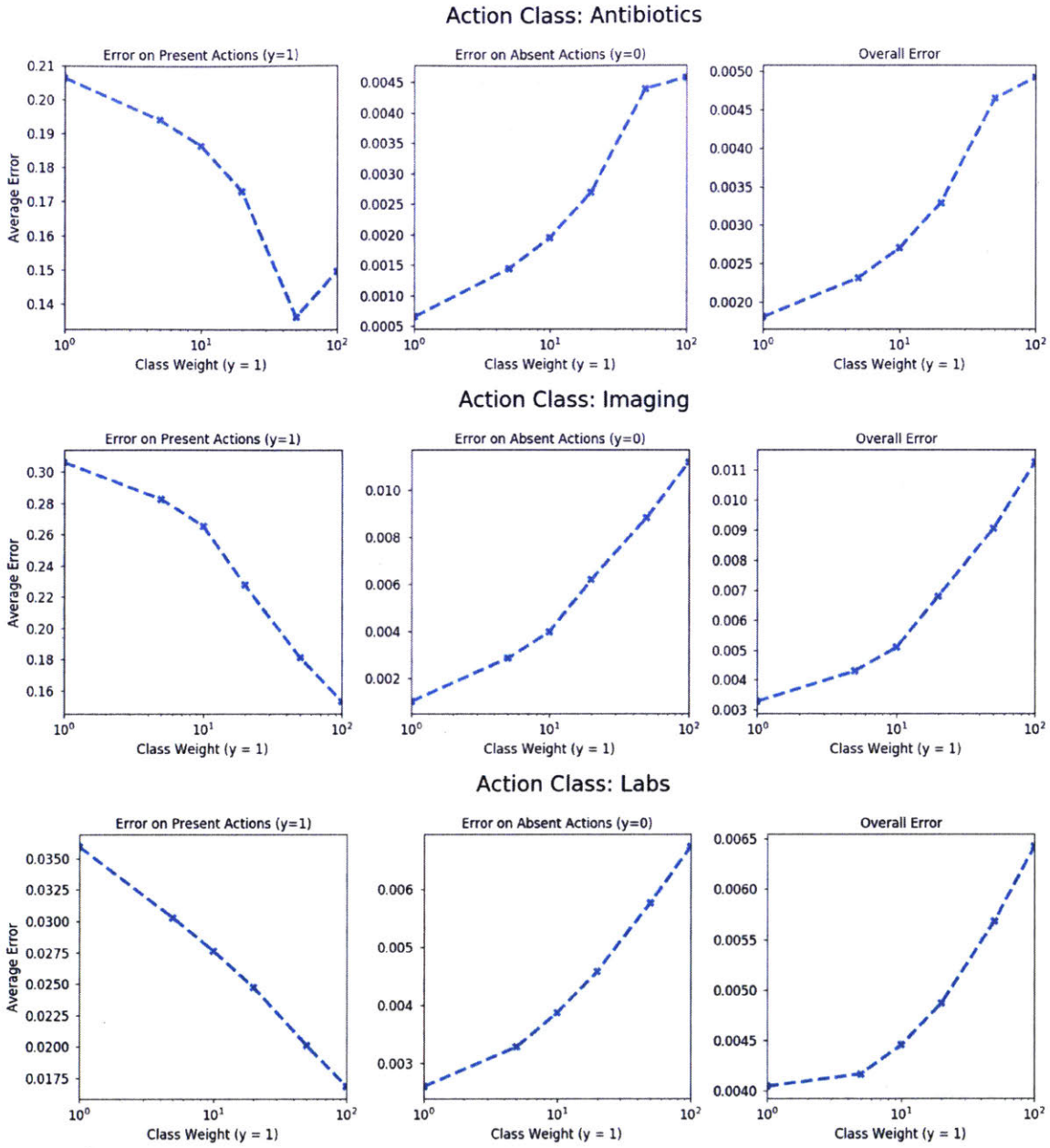


Figure 5-2: Mean error for different action classes: antibiotics (top), imaging (middle), and labs (bottom) over a range of asymmetric class cost parameters (1, 5, 10, 20, 50, 100). Performance error for different classes are shown ($y = 1$: left, $y = 0$: middle, all: right). The best asymmetric cost parameter was chosen based on the elbow in the curves (around 10).

5.4 Results

5.4.1 Care patterns differ across care units.

Figures 5-3–5-5 show the incidence of actions in different classes in each care unit and in the entire ICU population. Different lab tests and imaging tests are performed for patients in different care units, and antibiotic administration also differs by care unit.

For example, Figure 5-3 shows that antibiotics such as vancomycin and cefazolin are administered in many patients in the CSRU. Whereas vancomycin is also administered in the other care units, cefazolin is rarely administered in either the MICU or the CCU. In addition, patients in the MICU (green bars) and the CCU (blue bars) receive more diverse antibiotics compared to patients in the CSRU (orange bars).

Imaging tests also show differences across care units. Chest X-rays are frequently performed in the CSRU, whereas CT scans are not. In contrast, patients admitted to the TSICU receive a variety of different imaging tests, including chest X-rays, CT scans, and MRIs. Some imaging modalities are very rarely done; for example, angiographies, which can be invasive, are performed in only a small subset of patients. When they are performed, they usually occur for patients in the SICU.

Lab tests demonstrate many fewer differences across care units compared to the other two action classes. Chemistry lab tests, such as glucose, chloride, sodium, etc. are performed in all patients in all care units. However, some lab tests, such as blood gas tests, are performed more often in patients in the CSRU. Others, such as bilirubin and direct bilirubin (tests for assessing liver function), are performed much less frequently. Thus, while some lab tests are done regardless of patient attributes, others depend on patient state.

5.4.2 Model predictions outperform baseline incidence in capturing actions that are taken.

Table 5.3 describes the average error of our model predictions compared to the true event occurrence over all patients in the test set. Performance is broken down by

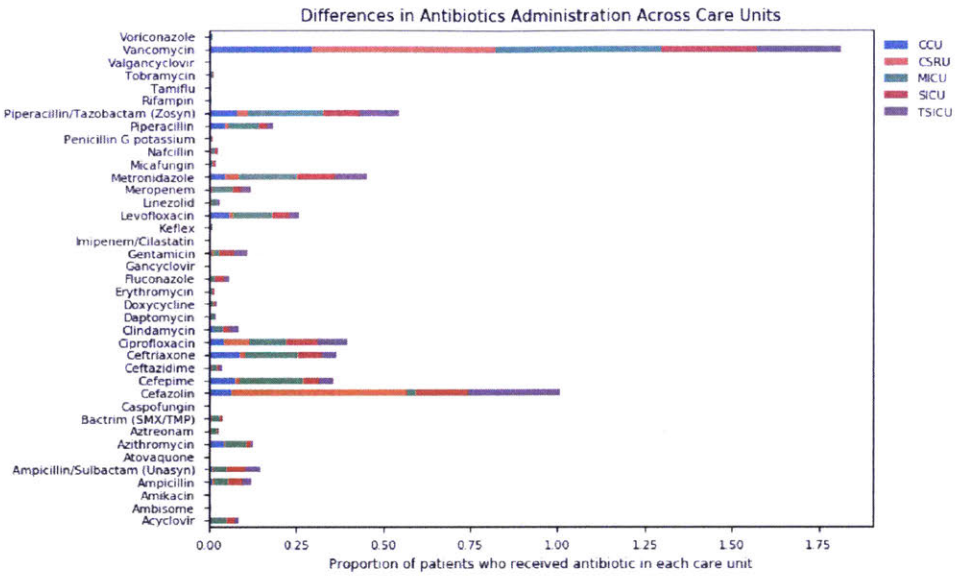


Figure 5-3: Incidence of antibiotics administration in different care units. Antibiotics such as vancomycin and cefazolin are given to many more patients than the other antibiotics. In addition, cefazolin is administered more frequently in the CSRU compared to other units. More diverse antibiotics are administered in the MICU and the CCU compared to the CSRU. This is evidenced by more bars with green (MICU) regions, compared to few bars with orange (CSRU) regions.

actions that were present ($y = 1$) and actions that were absent ($y = 0$). We additionally evaluated the aggregate error. Our model had the lowest error on events that were present. We are able to outperform baseline incidence measures in predicting actions taken on a patient, using only static characteristics and observations. This indicates that our model can learn meaningful correspondences between observations and actions. This trend holds for all three action classes that we considered.

Our model did not perform as well as the baselines on absent actions. And, because of the high class imbalance, the overall error is dominated by the error on absent actions. In general all models were able to predict absent actions much better than they were able to predict present actions.

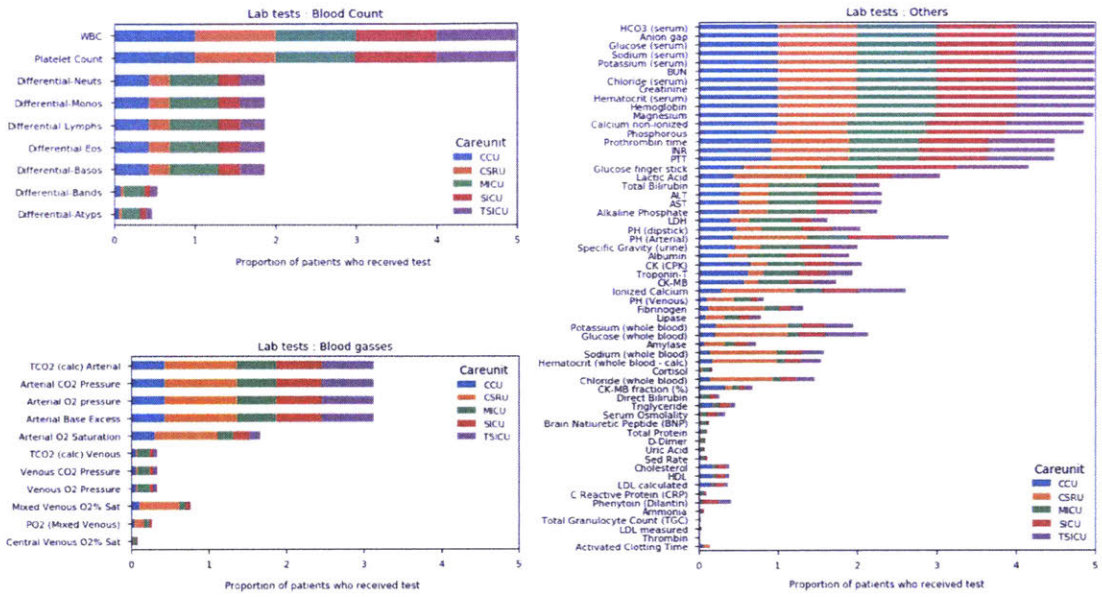


Figure 5-4: Incidence of lab tests in patients from each care unit. Tests are split into blood count tests (top left), blood gas tests (bottom left), and others (right). Many lab tests are performed in all patients (e.g., white blood cell (WBC) count, platelet count, anion gap, chloride, etc.). Some tests are performed in specific patient populations; for example, some blood gas tests are performed more frequently in the CSRU, and some chemistry tests (e.g., cholesterol, HDL) are performed in the CCU more than in the other units.

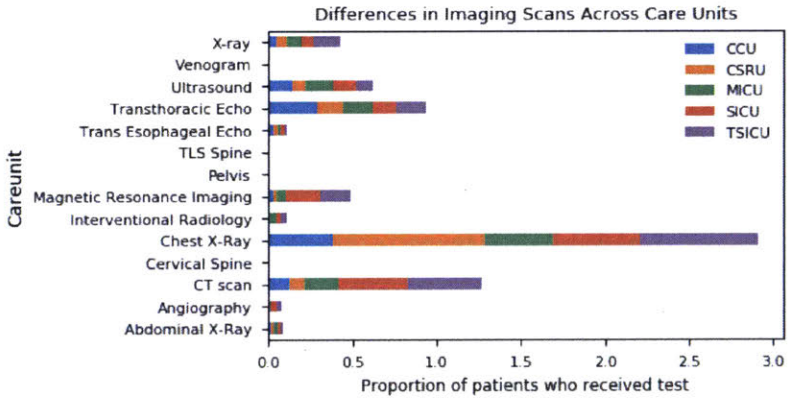


Figure 5-5: Incidence of imaging tests in patients from each care unit. Chest X-rays are more frequently done compared to other imaging modalities. CSRU patients primarily receive Chest X-rays, whereas patients in the other units receive more diverse imaging tests.

Table 5.3: Error on actions that were present ($y = 1$) and absent ($y = 0$) for different action classes. Because of the high class imbalance, the contribution to the total error of actions that are taken is only a small fraction of the overall error. All models (Ours) were trained with an asymmetric cost parameter, weighting errors in present actions by 10.

		Antibiotics		**Imaging**		**Labs**	
		Mean	Std	Mean	Std	Mean	Std
Present	Ours	**0.1866**	0.1007	**0.2696**	0.1027	**0.0278**	0.0085
	Baseline, All	0.2036	0.0970	0.3051	0.0926	0.0329	0.0070
	Baseline, Care units	0.1979	0.0963	0.2957	0.0949	0.0315	0.0075
Absent	Ours	0.0020	0.0017	0.0039	0.0022	0.0038	0.0014
	Baseline, All	**0.0004**	0.0001	**0.0007**	0.0002	**0.0015**	0.0004
	Baseline, Care units	**0.0004**	0.0002	0.0008	0.0004	**0.0015**	0.0005
All	Ours	0.0028	0.0015	0.0050	0.0020	0.0044	0.0011
	Baseline, All	**0.0016**	0.0011	**0.0030**	0.0016	0.0032	0.0008
	Baseline, Care units	**0.0016**	0.0011	**0.0030**	0.0016	**0.0031**	0.0008

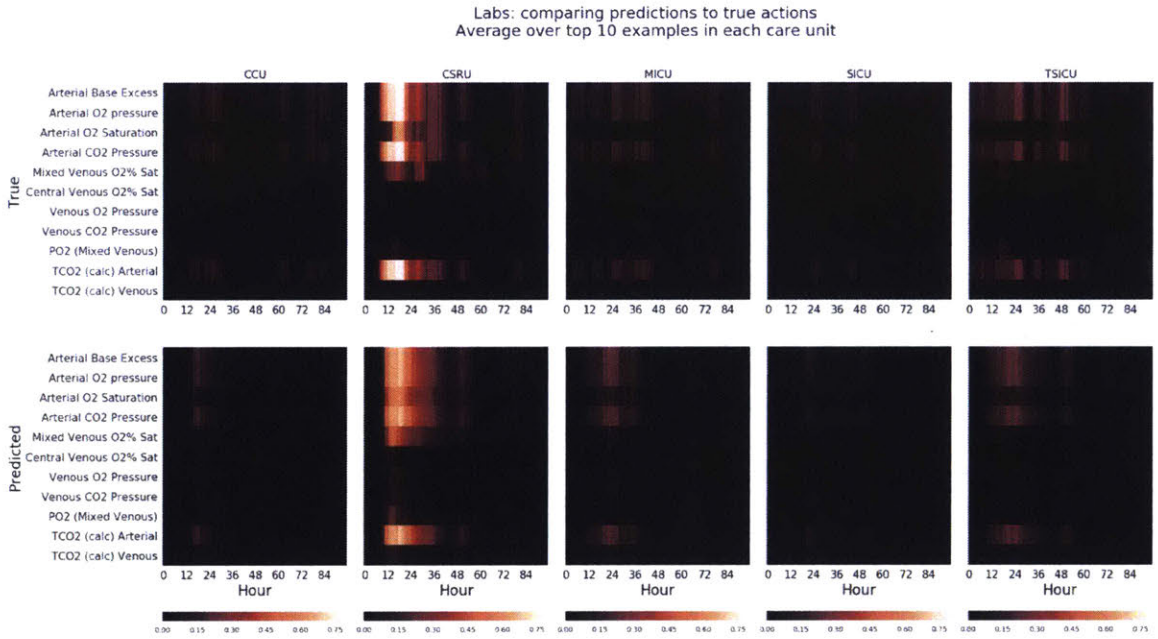


Figure 5-6: Blood gas lab tests: mean over best 10 examples by overall error. True actions (top) and model predictions (bottom) are shown

5.4.3 Model predictions capture differences in actions across care units and over time.

In the previous section, we presented quantitative measures of the ability of our models to capture care patterns for different action classes. In this section, we present qualitative analyses of examples from each action class where our models performed

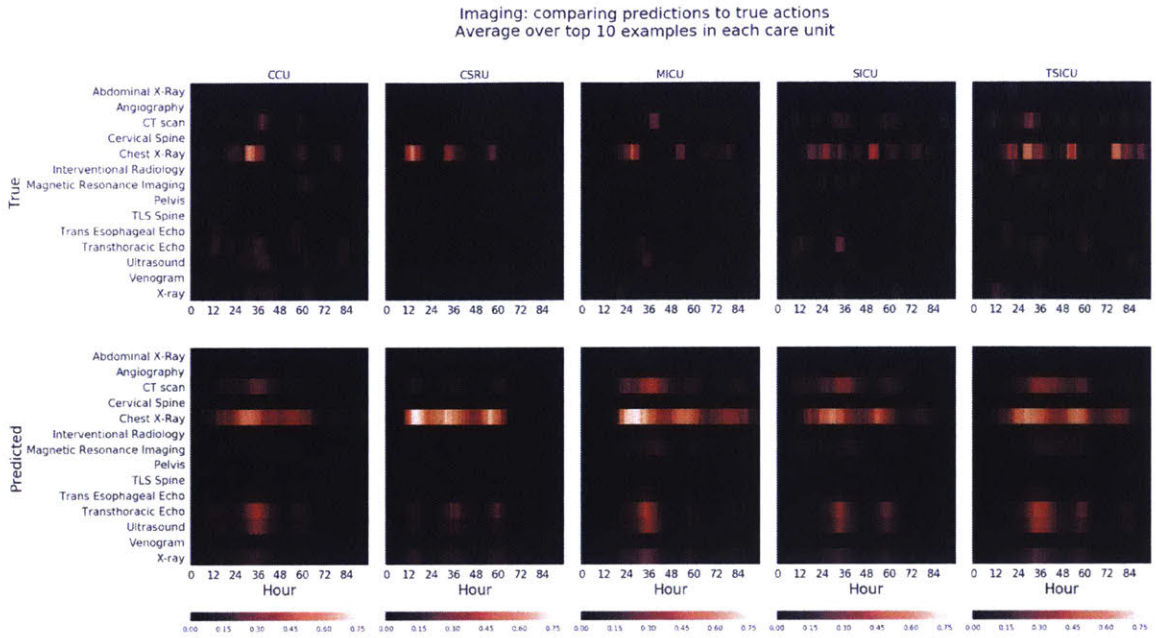


Figure 5-7: Imaging tests: mean over best 10 examples by error on present actions. True actions (top) and model predictions (bottom) are shown.

well. This analysis highlights examples of meaningful correspondences that were learned.

Figure 5-6 shows heatmaps for blood gas lab tests, depicting true action presence (top row) and average predictions (bottom row) for the ten patients with the lowest overall error from each care unit. These examples highlight the characteristics of patients for whom our model was able to accurately capture care patterns.

Our model predictions demonstrate that our approach of relating patient attributes to care patterns can effectively discover differences in practice across care units and over time. For example, from Figure 5-4, we know that blood gas tests are performed more frequently for patients in the CSRU compared to other care units. This is also evidenced by the true incidence of blood gas tests in the ten patients with lowest error (top panel). Our model's predictions (bottom panel) capture this trend well; whereas all other care units primarily have low-intensity regions, predictions for patients in the CSRU have high-intensity regions in the first day of the ICU stay for many arterial blood gas tests.

More generally, the images of our predictions (bottom panel) correspond to the

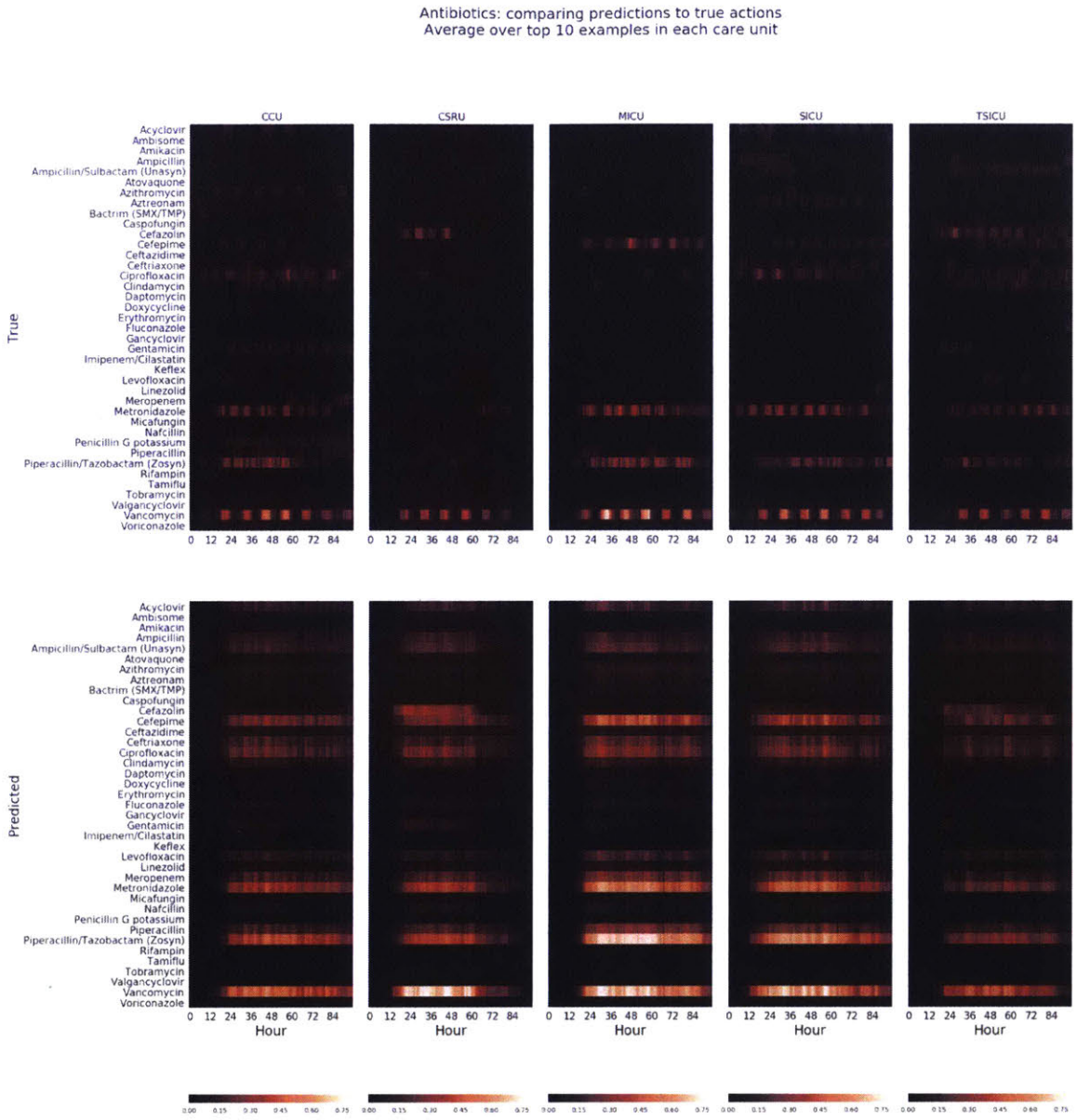


Figure 5-8: Antibiotics: mean over top 10 examples by error on present actions. True actions (top) and model predictions (bottom) are shown.

images of the true data (top panel) for all care units, both in terms of temporal trends and in terms of relationships between actions.

Figure 5-7 and Figure 5-8 show similar plots for imaging tests and antibiotic administration, respectively. These plots were constructed by taking the top 10 examples based on error over actions that were taken ($y = 1$), rather than overall error.

Figure 5-7 demonstrates that our model is able to capture shared attributes across

care units (e.g., chest X-rays performed at regular temporal intervals). In addition, it can capture unit-specific characteristics (e.g., brighter regions in the CT Scan are present in our predictions for the MICU, SICU, and TSICU, but not for the CSRU). This corresponds to a similar trend in the true data. Thus, our model is able to learn interactions between static characteristics (e.g., care unit on admission) with observations of patient state in order to predict corresponding care actions.

Figure 5-8 demonstrates that our model can capture unit-specific characteristics. For example, in the CSRU and in the TSICU, cefazolin is administered early in the first day. This is not true for the other care units. Our model predictions demonstrate that it is able to learn this association; in the predictions for the CSRU and TSICU, there is a horizontal band corresponding to cefazolin, and this band is not present in the other care units.

5.5 Discussion

In this chapter, we presented a preliminary formulation of learning correspondences between observations of patient state and care provider actions. We demonstrate how this correspondence can be learned using patient data in the ICU. We believe this direction holds a great deal of promise. The modeling approach we explore attempts to capture physician behavior. There are many questions we can ask about physician behavior given observations of patient state. For example, what actions are expected based on the observations? Does a given provider's actions deviate from the usual pattern of care, conditioned on observations of a particular patient?

There are many challenges to ensuring the correspondence that is learned captures meaningful relationships between observations and actions. We have presented a preliminary formulation, where we ensure that events that are categorized as observations and events that are categorized as actions are mutually exclusive. However, more specifically defining the nature of an observation and an action are necessary to accurately characterize how observations drive actions. We have excluded observations from numerical values (e.g., lab test results, recorded vital signs values)

and clinical notes. In addition, care actions taken in the past become observations of patient state; in our initial formulation, we did not include past care actions as additional observations.

Incorporating these data in a way that does not result in simplistic correspondences is a challenge. Our model does not require a forward-facing prediction in time; instead, it is trained on associations between observations and actions that occur close together in time. Thus, if we were to use lab test values as input observations of patient state, it would be easy to predict when a lab test (action) was performed. But, this is not a correspondence that provides insight into the care process.

Our initial formulation is also limiting in the way we evaluate performance. We take a holistic picture of care provided to a patient by evaluating mean-squared error over the entire space of an action category over time. However, certain actions are not available once others have been taken. For example, for a given patient, if one antibiotic has been administered, some of the others may no longer be options. In addition, once an antibiotic has been started, the option to start it is no longer available, but the option to stop it is. More carefully considering the set of possible actions at each time point (given the actions that have already been taken) would allow us to evaluate our predictions in a more realistic setting.

Learning the correspondences between observations and actions is a way of explicitly modeling the clinical decision-making process. Accurately characterizing this process can enable us to better understand what the expected course of care would be for a particular patient. We believe that our framework of learning how clinicians make decisions by relating observations to actions can lead to meaningful comparisons of physician behavior. These differences between actual care and expected care could then be related to patient outcomes.

Chapter 6

Summary and Conclusions

6.1 Summary

In this thesis, we presented three methods for leveraging correspondences in EHR data to improve clinical decision-making aids. Our work addresses a challenge to the use of machine learning models in health care in practice, and presents novel learning-based approaches to 1) summarize high dimensional health record data and 2) characterize care patterns using the relationship between observations of patient state and care provider actions.

In Chapter 3, we presented a method for building correspondences between structured data encodings across two different EHR systems by using an existing domain-specific concept vocabulary as an intermediate representation. We used natural language processing tools to annotate concepts in free-text descriptions of the structured item encodings. We demonstrated that machine learning models trained on one system and applied to another performed significantly better when features were encoded in the shared concept space, compared to the EHR-specific encodings. Our approach enables the portability of a machine learning risk model trained on one system to another.

In Chapter 4, we presented a method for learning correspondences between different modalities of health record data to summarize high-dimensional structured health record data. Summaries are written by care team members during the course of care,

in the form of clinical narrative notes. We used topic distributions over these existing clinical notes to guide our model. We demonstrated that our model was able to learn correspondences between structured health record data and topic distributions over existing clinical notes. In addition, we demonstrated that the generated topic representations are useful representations of patient state, as evidenced by improved performance on a set of downstream outcome prediction tasks. While these notes summarize care events and patient state, they are only intermittently available, and can sometimes contain errors or stale information about the patient. Our model can be used to suggest topics to care team members as they write a clinical note, and it is a first step towards automatically generating text-based summaries of the structured health record.

Finally, in Chapter 5, we learned correspondences between observations of patient state and actions taken by care team members. In the previous two chapters, we considered all events as observations of patient state. However, in Chapter 5, we took the perspective of the care provider. From a care provider's perspective, observations about patient state are made, and then actions are taken based on the provider's decision-making process. Our initial results suggest that we are able to learn the correspondence between observations of patient state and certain care action classes such as lab tests, antibiotic administration, and imaging tests. Our model captures meaningful differences in care patterns across different ICUs, and over the course of the patient stay.

6.2 Conclusion and Future Directions

In this thesis, we demonstrated that using correspondences across EHRs and within EHRs can lead to meaningful clinical insights.

We identified correspondences between subsets of data in the health record where there was both shared and unshared information. Different EHR systems still seek to capture similar clinical concepts. However, the patient data and feature encodings are not shared. Clinical narrative notes and structured health record data each record

observations of clinical care and state. However, notes summarize and are recorded infrequently, whereas structured data are detailed and frequently recorded. Finally, patient observations and care actions share an observer/actor: the care provider. But, actions capture decisions made by the provider, whereas observations capture patient characteristics. This framework of understanding where subsets of data intersect and diverge can highlight where learning correspondences can lead to clinical insight. In addition, learned correspondences between data modalities can lend insight into situations when data elements are redundant, and data collection processes could be streamlined.

In Chapters 4 and 5, we presented problem formulations that highlight the importance of high-level presentations of the large amounts of data available in EHRs. The amount of data clinicians need to ingest about even a single patient continues to grow. While many clinical decision-making aids summarize patient state through risk metrics, these risk-metrics are outcome-specific. Methods like the ones proposed in this thesis, and those in the area of patient phenotyping, present more general purpose outputs.

There are a number of ways we could build upon the work in this thesis.

In Chapter 3, we proposed a method to transition machine learning models across EHR systems. However, our evaluation was limited to a single EHR transition in a single hospital. In addition, because all dates in MIMIC are time-shifted, our work does not capture differences in care practice that may have existed around the time of the transition. Finally, MIMIC captures EHR-like data, but it is also a curated data warehouse. Future work could investigate how well our approach generalizes to systems closer to the point of care, and systems that may not have such comprehensive documentation of structured data items.

In Chapter 4, we proposed a method to generate topics for summaries of clinical care and patient state. Future work could investigate the feasibility of incorporating our work on generating such topics in a real EHR system. The interpretability of such topics is debated [85], and further evaluation of whether the topics learned from the clinical notes are meaningful to clinicians is needed. In addition, we trained a

single topic model over all categories of notes. However, physician notes and nursing notes differ in their structure, vocabulary, and length. Future work could investigate whether a category-specific topic model improves our ability to generate meaningful note topic distributions.

In Chapter 5, we proposed a method to characterize care patterns, by learning correspondences between observations of patient state and actions taken by care team members. We explored a single model architecture, and an evaluation metric that is sensitive to high class imbalance. Future work could investigate other methods of evaluating performance, as well as other model architectures that may be more well suited to learning correspondences between observations and care actions over time. For example, a convolutional architecture may be better suited to capturing relationships between observations and actions that are nearby in time than the LSTM we used.

The goal of machine learning models is to learn *generalizable* patterns in the data. As we discussed in the introduction to this thesis, clinical data are heterogeneous in patient characteristics and relevant care decisions. In this thesis, we sought to evaluate the generalizability of our conclusions on different patient populations by highlighting differences between care units. However, our modeling approaches do not explicitly account for differences between patient subtypes.

Another caveat to the generalizability of our claims is the use of MIMIC. First, MIMIC-III encompasses only critical care unit patients. Thus, we cannot examine the generalizability of our claims to other care settings. In addition, because MIMIC is de-identified and time-shifted, we are unable to evaluate if our approaches generalize over changes in care practice over time. Finally, MIMIC is from a single hospital. Our experiments are therefore limited to the ICUs at Beth Israel Deaconess.

However, increasing amounts of data are now available from other care settings and critical care units at other hospitals. Our models do not rely on data types that are available only in MIMIC; all of the data modalities we consider are part of most EHR systems. We believe the high-level ideas about enabling interoperability of clinical decision-making aids, learning correspondences across different modalities of

health care data, and understanding the relationship between observations of patient state and actions taken by care providers are generalizable to other institutions and to other care settings.

Machine learning to obtain data-driven insights about patient condition and clinical care has implications for the ongoing conversation about quality and cost of health care. Insights from retrospective health data, like the ones we discuss in this thesis, have the promise to positively impact the quality of care, improving outcomes and lowering cost.

Bibliography

[1] Centers for Medicare and Medicaid Services. National health expenditure fact sheet. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html, April 2018.

[2] Martin A Makary and Michael Daniel. Medical error-the third leading cause of death in the US. *BMJ: British Medical Journal (Online)*, 353, 2016.

[3] Jill Van Den Bos, Karan Rustagi, Travis Gray, Michael Halford, Eva Ziemkiewicz, and Jonathan Shreve. The $17.1 billion problem: the annual cost of measurable medical errors. *Health Affairs*, 30(4):596–603, 2011.

[4] Julia Adler-Milstein and Ashish K Jha. HITECH Act drove large gains in hospital electronic health record adoption. *Health Affairs*, 36(8):1416–1422, 2017.

[5] Office of the National Coordinator for Health Information Technology. Hospital progress to meaningful use by size, type, and urban/rural location. https://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-Progress-to-Meaningful-Use-by-size-practice-setting-area-type.php, August 2017. Accessed: 2018-04-23.

[6] Mark W Friedberg, Peggy G Chen, Kristin R Van Busum, Frances M Aunon, and Chau Pham. *Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy.* Rand Corporation, 2013.

[7] Stephen B Johnson, Suzanne Bakken, Daniel Dine, Sookyung Hyun, Eneida Mendonça, Frances Morrison, Tiffani Bright, Tielman Van Vleck, Jesse Wrenn, and Peter Stetson. An electronic health record based on structured narrative. *Journal of the American Medical Informatics Association*, 15(1):54–64, 2008.

[8] S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186, 2011.

[9] The merit-based incentive program. https://www.cms.gov/Medicare/Quality-Initrabilitiatives-Patient-Assessment-Instruments/

Value-Based-Programs/MACRA-MIPS-and-APMs/Merit-based-Incentive-Payment-System-pdf, November 2016.

[10] Quality Payment Program: MIPS Overview. https://qpp.cms.gov/mips/overview. Accessed: 2018-04-23.

[11] Office of the National Coordinator for Health Information Technology. Interoperability among U.S. Non-federal Acute Care Hospitals in 2015. https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-interoperability-2015.php, May 2016. Accessed: 2018-04-23.

[12] Hardeep Singh, Christiane Spitzmueller, Nancy J Petersen, Mona K Sawhney, and Dean F Sittig. Information overload and missed test results in electronic health record–based settings. *JAMA Internal Medicine*, 173(8):702–704, 2013.

[13] Justin M Weis and Paul C Levy. Copy, paste, and cloned notes in electronic health records. *Chest*, 145(3):632–638, 2014.

[14] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[15] Neil A Halpern, Debra A Goldman, Kay See Tan, and Stephen M Pastores. Trends in critical care beds and use among population groups and medicare and medicaid beneficiaries in the united states: 2000–2010. *Critical care medicine*, 44(8):1490, 2016.

[16] Constantine Manthous, Ingrid M Nembhard, and Andrea B Hollingshead. Building effective critical care teams. *Critical Care*, 15(4):307, 2011.

[17] Thomson Kuhn, Peter Basch, Michael Barr, and Thomas Yackel. Clinical Documentation in the 21st Century: Executive Summary of a Policy Position Paper From the American College of Physicians. *Annals of Internal Medicine*, 162(4):301–303, 2015.

[18] Benjamin J Miriovsky, Lawrence N Shulman, and Amy P Abernethy. Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. *Journal of Clinical Oncology*, 30(34):4243–4248, 2012.

[19] Leo Anthony Celi, Jeffrey David Marshall, Yuan Lai, and David J Stone. Disrupting electronic health records systems: the next generation. *JMIR Medical Informatics*, 3(4), 2015.

[20] Mike Wu, Marzyeh Ghassemi, Mengling Feng, Leo A Celi, Peter Szolovits, and Finale Doshi-Velez. Understanding vasopressor intervention and weaning: Risk

prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association*, 24(3):488–495, 2017.

[21] Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *Proceedings of AMIA Summits on Translational Science*, 2017:82, 2017.

[22] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. In *Proceedings of Machine Learning for Healthcare*, 2017.

[23] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. Learning to diagnose with LSTM recurrent neural networks. In *Proceedings of the International Conference on Learning Representations 2016.*

[24] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[25] Society of Thoracic Surgeons. Society of Thoracic Surgeons National Database, 2016.

[26] Society of Thoracic Surgeons. *STS Adult Cardiac Data Specifications: Version 2.61*, August 2007.

[27] Martin Dugas, Fleur Fritz, Rainer Krumm, and Bernhard Breil. Automated UMLS-based comparison of medical forms. *PloS one*, 8(7):e67883, 2013.

[28] David Baorto, Li Li, and James J Cimino. Practical experience with the maintenance and auditing of a large medical ontology. *Journal of Biomedical Informatics*, 42(3):494–503, 06 2009.

[29] Jenna Wiens, Wayne N. Campbell, Ella S. Franklin, John V. Guttag, and Eric Horvitz. Learning data-driven patient risk stratification models for clostridium difficile. *Open Forum Infectious Diseases*, 1(2), 2014.

[30] X. L. Dong and D. Srivastava. Big data integration. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1245–1248, April 2013.

[31] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8(2):I1, 2014.

[32] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(suppl 1):D267–D270, 2004.

[33] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. ACM, 2008.

[34] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[35] Christian Reich, Patrick B Ryan, Paul E Stang, and Mitra Rocca. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *Journal of Biomedical Informatics*, 45(4):689–696, 2012.

[36] Sebastian Mate, Felix Köpcke, Dennis Toddenroth, Marcus Martin, Hans-Ulrich Prokosch, Thomas Bürkle, and Thomas Ganslandt. Ontology-based data integration between clinical and research systems. *PloS ONE*, 10(1):e0116656, 2015.

[37] Yao Sun. Methods for automated concept mapping between medical databases. *Journal of Biomedical Informatics*, 37(3):162–178, 2004.

[38] Robert J Carroll, Will K Thompson, Anne E Eyler, Arthur M Mandelin, Tianxi Cai, Raquel M Zink, Jennifer A Pacheco, Chad S Boomershine, Thomas A Lasko, Hua Xu, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1):e162–e169, 2012.

[39] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 0:1–8, 2014.

[40] Sinno Jialin Pan, James T Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682, 2008.

[41] Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, and Shahram Ebadollahi. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 453–461. ACM, 2012.

[42] Alexander Van Esbroeck and Zeeshan Syed. Cardiovascular risk stratification with heart rate topics. In *Computing in Cardiology (CinC), 2012*, pages 609–612. IEEE, 2012.

[43] Chih-Chun Chia and Zeeshan Syed. Computationally generated cardiac biomarkers: Heart rate patterns to predict death following coronary attacks. In *SDM*, pages 735–746. SIAM, 2011.

[44] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. *Journal of the American Medical Association*, 270(24):2957–2963, 1993.

[45] Joon Lee, David M Maslove, and Joel A Dubin. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLoS ONE*, 10(5), 2015.

[46] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 75–84. ACM, 2014.

[47] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[48] Guy Divita, Qing T Zeng, Adi V Gundlapalli, Scott Duvall, Jonathan Nebeker, and Matthew H Samore. Sophia: a expedient UMLS concept extraction annotator. In *Proceedings of AMIA Annual Symposium*, page 467. American Medical Informatics Association, 2014.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[50] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[51] Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[52] Caleb W Hug and Peter Szolovits. ICU acuity: real-time models versus daily models. In *Proceedings of AMIA Annual Symposium*, page 260. American Medical Informatics Association, 2009.

[53] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83, 1945.

[54] Vinett Arora, J Johnson, D Lovinger, HJ Humphrey, and DO Meltzer. Communication failures in patient sign-out and suggestions for improvement: a critical incident analysis. *BMJ Quality & Safety*, 14(6):401–407, 2005.

[55] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[56] Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics*, 19(12):2227–2236, 2013.

[57] Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. Lifelines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–227. ACM, 1996.

[58] Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2014.

[59] Ayelet Goldstein and Yuval Shahar. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. *Journal of Biomedical Informatics*, 61:159–175, 2016.

[60] François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816, 2009.

[61] Jim Hunter, Albert Gatt, François Portet, Ehud Reiter, and Somayajulu Sripada. Using natural language generation technology to improve information flows in intensive care units. In *ECAI*, pages 678–682, 2008.

[62] Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

[63] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[64] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2506, 2016.

[65] Mehdi Moradi, Yufan Guo, Yaniv Gur, Mohammadreza Negahdar, and Tanveer Syeda-Mahmood. A cross-modality neural network transform for semi-automatic

medical image annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–307. Springer, 2016.

[66] Yohan Jo, Natasha Loghmanpour, and Carolyn Penstein Rosé. Time series analysis of nursing notes for mortality prediction via a state transition topic model. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1171–1180. ACM, 2015.

[67] Jen J Gong, Tristan Naumann, Peter Szolovits, and John V Guttag. Predicting clinical outcomes across changing electronic health record systems. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1497–1505. ACM, 2017.

[68] Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 52:199–211, 2014.

[69] Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of Biomedical Informatics*, 58:156–165, 2015.

[70] Ricardo Henao, James T. Lu, Joseph E. Lucas, Jeffrey Ferranti, and Lawrence Carin. Electronic health record analysis via deep poisson factor models. *Journal of Machine Learning Research*, 17(186):1–32, 2016.

[71] Vijay Huddar, Bapu Koundinya Desiraju, Vaibhav Rajan, Sakyajit Bhattacharya, Shourya Roy, and Chandan K Reddy. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, 4:7988–8001, 2016.

[72] Karla L Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78. ACM, 2015.

[73] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[74] Benjamin M Marlin, David C Kale, Robinder G Khemani, and Randall C Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398. ACM, 2012.

[75] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[76] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[77] François Chollet et al. Keras. `https://keras.io`, 2015.

[78] George Hripcsak and David J Albers. High-fidelity phenotyping: richness and freedom from bias. *Journal of the American Medical Informatics Association*, 2017.

[79] Zhengxing Huang, Xudong Lu, Huilong Duan, and Wu Fan. Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, 46(1):111–127, 2013.

[80] George Hripcsak, Patrick B Ryan, Jon D Duke, Nigam H Shah, Rae Woong Park, Vojtech Huser, Marc A Suchard, Martijn J Schuemie, Frank J DeFalco, Adler Perotte, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences*, 113(27):7329–7336, 2016.

[81] Jonathan H Chen, Mary K Goldstein, Steven M Asch, Lester Mackey, and Russ B Altman. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *Journal of the American Medical Informatics Association*, 24(3):472–480, 2017.

[82] Zhengxing Huang, Wei Dong, Peter Bath, Lei Ji, and Huilong Duan. On mining latent treatment patterns from electronic medical records. *Data Mining and Knowledge Discovery*, pages 1–36, 2014.

[83] George Hripcsak and David J Albers. Correlating electronic health record concepts with healthcare process events. *Journal of the American Medical Informatics Association*, 20(e2):e311–e318, 2013.

[84] Griffin M Weber and Isaac S Kohane. Extracting physician group intelligence from electronic health records to support evidence based medicine. *PloS one*, 8(5):e64933, 2013.

[85] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296, 2009.