Leveraging Text Representations for Clinical Predictive Tasks

by

Tristan Naumann

B.S., Columbia University (2009)
M.S., Columbia University (2010)

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

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

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

Certified by ..... Signature redacted .....
Peter Szolovits
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

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

Leveraging Text Representations for Clinical Predictive Tasks

by

Tristan Naumann

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

The increasing prevalence of digitized clinical data creates new opportunities to use machine learning to unlock clinical insights, and ultimately improve healthcare delivery. However, while data from Electronic Health Records (EHRs) have become common, they present unique challenges. Clinical data are noisy, sparse, irregularly sampled, and often biased in their recording of health state and care patterns. Further, much of the most important information used by care staff is recorded in unstructured text notes that are not easily deciphered by non-experts.

In this work, we present machine learning methods that distill large amounts of text-based clinical data into latent representations. These representations are then used to predict a variety of important outcomes. In particular, we focus on prediction tasks that can provide evidence-based risk assessment and forecasting in settings with guidelines that have not traditionally been data-driven. We consider several abstractions for clinical narrative text, and evaluate their utility on common predictive tasks, such as mortality and readmission. We argue that a "good" representation will improve performance on these tasks and that multiple representations may be necessary, as different models excel on differing tasks.

We present three case studies in which we use representations of clinical text to improve performance of clinical prediction tasks. First, we augment predictive models that used baseline clinical features by including features from clinical progress notes [31].These features are derived using Latent Dirichlet Allocation (LDA) and incorporated as features using per-patient topic membership. Notably, this representation has the benefit of interpretable topics over which each patient can be represented as a distribution.

Second, we explore the expressive power of clinical prose by evaluating the performance of several common models on both downstream clinical tasks and their ability to identify information contained in patients' notes [7]. This stands in contrast to much prior work that positions the utility of a given model solely with respect to its ability to improve downstream clinical performance. Such extrinsic evaluations are blind to much of the insight contained in the notes, thus motivating the need for intrinsic evaluations.

Finally, we use the text-based metadata associated with EHR encodings to allow the

transfer of predictive models from one database to another [35]. Existing machine learning methods typically assume consistency in how semantically equivalent information is encoded. However, the way information is recorded differs across institutions and over time, often rendering potentially useful data obsolescent. To address this problem, we map database-specific representations of the information to a shared set of semantic concepts, thus allowing models to be built from or transition across different databases.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Following many endeavors, it becomes apparent that three phrases should have been said more often than they were: "Thank you," "I'm sorry," and "I love you." And so, I would like to acknowledge some people who deserved to hear them far more frequently.

I thank my family for their love and support. My father, Otto Naumann, has always kept an excellent sense of humor about graduate studies, and made getting a second Ph.D. look effortless while I was still working toward a first. My mother, Heidi Shafranek, has offered unwavering encouragement, and her commitment to helping me see beyond work has been essential. My siblings, Ambrose and Siena, make me incredibly proud, and inspire me to constantly work harder in order to be the older brother that they deserve.

Moving to Boston meant seeing some people less often than I would have liked. Among them, Emily Sneeringer has endured more cross-country flights than anyone should. She's met me with unconditional patience, and encouraged me to continue even when I was unsure.

Moving also brought new friends, who have made living here wonderful. Jonathan Battat made me feel immediately welcome and found a place for us to call home. Nora Kelleher, Tom Pollard, and Jessica Liu have kept that same place filled with happiness. Jen Gong has been an incredible collaborator and friend, especially as we approached submission deadlines that always came later than I was led to believe.

Marzyeh Ghassemi has been the best officemate, collaborator, and friend imaginable—regardless of how long she claims it took us to first talk. She's provided immense support and the place she holds will be empty. She and her husband, Eric Munson, have supplied uncountable meals, and her children Raziyeh, Abbas, and Somayeh are an infinite source of happiness.

Finally, I thank my advisor, Peter Szolovits, and my thesis committee members, John Guttag and Anna Rumshisky. Their support, flexibility, and advice made this experience fulfilling.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Electronic health record (EHR) systems have become abundant in acute care hospitals (97% in 2014 [16]) and office-based practices (78% in 2015 [72]). While the data contained in EHRs are collected for the primary purpose of facilitating day-to-day operations, the increasingly large amount of data available present opportunities for their secondary use in retrospective analyses that can improve both the present understanding of patient physiology and clinical practice. Using machine learning to derive such insights has become an emerging topic of interest for researchers hoping to unlock the potential offered by EHRs. Indeed, recent successes include work in detecting lymph node metastases from breast pathology [33], autism subtyping by clustering comorbidities [24], large-scale phenotyping from observational data [77], and many other areas.

Applications of machine learning in healthcare create new opportunities to advance the field of medicine. Insights derived from observational data complement existing methods of clinical knowledge generation. For example, randomized controlled trials (RCTs) remain the gold standard in assessing treatment effects, but a large fraction of clinical decisions are not, or cannot be, based on high-quality RCTs [58]; thus many clinicians have limited evidence to guide their decisions. Even for those treatments that are based on an RCT, it is often the case that inclusion criteria result in a narrow cohort, which may not be representative of the more heterogeneous population that receives the treatment [90]. Here, machine learning

can provide a means to inform the priority of high-cost RCTs, highlighting those that yield the greatest impact, and even confirm the validity of their findings when applied in practice, often to more diverse patient populations.

The data contained in EHRs consist of many modalities, including high-frequency signals from medical instrumentation, sporadic results from lab tests, and clinical text from care staff. Among these modalities, clinical text is perhaps the most descriptive, consisting of clinical narratives written by care staff and text-based metadata associated with EHR encodings (e.g., the human-readable labels associated with laboratory tests). These unstructured, free-text data have become an emerging topic of interest for researchers hoping to unlock the potential offered by EHRs.

In this thesis, we argue that the use of free-text data is critical to delivering on the full potential offered by EHR data, and demonstrate the use of this modality in several case studies. These case studies use data from the intensive care unit (ICU), where the potential impact of EHR data is magnified due to the high cost of care and lack of evidence-based interventions. Critical care medicine in the United State had grown to cost over $80 billion annually by 2005 [40, 39], and these costs are incurred even when treatment is perceived to be futile by an external focus group of clinicians [45]. Further, a majority of the treatments commonly provided in ICUs have not been subject to an RCT [69], with some estimates as low as 10–20% of treatments backed by an RCT [71]. Among those treatments that are provided, many do not have a demonstrable impact on improving outcomes [74]. Meanwhile ICUs generate an abundance of data as a byproduct of the continuous monitoring required to support critical care. As a result, the ICU is a data rich environment, making it an ideal proving ground for improving our current standard of care, and doing so at a lower cost.

The case studies in this thesis present methodologies for leveraging free-text EHR data, and serve as steps toward the broader goal of improving care. The effective use of EHR text data requires overcoming issues that complicate the application of traditional machine learning methods; namely, those related to 1) challenges pervasive across EHR data modalities, and 2) challenges unique to clinical text.

1.1 Challenges of EHR Data

EHR data are not collected for the purpose of analysis; instead, they are intended to facilitate day-to-day operations, including the provision of care and billing. As a result, EHR data reflect the underlying care processes, and have a number of a properties that complicate traditional machine learning methods. Specifically, these data are irregularly sampled, noisy, sparse, and often biased in their recording of health state and care patterns. Further, these properties appear across multiple modalities, which must often be analyzed jointly to form an understanding of human health [97].

Consider the recording of several common data types. In an ICU, patient vitals are often recorded with high frequency and summarized hourly, while lab results are made available sporadically based on when they are ordered, static demographic data are typically recorded once per stay, and notes might be written whenever necessary and aggregated during each nursing shift. Such differences in data modality and sampling rate complicate learning, even when data are otherwise structured.

Each data type also typically reflects a varying degree of sparsity. For example, it may be the case that a lab is ordered but the resulting value not recorded, or a vital may be unrecorded for an extended period of time because of issues with the instrumentation. This sparsity often reflects bias in the collection of data since there are few measurements performed for all patients; instead, measurements are taken to facilitate diagnoses and thus the fact that something is measured may be meaningful on its own.

Additionally, uncertainty is pervasive both with respect to data that are collected, and the labels used for prediction targets. Uncertainty in data collection may take the form of noisy measurements—to which many machine learning methods are, or can be made, robust—but more often is the result of the generating process. For example, consider the time recorded with a given measurement. While it may have been recorded at the time the measurement or sample was taken, it may also have been recorded much later when a result became available, or care staff had time to record it. Uncertainty in the labels used for prediction targets also complicates traditional machine learning methods. For example, it may be tempting to use

billing codes as a prediction target in identifying diagnoses, but these codes may be recorded to reflect a chronic condition rather than the condition that actually led to the provision of care. Similarly, billing codes for some conditions may be recorded when a patient receives diagnostic testing, whether or not the tests reveal the presence of that condition.

1.2 Challenges of Clinical Text

Clinical text presents unique challenges. Many natural language processing (NLP) techniques that are well-studied in the general domain perform poorly in clinical settings where the underlying text reflects a distinct vocabulary, contains context-specific abbreviations and statements, and clinical practices (e.g., copy-and-paste) result in redundant information.

A distinct vocabulary means that many state-of-the-art resources have limited utility when transferred to the clinical domain because such resources don't reflect the underlying distributions of clinical text. This constraint can be mitigated by creating similar resources for clinical data, or fine-tuning existing models from the general domain on clinical data; however, the sensitive nature of clinical text means that data are not often available in the same quantity as the general domain (i.e., web scale).

Further, clinical statements and abbreviations are frequently dependent on context. For example, a statement like "the patient's condition worsened" will have very different meaning depending on whether a patient has a cold or has been intubated. This context dependence is exemplified in the use of abbreviations. A common abbreviation like "s/p" (*status post*, or "condition after") might be shared across specialties. However, an abbreviation like "T1" is more ambiguous. An oncologist may use "T1" to describe tumor size and its spread into nearby structures; whereas a radiologist may use "T1" to describe the image weighting in MRI sequences.

Finally, clinical narratives are used to communicate among care staff and often contain redundant information. Duplication can occur when care staff explicitly copy-and-paste existing information, a practice so pervasive that it has attracted significant academic attention [70, 41, 103, 101]. Redundant information can also be introduced implicitly as a means

of placing emphasis on those factors that are most important in the provision of care.

1.3 Contributions

In this thesis, we evaluate several methods to create patient representations using clinical text, and use the resulting features to predict outcomes of interest. In doing so our main contributions fall into two areas: 1) leveraging representations of clinical text to improve the performance of existing prediction tasks, and 2) exploring the information captured by these underlying representations.

These contributions represent steps toward a broader goal of using machine learning to improve healthcare. The first of these contributions serves to motivate the future use of clinical text in predictive applications; that is to say, our work demonstrates that clinical text provides additional, complementary information to structured EHR data, and thus should be used to improve predictive performance. The second contribution is necessary to better understand why a given representation is able to improve performance, and, to the extent possible, build trust in the underlying model. We demonstrate these contributions through several works in which we address either one or both of these issues to find appropriate representations that facilitate downstream predictive tasks.

The works herein were informed by several common considerations: leveraging domain knowledge, defining a meaningful clinical problem, and using representations of clinical text to do so. First, using appropriate domain knowledge is important to avoid results that are representative of the data, but clinically irrelevant. Indeed, applying machine learning methods out-of-the-box without sufficient domain knowledge often leads to unintended results. For example, Caruana et al. [13] showed this in work that considered applying a deep learning method to predicting the risk of dying of pneumonia. Notably, the authors find that a model learned that patients with asthma had a lower risk of dying, a finding that cannot be reconciled with existing clinical knowledge. However, this result is a reflection of the differing care provided to those patients who entered with asthma, something that becomes evident with the addition of clinical domain knowledge.

Second, defining the *right* clinical problem is an important, and often difficult, task. For example, in an ICU setting it may be tempting to predict in-hospital mortality using all information from a patient's record [43]. However, doing so naively, one might include a patient's billing codes. At first glance, such an inclusion seems sensible because billing codes contain information about diagnoses that were provided during the course of care. However, many of these codes are recorded after care has been provided (i.e., they are curated from other EHR data). As a result, such a prediction would depend on data that becomes available only after the point at which the prediction can be made meaningfully; thus, compromising its utility entirely.

Finally, in each of the works we leverage representations of clinical text. While unstructured clinical text presents unique challenges, it also presents unique opportunities. By its nature, the lack of structure is common across EHR systems. As a result, successful representations can often be transferred across EHR systems, or facilitate the transfer of other models across EHR systems. These two properties are critically important in enabling us to amass sufficient data for machine learning methods—an essential undertaking since simple machine learning methods often perform better than their more complex counterparts when sufficient data are available [38, 2]. Further, clinical text are used to communicate among care staff. In the case of clinical notes, we intuitively expect text to contain both a summary of the most relevant information from other signals, and subjective observations that cannot otherwise be instrumented. Thus, clinical notes constitute a channel that both highlighting important information and suppling new information.

1.4 Organization

This thesis is organized as follows:

Ch. 2 **Background**: The works that follow share several common clinical resources, as well as a common data source. We begin by providing a high-level overview of these resources prior to a more detailed discussion accompanying each work. Specifically, these works rely on the publicly-available MIMIC critical care databases [81, 48].

Ch. 3 **Representations for Predicting Clinical Outcomes**: We derive representations of clinical notes that are predictive of in-hospital mortality and post-discharge mortality in an ICU setting, and psychiatric readmission in a psychiatric care setting [31].

Ch. 4 **Representations for Predicting Intrinsic Note Information**: We explore the power of expressive clinical prose by evaluating the performance of several common models on both downstream clinical tasks and their ability to identify information contained in each note [7].

Ch. 5 **Representations for Predicting Outcomes Across Changing EHR Systems**: We map database-specific representations of information to a shared set of semantic concepts using the human-readable, text-based metadata associated with EHR encodings, thus allowing models to be built from or transition across different databases [35].

Relevant literature is highlighted alongside each of the works presented.

Chapter 2

Background

We make use of several resources frequently encountered in machine learning for healthcare. Their description here is intended to serve as an introduction for readers without a background in this domain. While we defer the particulars of each resource's usage to accompany its presentation, their description here also reduces repetition in subsequent chapters. These common resources fall broadly into several categories: 1) existing acuity scores, 2) data sources, and 3) clinical natural language processing (NLP) tools.

2.1 Acuity Scores

The intensive care unit (ICU) admits severely ill patients in order to provide life-saving treatment, such as mechanical ventilation. ICUs frequently have a very high staff to patient ratio in order to facilitate continuous monitoring of all patients and ensure that any deterioration in patient condition is detected and corrected before it becomes fatal; an approach that has been demonstrated to improve outcomes [51]. As a result, the ICU is a data rich environment.

A major effort has been placed in utilizing this data to both quantify patient health and predict future outcomes. The APACHE system was first published in Knaus et al. [54], and provided predictions for patient mortality based upon data collected in the ICU. While the

initial system was based on expert rules, later updates used data driven methods [104]. Other prediction systems have been developed as well, including the Acute Physiology Score (APS) III [56], Simplified Acute Physiology Score (SAPS) [60], SAPS II [62], the Sequential Organ Failure Assessment (SOFA) score [95], the Logistic Organ Dysfunction Score (LODS) [61], and the Oxford Acute Severity of Illness Score (OASIS) [47].

For a thorough review of severity of illness scores in the ICU, see Strand and Flaatten [88] and Keegan et al. [53]. Among the descriptions found there, it is important to note that these models were identified to lack sufficient calibration for use on the individual level [55], and some research goals were shifted to quantify the performance of ICUs and hospitals in aggregate, only. Nevertheless, a subset of these scores are commonly used on an individual level, providing a competitive baseline for predictive systems.

Among these many scores, the works contained in this thesis most heavily make use of SAPS II [62], either as a baseline or as a feature to inform models of patients' physiological states. The SAPS II score includes information from routine physiological measurements made during the first 24 hours, and is comprised of 17 variables: age, type of admission (scheduled surgical, unscheduled surgical, or medical), three underlying disease variables (acquired immunodeficiency syndrome, metastatic cancer, and hematologic malignancy), and 12 physiological variables. These physiological variables are comprised of a small number of parameters, and common SAPS II calculators use heart rate, systolic blood pressure, temperature, Glasgow Coma Scale [49], mechanical ventilation or CPAP, PaO2, FiO2, urine output, blood urea nitrogren, sodium, potassium, bicarbonate, bilirubin, and white blood cell count. The result is an integer score from 0–163 and a predicted mortality between 0% and 100%, which is used to determine the morbidity of a patient compared to other patients, or more frequently the morbidity of a population compared to other populations.

2.2 Data

Each of the works included in this thesis use data from the publicly-available Medical Information Mart for Intensive Care (MIMIC) critical care databases. These databases contain

de-identified electronic health record (EHR) data for patients seen at the Beth Israel Deaconess Medical Center (BIDMC). The data were collected as a collaboration among the MIT Laboratory for Computational Physiology (LCP), BIDMC, and Philips Health Care. The creation and use of the MIMIC database was approved by the Institutional Review Boards of both BIDMC and MIT (IRB Protocol 2001-P-001699/3).

The data contained in MIMIC include information pertaining to patient physiology such as demographics, hourly vital sign measurements, laboratory test results, procedures, medications, and notes. Additionally, MIMIC contains information about patient outcomes such as mortality and readmission, both in and out of the hospital—the later is achieved by joining with Social Security records. This abundance of data encompasses a large population of ICU patients, and is made freely available to researchers worldwide;[1] thus MIMIC has become one of the preeminent data sources in machine learning for healthcare. Perhaps most importantly, the use of an open data source facilitates reproducibility that is not otherwise easily achieved when using private data sets [46].

While MIMIC is used extensively throughout this thesis, the individual works were conducted over a period of time, and consequently each uses a different version of MIMIC. Specifically, Chapter 3 uses MIMIC-II [81], while Chapter 4 and Chapter 5 use versions of MIMIC-III [48].

2.2.1 MIMIC-II

MIMIC-II v2.6 is used for the work described in Chapter 3. Data in MIMIC-II were collected at BIDMC from 2001–2008, and cover 26,870 ICU patients. In addition to patient physiological recordings, MIMIC-II provides common acuity scores (e.g., SAPS II [62], which is described in Section 2.1), and billing codes given by International Classification of Diseases, Ninth Revision (ICD-9) diagnoses. Some derived data are also provided based on indicators in the records. Notably, the work in Chapter 3 makes use of medical co-morbidities called the Elixhauser score (EH) for 30 co-morbidities as calculated from the ICD-9 codes [26].

[1]The latest version of MIMIC can be downloaded from `https://mimic.mit.edu/`.

Patient mortality outcomes were also queried to determine which patients died in-hospital, or lived past the most recent query of Social Security records.

2.2.2 MIMIC-III

MIMIC-III is a successor of MIMIC-II, containing additional patient data and an updated data layout. The work presented in Chapter 4 makes use of MIMIC-III v1.4, and the work found in Chapter 5 makes use of MIMIC-III v1.3. While MIMIC-III v1.4 marked a major release with respect to data quality, the underlying population was the same as v1.3, consisting of over 58,000 hospital admissions for nearly 38,600 adult patients. MIMIC-III contains intensive care unit (ICU) data from the Beth Israel Deaconess Medical Center collected over the years 2001–2012. It is openly accessible to researchers and provides detailed patient information, including regularly sampled vital signs, demographics, lab test results, and time-stamped treatments and interventions. This data spans two EHR versions, CareVue (2001–2008) and MetaVision (2008–2012). There are approximately 9,000 items specific to CareVue and approximately 3,000 items specific to the MetaVision data.

2.3 Clinical NLP Tools

Many clinical natural language processing (NLP) tools have been developed. While this thesis does not cover advances in these tools, it does make use of some commonly used tools. The most widely-used clinical NLP tool is perhaps the clinical Text Analysis Knowledge Extraction System (cTAKES) [84], which is used in Chapter 5. cTAKES relies heavily on dictionary-based lookups from the Unified Medical Language System (UMLS) [8], a collection of medical ontologies [9]. An ontology consists of a set of concepts (*entities*), and *relations* between entities. Although general domain ontologies (e.g., Bollacker et al. [10]) and tools for identifying equivalent semantic concepts (e.g., Finkel et al. [28]) exist, these tools do not work well with the highly domain-specific vocabulary present in clinical text.

By relying on dictionary-based lookups from the UMLS, cTAKES is able to achieve high

recall (at the cost of low precision) by identifying all phrases that have any potential to be a relevant concept. While this property may be desirable for search-related tasks, its lack of relevance to many downstream clinical decision-making tasks has been noted as the reason for the development of additional tools, such as Sophia [22], the Eligibility Criteria Information Extraction (EliIE) [52], and the Clinical Language Annotation, Modeling, and Processing Toolkit [87]. Similarly, this limitation combined with a desire for out-of-the-box usability motivated the creation of projects such as CliNER [6].

While this thesis uses only cTAKES, many of the others were used in experiments and projects that are not reported. These earlier experiments informed the decisions made in works presented here. For example, the 2010 i2b2/VA Workshop on NLP Challenges for Clinical Records [93] promoted the development of 22 systems towards the task of concept extraction from discharge summaries. The winning system of the workshop challenge used a discriminative semi-Markov HMM, trained using passive-aggressive online updates [20]. Many other top performing methods used a Conditional Random Field (CRF) to model the sequence learning problem [79].

In the years following the shared task workshop, the dataset proved very useful as a research benchmark. Numerous systems and methods that have been developed can be compared against one another using this dataset. Early successful attempts utilized the strengths of workshop participants (sequential models, such as a CRF) and added generalized word representations using distributional semantics [29, 50, 102]. Since then, deep learning and recurrent neural networks have increased in popularity. Much like general domain NLP, clinical NLP has also been shown to benefit from deep learning models that can better learn complex patterns from the data, leading to many LSTM-based approaches to clinical concept extraction [14, 91]. Recently, Dernoncourt et al. [21] proposed a word- and character-level LSTM model for the de-identification task that outperformed all existing baselines.

Chapter 3

Representations for Predicting Clinical Outcomes

Accurate knowledge of a patient's disease state and trajectory is critical in a clinical setting. As modern electronic healthcare records contain an increasingly large amount of data, the ability to automatically identify the factors that influence patient outcomes stand to greatly improve the efficiency and quality of care.

The following chapter presents work that appeared at KDD and was done in collaboration with Marzyeh Ghassemi, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. In it we explore the use of text representations for predicting clinical outcomes, and demonstrate the value of incorporating information from clinical notes, via latent topic features (viz. Latent Dirichlet Allocation), for the task of predicting patient mortality. We evaluate our representation in three prediction regimes: (1) baseline prediction, (2) dynamic (time-varying) outcome prediction, and (3) retrospective outcome prediction. The baseline and retrospective prediction regimes establish lower and upper bounds on performance, respectively. The dynamic outcome prediction uses latent topic features derived from increasingly large subsets of the clinical notes as a semi-continuous indicator of patient state. We focus on the dynamic (time-varying) setting because models from this regime could facilitate an on-going severity stratification system that helps direct

care-staff resources and inform treatment strategies.

We found that latent topic-derived features were effective in determining patient mortality under three timelines: in-hospital, 30 day post-discharge, and 1 year post-discharge mortality. Our results demonstrated that the latent topic features that are important in predicting in-hospital mortality are very different from those that are important in post-discharge mortality. In general, latent topic features were more predictive than structured features, and a combination of the two performed best.

3.1 Overview

In a fragmented healthcare system of patients, doctors, caregivers, and specialists, an accurate knowledge of a patient's disease state is critical. Electronic monitoring systems and health records facilitate the flow of information among these parties to effectively manage patient health. However, information is not knowledge, and often only some of the information will be relevant in the context of providing care. Expert physicians need to sift through these extensive records to discover the data most relevant to a patient's current condition. As such, systems that can identify these patterns of relevant characteristics stand to improve the efficiency and quality of care.

This work focuses on the task of on-going mortality prediction in the intensive care unit (ICU). The ICU is a particularly challenging environment because each patient's severity of illness is constantly evolving. Further, modern ICUs are equipped with many independent measurement devices that often produce conflicting (and even false) alarms, adversely affecting the quality of care. Consequently, much recent work in ICU mortality models [62, 95, 47] has aimed to consolidate data from these devices (primarily structured data and physiological waveforms) and to transform these information streams into predictions. However, these works omit perhaps the most informative sources recorded about patients: free-text clinical notes and reports.

The narrative in the clinical notes, recorded by expert care staff, is designed to provide trained professionals a quick glance into the most important aspects of a patient's state. We

expect that combining features extracted from these notations with standard physiological measurements will result in a more complete representation of patients' states, thus affording improved outcome prediction. Unfortunately, free-text data are often more difficult to include in predictive models because they lack the structure required by most machine learning methods. To overcome the obstacles inherent in clinical text, latent variable models such as topic models [4, 1] can be used to infer intermediary representations that can in turn be used as structured features for a prediction task.

We demonstrate the value of incorporating information from clinical notes, via latent topic features, in the task of in-hospital mortality prediction as well as 30 day and 1 year post-discharge mortality prediction. Specifically, we evaluate mortality prediction under three prediction regimes: (1) baseline regime, which used structured data available on admission (2) time-varying regime, which used baseline features together with dynamically accumulated clinical text using increasigly large subsets of the patient's narrative record, and (3) retrospective regime, which used all clinical text generated from a hospital stay to supplement the baseline features. In each, our prediction task differs from the familiar time-varying situation whereby data accumulates; since fewer patients have long ICU stays, as we move forward in time fewer patients are available and the prediction task becomes increasingly difficult. In all targeted outcomes, we demonstrate that adding information from clinical notes improves predictions of mortality.

3.2 Related Work

Mortality models for acute (i.e., ICU) settings constitute a broad area of research. Siontis et al. [85] reviewed 94 studies with 240 assessments of 118 mortality prediction tools from 2009 alone. Many of these studies evaluated established clinical decision rules for predicting mortality, such as APACHE [56], SAPS II [62], and SOFA [95] (described in Section 2.1). The more recent OASIS score [47] uses machine-learning algorithms to identify the minimal set of variables capable of yielding an accurate severity of illness score (AUC 0.88).

Work by Hug and Szolovits [44] used several hundred structured clinical variables to create

a real-time ICU acuity score that reported an AUC of 0.88-0.89 for in-hospital mortality prediction. Notably, most of the predictive power of their models was from data gathered within the first 24 hours of the ICU stay. For example, their baseline computed acuity score (SAPS I) reported an AUC of 0.809 for in-hospital mortality prediction based on information during the first 24 hours of ICU stays in 1,954 patients. Their real-time acuity score (RAS) had AUCs of 0.875 on day 1, 0.880 on day 2, 0.878 on day 3, 0.871 on day 4, and 0.853 on day 5.

Several recent works have used information from clinical notes in their model formulations. Saria et al. [83] combined structured physiological data with concepts from the discharge summaries to achieve a patient outcome classification F1 score of 88.3 with a corresponding reduction in error of 23.52%. Similarly, Ghassemi et al. [30] described preliminary results indicating that topic models extracted from clinical text in a subgroup of ICU patients were valuable in the prediction of per-admission mortality. They found that common topics from the unlabeled clinical notes were predictive of mortality, and an RBF SVM achieved a retrospective AUC of 0.855 for in-hospital mortality prediction using only learned topics. Finally, Lehman et al. [64] applied Hierarchical Dirichlet Processes to nursing notes from the first 24 hours for ICU patient risk stratification. They demonstrated that unstructured nursing notes were enriched with clinically meaningful information, and this information could be used for clinical support. Using topic proportions, the average AUC for hospital mortality prediction was 0.78 ($\pm$ 0.01). In combination with the SAPS I variable, their average AUC for hospital mortality prediction was 0.82 ($\pm$ 0.003). While each work was performed using a different cohort—and, therefore, cannot be directly compared—their reported performances inform our expectations.

3.3 Methods

Figure 3-1 gives a general overview of our experimental process. First, we extract clinical baseline features, including age, sex, and SAPS II score, from the database for every patient. We also extract each patient's de-identified clinical notes. We use these notes as the

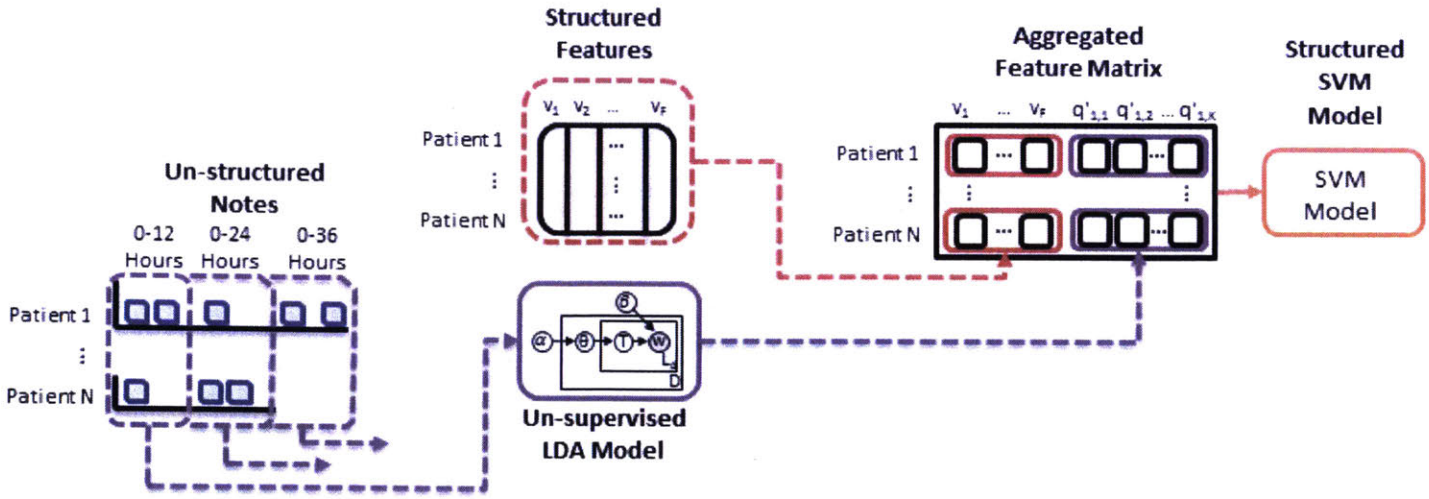


Figure 3-1: Overall flow of experiment. 1) Clinical baseline features are extracted from the database for every patient (e.g., age, sex, admitting SAPS II score) and derived features are computed (e.g., maximum/minimum SAPS II score) to form the *Structured Features* matrix v ($v_{p,f}$ is the value of feature f in the p^{th} patient). 2) Each patient's de-identified clinical notes are used as the observed data in an LDA topic model (i.e., *Un-supervisted LDA Model*), and a total of 50 topics are inferred to create the per-note topic proportion matrix q. 3) Per-note latent topic features are aggregated in extending 12 hour windows (e.g., notes within 0-12 hours, notes within 0-24 hours, etc.) and used to form matrix q' where $q'_{m,k}$ is the overall proportion of topic k in time-window m. 4) Depending on the model and time window being evaluated, subsets of the feature matrix v and matrix q' are combined into an *Aggregated Feature Matrix*. 5) A linear kernel SVM is trained to create classification boundaries for three clinical outcomes: in-hospital mortality, 30 day post-discharge mortality, and 1 year post-discharge mortality (i.e., *Structured SVM Model*).

observed data in an LDA topic model, and infer a total of 50 topics.[1] We normalize the word counts associated with each note, so that each note is represented by a 50-dimensional vector, summing to 1. These per-note topic distributions are then aggregated on a 12 hour semi-continuous timescale (e.g., notes within 0-12 hours, notes within 0-24 hours, etc.). A linear kernel SVM is trained to create classification boundaries with combinations of the structured clinical features and latent topic features to predict in-hospital mortality, 30 day post-discharge mortality, and 1 year post-discharge mortality.

[1]We selected 50 empirically after considering several parameterizations for the number of topics.

3.3.1 Data and Pre-Processing

We used ICU data from the MIMIC-II v2.6 database [81], a publicly-available, de-identified medical corpus described in Section 2.2.1. In addition to clinical baseline features, we extract International Classification of Diseases-Ninth Revision (ICD-9) diagnoses, and Disease-Related Group. Medical co-morbidities were represented by the Elixhauser scores (EH) for 30 co-morbidities as calculated from these ICD-9 codes. Patient mortality outcomes were also queried to determine which patients died in-hospital, or lived past the most recent query of Social Security records.

We extracted all clinical notes recorded prior to the patient's first discharge, including notes from nursing, physicians, labs, and radiology. The discharge summaries themselves were excluded because they typically stated the patient's outcome explicitly. Vocabularies for each note were generated by first tokenizing the free text and then removing stopwords using the Onix stopword list.[2] A TF-IDF metric [82] was applied to determine the 500 most informative words in each patient's notes, and we then limited our overall vocabulary to the union of the most informative words per-patient. This pre-processing step reduced the overall vocabulary down to 285,840 words from over 1 million words while maintaining the most distinctive features of each patient.[3]

Patients were excluded if they had fewer than 100 non-stop words or were under the age of 18. Specific notes were excluded if they occurred after the the end of the day in which a patient died or was discharged (e.g., radiology or lab reports whose results were reported afterwards). The resulting cohort consisted of 19,308 patients with 473,764 notes. We held out a random 30% of the patients as a test set. The remaining 70% of patients were used to train our topic models and mortality predictors. Table 3.1 summarizes the number of notes and patients in the training and test sets.

[2]Onix Text Retrieval Toolkit, API Reference, http://www.lextek.com/manuals/onix

[3]Some medical term canonicalization parsers were also examined, but we found their outputs to be fairly unreliable for this task.

Table 3.1: Cohort Composition

	Train	Test	Total
Patients	13,524	5,784	19,308
Notes	331,635	142,129	473,764

3.3.2 Structured and Derived Features

In total, we extracted and derived 36 structured clinical variables for each patient: the age, gender, SAPS II score on admission, minimum SAPS II score, maximum SAPS II score, final SAPS II score, and the 30 EH comorbidities. Data were scaled to avoid the range of a feature impacting its classification importance. This formed a feature matrix v, where the element $v_{p,f}$ was the value of feature f in the p^{th} patient.

3.3.3 Topic Inference

Instead of considering each note separately, we used the all notes that occurred in a particular time period as features for that period. We examined the distribution of note times, and found three peaks in note entry for any given day in a patient's stay (e.g., day 1, day 2, etc.): around 06:00, 18:00 and 24:00.[4] Given this distribution, we used 12-hour windows for our time windows.

Topics were generated for each note using Latent Dirichlet Allocation [4, 36]. Our initial experiments found no significant difference in held-out prediction accuracy across a range of 20 to 100 topics. We set hyperparameters on the Dirichlet priors for the topic distributions (α) and the topic-word distributions (β) as $\alpha = \frac{50}{numberTopics}, \beta = \frac{200}{numberWordsInVocab}$. Topic distributions were sampled from an MCMC chain after 2,500 iterations. This topic-modeling step resulted in a 50-dimensional vector of topic proportions for each patient for each note.

We concatenated the topic vectors into a matrix q where the element $q_{n,k}$ was the proportion of topic k in the n^{th} note. Of particular interest was whether certain topics were enriched for in-hospital mortality and long-term survival. We used an enrichment measure

[4]The increases in note submission at 06:00 and 18:00 were likely due to the current 12 hour nursing shift cycle. The large number of notes submitted at end-of-day were likely due to a previously common 14:00 - midnight nursing shift.

defined by Marlin et al. [67], where the probability of mortality for each topic is calculated as $\theta_k = \frac{\sum_{n=1}^{N} q_{n,k} * y_k}{\sum_{n=1}^{N} q_{n,k}}$, where y_n is the noted mortality outcome (0 for a patient that lives, and 1 for a patient that dies). These enrichment measures are reported in Section 3.4.1.

The time windows were used to construct feature vectors for each prediction task, where (at each step) we extended the period of consideration forward by 12 hours. From the previously constructed per-note matrix q that describes the distribution over topics in each note, we collapse into another matrix q' where $q'_{m,k}$ describes the overall proportion of topic k in time-window m. The element $q'_{m,k}$ is given by the mean of that topic's proportions of all the notes in time-window m: $\text{mean}_{n \in m} q_{n,k}$.

3.3.4 Prediction

We considered three prediction regimes with the inferred topic distributions: baseline prediction, dynamic (time-varying) outcome prediction and retrospective outcome prediction for the outcomes of in-hospital, 30-day, and 1-year mortality.

A separate linear SVM [15] was trained for each of the three outcomes, and each set of model features evaluated.[5] The loss and class weight parameters for the SVM were selected using five-fold cross-validation on the training data to determine the optimal values with AUC as an objective. The learned parameters were then used to construct a model for the entire training set, and make predictions on the test data.

All outcomes had large class-imbalance (mortality rates of 10.9% in-hospital, 3.7% 30 day post-discharge, and 13.7% 1 year post-discharge[6]). To address this issue, we randomly sub-sampled the negative class in the training set to produce a minimum 70%/30% ratio between the negative and positive classes. Test set distributions were not modified to reflect the reality of class imbalance during prediction, and reported performance reflects those distributions.

First, we established a static baseline model using only structured features present at

[5]The choice of linear kernel SVM was motivated by a fast implementation, though other choices (e.g., logistic regression) would be reasonable as well.

[6]This includes those who die within the first 30-days post-discharge, so two of the prediction targets have overlap.

admission (i.e., clinical baseline features and derived features thereof). We then ran dynamic outcome prediction in intervals of 12 hours at each step by including larger sets of patient notes in a step-wise manner. We finally performed retrospective outcome predictions, where we included structured features and all notes written during the stay as a static entity for prediction. Significantly, predictions of mortality with this type of feature set are a retrospective exercise only: it is not possible to first select all notes that occur before a patient's death, and then predict in-hospital mortality, because the time of mortality is not known *a priori*. The observer would have to "know" that the patient's hospital record was about to finish (either by death or discharge). The following settings were evaluated:

- *Admission Baseline Model*: A baseline model using the structured features of age, gender, and the SAPS II score at admission. These baseline features are extracted from the data present at patient admission only. (3 features total)

- *Time-varying Topic Models 1–20*: Outcome prediction performed by including notes in a step-wise fashion, extending the period of consideration forward by 12 hours at each step. For example, Time-varying Topic Model 1 includes topic features derived from all notes written during the first 12 hours of a patient's stay in the ICU, while Time-varying Topic Model 20 includes those derived from the first 240 hours. (50 features total)

- *Combined Time-varying Model 1–20*: Outcome prediction using the same setup as Time-varying Topic Models 1–20, but with the static structured features from Admission Baseline Model (gender, age, admitting SAPS score) included. (53 features total)

- *Retrospective Derived Features Model*: A retrospective model using the structured features of age, gender, admitting SAPS II score, the minimum SAPS II score, the maximum SAPS II score, the final SAPS II score, and all EH comorbidities. (36 features total)

- *Retrospective Topic Model*: A retrospective model using topics derived from all notes written during a patient's stay in the ICU. (50 features total)

- *Retrospective Topic + Admission Model*: A retrospective model combining structured features from Admission Baseline Model (gender, age, admitting SAPS scores) with latent topic features from Retrospective Topic Model. (53 features total)

- *Retrospective Topic + Derived Features Model*: A retrospective model combining structured features from Retrospective Derived Features Model (gender, age, admitting/min/-max/final SAPS scores, EH comorbidities) with latent topic features from Retrospective Topic Model. (86 features total)

We compare the prediction results for all models on each of the outcomes in Figure 3-3 and Table 3.5. We again emphasize that retrospective models are retrospective exercises only to establish the isolated and combined prediction ability of clinical notes and features. We also note that our *Time-varying Topic Model* is time-varying only in its application. We do not use other possible latent variable models such as "Dynamic topic models" [5], because we do not want to model the time evolution of topics, but rather the time evolution of membership to a given set of topics.

3.4 Results

3.4.1 Qualitative Enrichment

Table 3.2 lists the top words for the topics which had the largest enrichment ($\theta_k = \frac{\sum_{n=1}^{N} q_{n,k} * y_k}{\sum_{n=1}^{N} q_{n,k}}$) for in-hospital mortality, the smallest enrichment for in-hospital mortality, and the highest enrichment for 1 year mortality. The relative distributions of the in-hospital mortality probabilities for each of the 50 topics are shown in Figure 3-2. There was a wide variation in the in-hospital mortality concentration for the different topics, ranging from 3% – 30%. (See Table 3.3 for a listing of the top ten words for all topics.)

Table 3.2: Top ten words in topics enriched for in-hospital mortality, hospital survival (any number of days post-discharge), and 1 year post-discharge mortality.

	Topic	Top Ten Words	Possible Topic
In-hospital Mortality	27	name, family, neuro, care, noted, status, plan, stitle, dr, remains	Discussion of end-of-life care
	15	intubated, vent, ett, secretions, propofol, abg, respiratory, resp, care, sedated	Respiratory failure
	7	thick, secretions, vent, trach, resp, tf, tube, coarse, cont, suctioned	Respiratory infection
	5	liver, renal, hepatic, ascites, dialysis, failure, flow, transplant, portal, ultrasound	Renal Failure
Hospital Survival	1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp	Cardio-vascular surgery
	40	left, fracture, ap, views, reason, clip, hip, distal, lat, report	Fracture
	16	gtt, insulin, bs, lasix, endo, monitor, mg, am, plan, iv	Chronic diabetes
1-Year Mortality	3	picc, line, name, procedure, catheter, vein, tip, placement, clip, access	PICC[7] line insertion
	4	biliary, mass, duct, metastatic, bile, cancer, left, ca, tumor, clip	Cancer treatment
	45	catheter, name, procedure, contrast, wire, french, placed, needle, advanced, clip	Coronary catheterization

The topics enriched for in-hospital mortality presented a detailed view of the possible causes of death in the ICU. For example, patients in a modern ICU rarely die suddenly. Often patient life is sustained for some time in order for their family to express their wishes regarding terminal care and death. This could be one interpretation for Topic 27, which pertains to the discussion of end-of-life care options. Other topics with in-hospital mortality enrichment pertained to top causes of ICU mortality: respiratory infection (Topic 7), respiratory failure (Topic 15), and renal failure (Topic 5).

Hospital survival was also marked by topics which seem relevant to factors tied closely to the ability to recover from physiological insults: patients who are admitted for cardiovascular surgery (Topic 1) are often not allowed as surgical candidates until they are otherwise in very good health; patients who are able to respond to their care staff and the ICU environment (Topic 26, Table 3.3) are adequately dealing with the known stress of ICU admission; patients

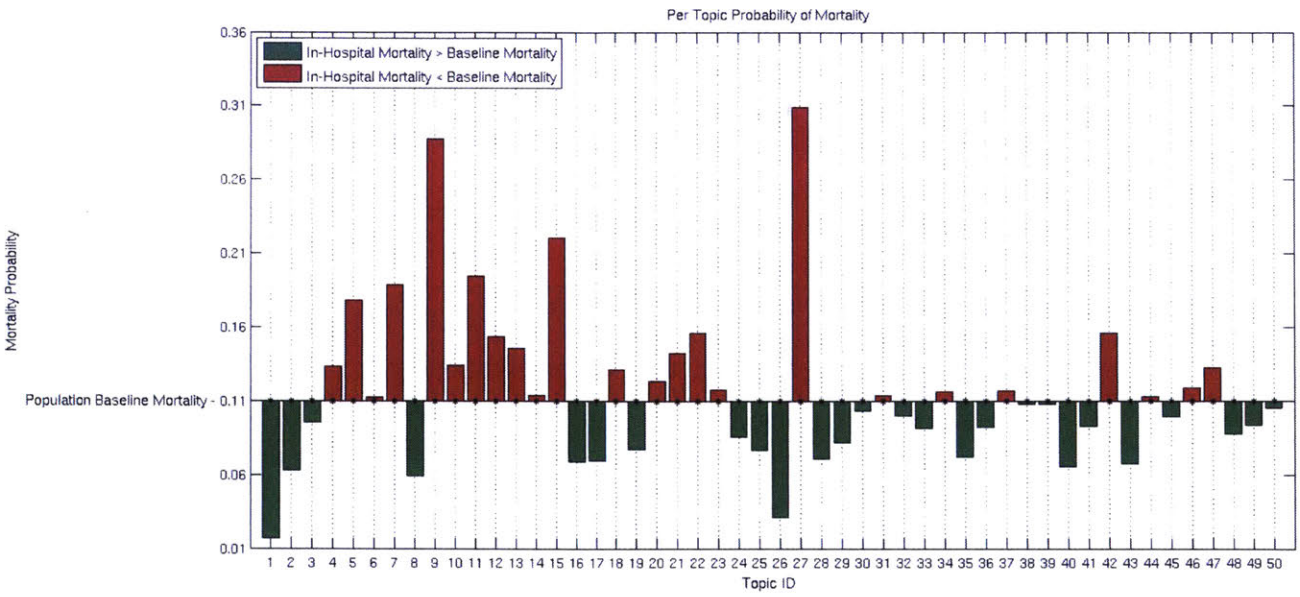


Figure 3-2: The probability of in-hospital mortality for each topic, indicating that topics represent differences in outcome. Probabilities are calculated as $\theta_k = \frac{\sum_{n=1}^{N} q_{n,k} * y_n}{\sum_{n=1}^{N} q_{n,k}}$ (see section 3.3.3). Each bar shows the prevalence of a given topic k in the mortality category, as compared to the set of all patients. Bars are shown as above (in red) or below (in green) the baseline in-hospital mortality based on the value of θ_k for each topic k.

with trauma-based injuries such as fracture and pneumothorax (Topics 8, 40); and patients with chronic conditions like diabetes (Topic 16).

The topics enriched for 1 year post-discharge mortality suggested that patients who are discharged but die within a year have conditions with a low chance of long-term survival. For example, cancer (Topic 4), the need for long-term IV access while in the ICU (Topic 3), and the use of coronary catheterization (Topic 45) to diagnose lesions in coronary arteries or other valvular/cardiac issues.

3.4.2 Prediction

We evaluated the predictive power of each model and outcome pair. Figure 3-3 shows the AUCs achieved by each model for the three targeted outcomes. Table 3.5 lists a more complete set of the SVM classification metrics.

Table 3.3: Top ten most probable words for all topics.

Topic Number	Top Ten Words
1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp
2	ccu, cath, mg, am, sp, groin, bp, cardiac, hr, cont
3	picc, line, name, procedure, catheter, vein, tip, placement, clip, access
4	biliary, mass, duct, metastatic, bile, cancer, left, ca, tumor, clip
5	liver, renal, hepatic, ascites, dialysis, failure, flow, transplant, portal, ultrasound
6	ct, contrast, pelvis, abdomen, fluid, bowel, clip, free, wcontrast, iv
7	thick, secretions, vent, trach, resp, tf, tube, coarse, cont, suctioned
8	chest, pneumothorax, tube, reason, clip, sp, ap, left, portable, ptx
9	remains, family, gtt, line, map, cont, levophed, cvp, bp, levo
10	name, neo, gtt, stitle, dr, sbp, resp, cont, wean, aware
11	remains, increased, temp, hr, pt, cc, ativan, cont, mg, continues
12	micu, code, stool, hr, bp, social, note, id, received, cchr
13	chest, pulmonary, bilateral, edema, portable, clip, reason, ap, pleural, effusions
14	resp, cough, sats, mask, sob, wheezes, nc, status, mg, neb
15	intubated, vent, ett, secretions, propofol, abg, respiratory, resp, care, sedated
16	gtt, insulin, bs, lasix, endo, monitor, mg, am, plan, iv
17	drainage, pain, abd, fluid, draining, drain, incision, sp, intact, pt
18	heparin, afib, ptt, am, gtt, mg, rate, hr, pvcs, iv
19	name, pacer, namepattern, placement, heart, pacemaker, ventricular, av, rate, chest
20	left, lung, effusion, lobe, pleural, lower, chest, upper, ct, opacity
21	skin, noted, care, left, applied, changed, draining, coccyx, wound, edema
22	tube, placement, tip, line, portable, ap, reason, position, chest, ng
23	noted, shift, name, pt, patent, patient, foley, agitated, soft, mg
24	hct, pt, gi, blood, bleeding, am, stable, unit, bleed, noted
25	name, am, mg, able, bp, time, night, times, doctor, confused
26	pain, co, denies, oriented, neuro, plan, diet, po, pt, floor
27	name, family, neuro, care, noted, status, plan, stitle, dr, remains
28	clip, reason, ro, medical, examination, evidence, impression, underlying, condition, normal
29	neuro, sbp, bp, commands, iv, cough, soft, status, lopressor, swallow
30	skin, stable, social, family, intact, tsicu, id, note, support, endo
31	woman, female, husband, name, pain, patient, pm, am, hospital, noted
32	diagnosis, admitting, name, reason, please, examination, yearold, eval, findings, underlying
33	name, neck, soft, patient, noted, anterior, epidural, level, posterior, namepattern
34	ct, contrast, chest, lymph, optiray, images, lesions, iv, nodes, lobe
35	left, stenosis, disease, clip, reason, carotid, severe, report, radiology, final
36	femoral, foot, left, leg, iliac, groin, lower, patent, graft, extremity
37	acute, reason, head, clip, evidence, eval, name, wo, status, ct
38	aortic, aorta, cta, wwo, dissection, recons, contrast, left, aneurysm, chest
39	left, ivc, filter, vein, pulmonary, veins, dvt, clip, inferior, upper
40	left, fracture, ap, views, reason, clip, hip, distal, lat, report
41	spine, cervical, spinal, clip, thoracic, fall, lumbar, vertebral, contrast, reason
42	hemorrhage, head, ct, left, frontal, contrast, subdural, hematoma, clip, bleed
43	ct, trauma, contrast, injury, fracture, fractures, pelvis, clip, wcontrast, sp
44	contrast, brain, head, left, mri, images, mra, stroke, clip, cerebral
45	catheter, name, procedure, contrast, wire, french, placed, needle, advanced, clip
46	artery, left, common, distal, catheter, internal, branches, flow, name, middle
47	vein, stent, catheter, name, mm, portal, tips, balloon, venous, sheath
48	service, distinct, procedural, artery, sel, carotid, left, cath, name, clip
49	catheter, name, performed, embolization, contrast, bleeding, procedure, mesenteric, extravasation, clip
50	artery, carotid, left, aneurysm, injection, vertebral, internal, evidence, clip, cerebral

As shown in Table 3.5, the prevalent class imbalance resulted in a bias toward low specificities in the *Admission Baseline Model*. The balance between sensitivity and specificity generally leaned towards favoring higher specificities for in-hospital and 30 day mortality

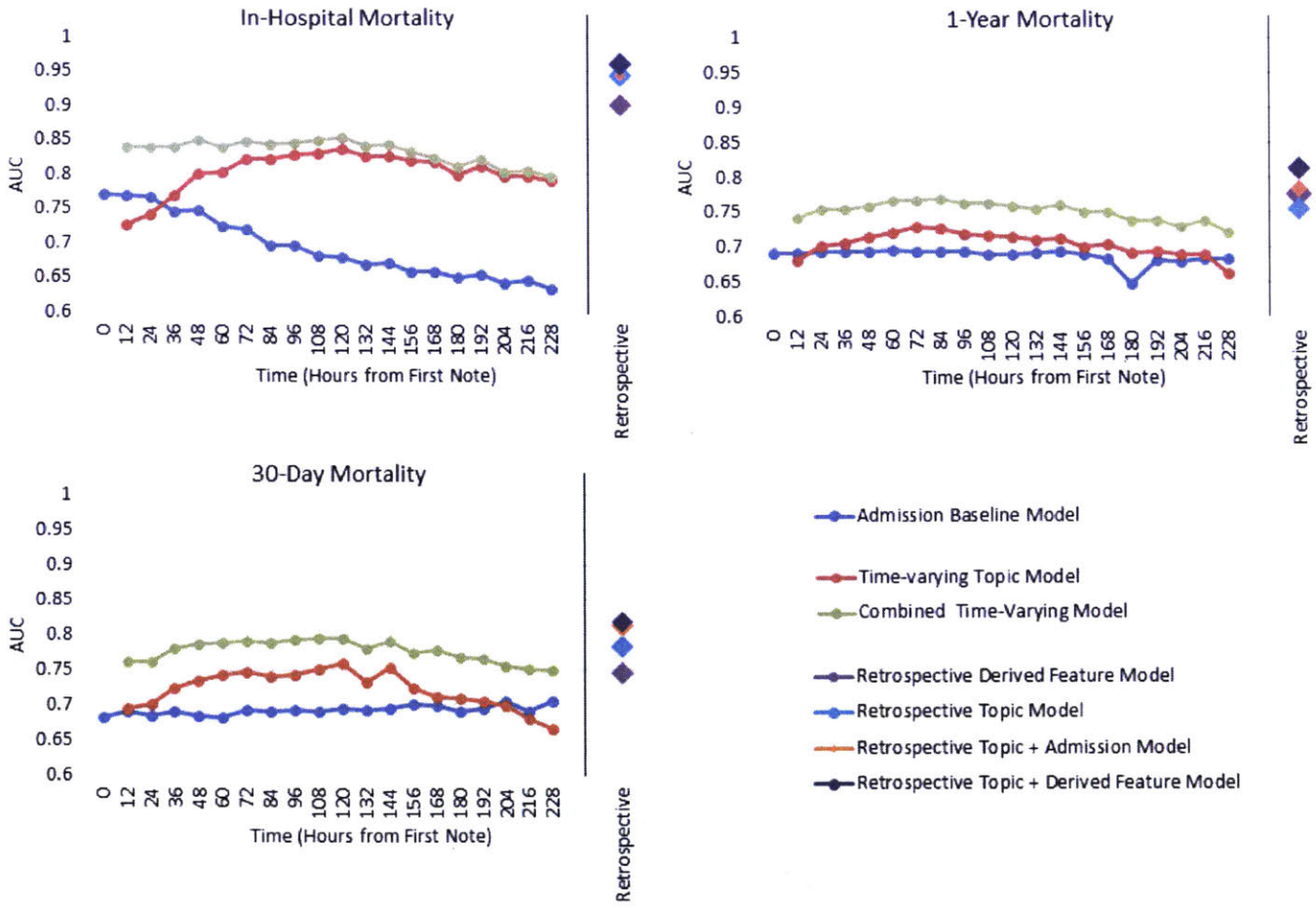


Figure 3-3: Linear SVM model performance measured via AUC on three outcomes: in-hospital mortality, 30 day post-discharge mortality, and 1 year post-discharge mortality. In each case, the features used are described in detail in Section 3.3.4. Our prediction task is different from the usual situation where data is accumulated over time. Since fewer patients have long ICU stays, in this case, we actually lose data points as time goes on, making the prediction task harder. For example, at time 0 there are 5,784 patients (5,157 controls/627 positives for in-hospital mortality) in the test set. By 72 hours, this had dropped to 5,084 patients (4,591 controls/493 positives for in-hospital mortality) and at 144 hours to 3,496 patients (3,141 controls/355 positives for in-hospital mortality). (Table 3.4)

prediction as time moved forward in the *Time-varying* models, but this was not uniformaly true in all cases. In general, the *Retrospective Derived Features Model* had a high sensitivity and low specificity, the *Retrospective Topic Model* had good specificity, and the combined models tended to have a more even set of both measures.

For 30 day and 1 year post-discharge mortality prediction, the *Admission Baseline Model*

Table 3.4: Patient cohort size at each time tested by time-varying models. Note that patients are removed from a prediction time if they are discharged or die prior to that time.

		Cohort Size (Control, Positive)		
Time (Hours)	**Total**	**In-Hospital**	**30 Day**	**1 Year**
0	5784	5157, 627	5597, 187	5058, 726
12	5784	5157, 627	5597, 187	5058, 726
24	5749	5128, 621	5563, 186	5026, 723
36	5563	4998, 565	5382, 181	4855, 708
48	5497	4937, 560	5318, 179	4795, 702
60	5161	4664, 497	4986, 175	4480, 681
72	5084	4591, 493	4911, 173	4407, 677
84	4691	4241, 450	4524, 167	4043, 648
96	4587	4140, 447	4421, 166	3945, 642
108	4116	3710, 406	3963, 153	3530, 586
120	4030	3626, 404	3877, 153	3448, 582
132	3570	3210, 360	3427, 143	3023, 547
144	3496	3141, 355	3354, 142	2956, 540
156	3026	2707, 319	2898, 128	2533, 493
168	2967	2652, 315	2840, 127	2479, 488
180	2580	2291, 289	2468, 112	2138, 442
192	2541	2254, 287	2431, 110	2109, 432
204	2215	1953, 262	2117, 98	1825, 390
216	2186	1925, 261	2090, 96	1802, 384
228	1925	1681, 244	1837, 88	1575, 350

was very steady, averaging an AUC of 0.68 over all time windows for both outcomes. The *Combined Time-varying Model* achieved an average/best performance of 0.77/0.8 for 30 day mortality and 0.75/0.77 for 1 year mortality. In both outcomes the *Time-varying Topic Model* performed strictly better than the *Admission Baseline Model* until the available patient subset became minimal (the 204 -216 hour windows), and the *Combined Time-varying Model* was always better than either alone.

As expected, the four *Retrospective* models were generally more predictive than any of the *Time-varying* models. *Retrospective* models tended to increase performance as more features were added. For in-hospital and 30 day mortality prediction, the *Retrospective Topic Model* performed better than the *Retrospective Derived Features Model* (AUCs increased from 0.90 to 0.94 and 0.75 to 0.78 respectively). For 1 year mortality this was reversed (AUC decreased

Table 3.5: Detailed model prediction results for three outcomes: in-hospital mortality, 30 day post-discharge mortality, and 1 year post-discharge mortality. This also appears in Figure 3-3.

Outcome Predicted	Model Used	AUC	Sens.	Spec.
In-Hospital Mortality	Admission Baseline Model	0.771	0.999	0.010
	Time-varying Topic Model 1	0.728	0.858	0.471
	...		...	
	Time-varying Topic Model 10	0.838	0.686	0.829
	...		...	
	Time-varying Topic Model 20	0.791	0.525	0.853
	Combined Time-varying Model 1	0.840	0.638	0.85
	...		...	
	Combined Time-varying Model 10	0.854	0.666	0.844
	...		...	
	Combined Time-varying Model 20	0.798	0.299	0.950
	Retrospective Derived Features Model	0.901	0.997	0.108
	Retrospective Topic Model	0.944	0.856	0.892
	Retrospective Topic + Admission Model	0.944	0.821	0.910
	Retrospective Topic + Derived Features Model	0.961	0.915	0.870
30 Day Mortality	Admission Baseline Model	0.683	0.995	0.075
	Time-varying Topic Model 1	0.695	0.150	0.944
	...		...	
	Time-varying Topic Model 10	0.759	0.817	0.551
	...		...	
	Time-varying Topic Model 20	0.665	0.602	0.579
	Combined Time-varying Model 1	0.761	0.348	0.885
	...		...	
	Combined Time-varying Model 10	0.796	0.641	0.770
	...		...	
	Combined Time-varying Model 20	0.75	0.011	0.991
	Retrospective Derived Features Model	0.745	0.941	0.220
	Retrospective Topic Model	0.783	0.342	0.909
	Retrospective Topic + Admission Model	0.813	0.872	0.633
	Retrospective Topic + Derived Features Model	0.818	0.096	0.985
1 Year Mortality	Admission Baseline Model	0.692	0.997	0.021
	Time-varying Topic Model 1	0.681	0.218	0.907
	...		...	
	Time-varying Topic Model 10	0.715	0.321	0.870
	...		...	
	Time-varying Topic Model 20	0.662	0.834	0.379
	Combined Time-varying Model 1	0.743	0.705	0.665
	...		...	
	Combined Time-varying Model 10	0.760	0.512	0.812
	...		...	
	Combined Time-varying Model 20	0.722	0.451	0.804
	Retrospective Derived Features Model	0.776	0.999	0.045
	Retrospective Topic Model	0.755	0.358	0.890
	Retrospective Topic + Admission Model	0.784	0.314	0.919
	Retrospective Topic + Derived Features Model	0.813	0.464	0.887

from 0.78 to 0.76).

In the in-hospital mortality setting, it seemed that admission features were not needed once latent topic features are known, but the derived features did provide extra information.[8] However, in the 30 day setting, latent topic features were similarly improved by either

[8]Adding the admission features did not improve the *Retrospective Topic Model*, but adding the derived features boosted AUC slightly to 0.96.

the admission features or the derived features.[9] This is likely because the derived features included EH comorbidities derived from the ICD-9 codes, and the ICD-9 codes themselves are often transcribed after a patient's discharge with the most actionable (or billable) conditions a patient presented. It is possible that these features are most relevant to in-hospital mortality risks (e.g., EH scores for myocardial infarction, congestive heart failure, etc.).

3.5 Discussion

Models that incorporated latent topic features were generally more predictive than those using only structured features, and a combination of the two feature types performed best. Notably, the combination provides a robustness that is able to perform well initially, leveraging primarily the structured information, and then continues to improve over the first 24 hours by incorporating the latent topic features. This resilience is particularly important since we observed that the first 24 hours of clinical notes appear to be the most meaningful toward predicting in-hospital mortality, while the baseline begins to steadily decrease.

Our observation of the importance of early data agrees with other reported results. Recall that, using topics derived from the first 24 hours of notes only, Lehman et al. obtained an average AUC for in-hospital mortality prediction of 0.78 ($\pm$ 0.01), and this was increased to 0.82 ($\pm$ 0.003) with the SAPS I variable. Further, Hug et al. obtained an AUC of 0.809 for in-hospital mortality prediction based on information during the first 24 hours of ICU. As such, we examined our results for in-hospital mortality when using topics derived from the first 24 hours of notes only (prediction time of 36 hours in Figure 3-3), and obtained corresponding AUCs of 0.77 for the *Time-varying Topic Model*, and 0.841 for the *Combined Time-varying Model*. Compared to Lehman et al's result, this implies that (with enough data) neither the extra hierarchical machinery added with HDPs nor the knowledge-based cleansing of medical terms before modeling improve prediction results (i.e., an AUC of 0.78 vs. 0.77). Compared to Hug et al's results, this implies that the addition of clinical text

[9] Adding the admission features to the *Retrospective Topic Model* improved AUC to 0.81 but adding the derived features did not improve AUC further.

provides reasonable performance boosts to the power of gold-standard structured information like SAPS II score (i.e., an AUC of 0.809 vs. 0.841).

Further, when predicting in-hospital mortality, we observed that the *Admission Baseline Model*'s predictive power (i.e., information acquired on admission) becomes much less valuable to predicting mortality as patients stay longer. This is likely because those who are not discharged within the first day of hospital admission are significantly sicker than those who are. Note that the average ICU stay time in the MIMIC-II database is 3 days across all units, and Figure 3-3 shows that after this time there was no additional predictive power gained by adding the structured admission information to the latent topic features (i.e., the *Time-varying Topic Model* and the *Combined Time-varying Model* converge).

This convergence draws attention to another interesting observation. Namely, both of the *Time-varying* models trended up in their ability to predict in-hospital mortality until $\sim$120 hours, and then trended down until the end of prediction. While initially counterintuitive, this is likely because of the loss of a significant number of patients (from both death and discharge) in the available patient cohort. For example, the test set population goes from 4,030 patients (3,626 control/404 positive for in-hospital mortality) to 3570 patients at this point (3,210 control/360 positive for in-hospital mortality).

Additionally, the predictive power of each topic changed depending on the target outcome. This appeals to intuition because, in a modern ICU, conditions that lead to in-hospital mortality are very different from those that would allow for a live discharge leading to a 30 day or 1 year mortality. As such, information about which topics tend to bias a patient towards any set of outcomes in useful for clinicians, when compared to the typical "black-box" approach to feature selection.

Finally, much work focuses on retrospective prediction of mortality outcomes. We also performed these predictions to compare the relative predictive power of different feature types and were able to achieve retrospective AUCs of 0.9, 0.94 and 0.96 for in-hospital mortality prediction using the *Retrospective Derived Feature Model*, *Retrospective Topic Model*, and combined *Retrospective Topic + Dervied Features Model*. However, we re-emphasize that

predictions of mortality with retrospective feature sets are not helpful or relevant for clinical staff because statistical functions of signals or features (e.g., min/max) and other structured data (such as ICD-9 codes and EH comorbidities) are not known *a priori*. Instead, these models are useful to establish upper bounds on what can be predicted from such data, and to compare to existing literature.

3.6 Conclusions

Modern electronic healthcare records contain an increasingly large amount of data including high-frequency signals from biomedical instrumentation, intermittent results from lab tests, and text from notes. Such voluminous records can make it difficult for care-staff to identify the information relevant to diagnose a patient's condition and stratify patients with similar characteristics.

Standard approaches to hospital mortality prediction use features such as gender, age, SAPS and SOFA score. In this work, we examined the utility of augmenting these standard features with textual information—specifically in the form of topic-based features—for predicting mortality in the ICU. Features extracted by latent variable models are attractive in this clinical application because scientific understanding is as important as clinical utility.

Qualitatively, the discovered topics correlated with known causes of in-hospital and post-discharge death. Further, adding latent topic features to structured clinical features increased classification performance in a variety of prediction scenarios: in-hospital mortality, 30-day mortality, and 1-year mortality.

The models and results explored in this work could ultimately be useful for interpretable models of disease and mortality.

Chapter 4

Representations for Predicting Intrinsic Note Information

The narrative prose contained in clinical notes is unstructured and unlocking its full potential has proved challenging. Many studies incorporating clinical notes have applied simple information extraction models to build representations that enhance a downstream clinical prediction task, such as mortality or readmission. Improved predictive performance suggests a "good" representation. However, these extrinsic evaluations are blind to most of the insight contained in the notes.

The following chapter presents work that appeared at the AMIA Informatics Summit and was done in collaboration with Willie Boag, Dustin Doss, and Peter Szolovits. In it, we explore the use of text representations for predicting intrinsic note information. Specifically, in order to better understand the expressive power of clinical prose, we investigate both intrinsic *and* extrinsic methods for understanding several common note representations.

4.1 Overview

Electronic Health Records (EHRs) contain an abundance of data about patient physiology, interventions and treatments, and diagnoses. The amount of data can be overwhelming in

the intensive care unit (ICU), where patients are severely ill and monitored closely. In this setting, it can be difficult to reconcile data from multiple sources; instead, care staff rely on clinical notes to provide summaries that capture important events and results. This unstructured, free text thus contains important observations about patient state and interventions, in addition to providing insight from caregivers about patient trajectory.

The secondary use of EHR data in retrospective analyses facilitates a better understanding of factors, such as those contained in clinical notes, that are highly predictive of patient outcomes [31, 11, 80, 37]. Additionally, the free-text nature of clinical notes means that data extraction does not rely heavily on each EHR's implementation, making methods for clinical notes portable across different EHRs. However, there are many ways to represent the information contained in text, and it is unclear how to best represent clinical narratives for the purpose of predicting outcomes.

Many efforts to leverage clinical notes for outcome prediction focus on improving the performance of a final prediction task [31, 30, 80, 64]. *Post hoc* feature analysis can assist in discovering those features that are most predictive, but it provides only a partial solution toward improving our understanding. We would like to know what facts and derived features matter most in affecting the predictive abilities of the models we build from them. This will allow us not only to improve performance but to understand what representations of the identified features are most useful.

For example, it has been shown that a patient's EHR-coded race and social history can help to identify a Gonorrhea infection accurately [92]. Therefore, if we are trying to use text analysis tools to make such an identification, we would like to know if those tools are able to determine a patient's race and social history accurately from the notes. While it is seemingly counter-intuitive to predict EHR-coded information using clinical notes, doing so provides insight into what is, and isn't, reflected in a given note's representation. Such awareness is important when designing representations for downstream prediction tasks because it exposes assumptions both about information is contained in the notes and what sophisticated models may be able to accomplish.

Toward the goal of understanding and improving note representations for downstream prediction performance, we consider several common representations and evaluate them on a variety of tasks. We explore performance on "easy" tasks, such as age, gender, ethnicity, and admission type, each of which are readily accessible as EHR-coded data. Additionally, we use the same representations on common prediction tasks, such as in-hospital mortality and length of stay. We show that 1) no single representation outperforms all others, 2) a simple representation tends to outperform more complex representations on "easy" tasks while the opposite is true for "common" prediction tasks, and 3) some seemingly "easy" tasks, such as ethnicity, are difficult for all of the representations considered.

4.2 Related Work

Work leveraging clinical notes for prediction can be broadly categorized into those focusing on clinical prediction tasks and those focusing on the representation of text.

Clinical Prediction Tasks: Several existing works have demonstrated the utility of clinical narratives in forecasting outcomes. A standard approach for converting narrative prose to structured vector-based features uses unsupervised topic modeling to represent each note as a distribution over various topics. Lehman et al. [64] and Ghassemi et al. [30, 31] use note-derived features in a framework to predict mortality. In recent work, Caballero Barajas and Akella [11] use generalized linear dynamic models on top of latent topics to detect an increase in the probability of mortality before it occurs. Luo and Rumshisky [66] use a supervised topic modeling approach to improve prediction of 30-day mortality. Grnarova et al. [37] use convolutional neural networks (CNNs) to construct document representations for the task of mortality prediction. Although the authors perform this prediction using all data from a patient's stay, their results show that both *doc2vec* [59] and their CNN approach improve performance relative to a topic representation. Further, Cohen et al. [19] explore the use of redundancy-aware topic models to combat the prevalent issue of copying notes forward in a patient's clinical record; however, they do not apply this model in a downstream prediction

task. Similarly, Pivovarov et al. [77] explore the use of topic models in the discovery of probabilistic phenotypes, but do not use these phenotypes to make predictions.

Text Representations: In the general domain, it has been observed that data often drives performance. Even simple models can often outperform complex models when they have sufficiently with access to sufficiently large data. Banko and Brill [2] observe this effect directly in the general natural language processing domain, noting that many methods continue to be optimized on small datasets and prove ineffective when applied to datasets orders of magnitude larger. Similarly, Halevy et al. [38] discover that "for many tasks, words and word combinations provide all the representational machinery we need to learn from text." Previously, limited access to clinical narratives have limited the applicability of this observation to the clinical domain.

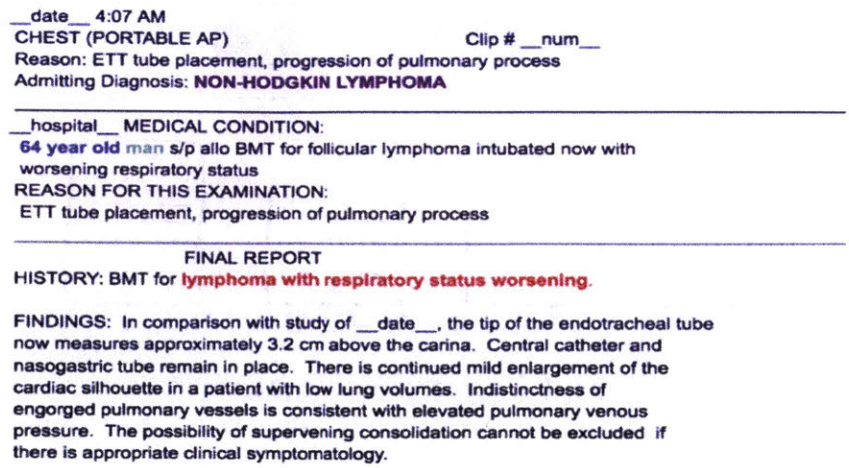
4.3 Data

This work uses data from the publicly-available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) database, version 1.4 [48], described in Section 2.2.2. A typical clinical note might look like the one shown in Figure 4-1, which shows the radiology report of a 64-year-old patient with poor respiratory status.

We consider only patients older than 15 years, who were in the ICU for at least 12 hours. Young patients are excluded since they typically exhibit different physiology from an adult population. Further, we include only each patient's first ICU stay, thus precluding training and testing on data from the same patient. Because of recording and measurement issues in the database, we exclude any ICU stays that do not conform to the common sense ordering of

$$hosp\_admission \leq icu\_intime \leq icu\_outtime \leq hosp\_dischtime$$

Finally, we consider *Nursing and Nursing/Other*, *Radiology*, and *Physician* notes, because other categories occurred relatively infrequently. For each ICU stay, we extract the first 24



__date__ 4:07 AM
CHEST (PORTABLE AP) Clip # __num__
Reason: ETT tube placement, progression of pulmonary process
Admitting Diagnosis: NON-HODGKIN LYMPHOMA

__hospital__ MEDICAL CONDITION:
64 year old man s/p allo BMT for follicular lymphoma intubated now with worsening respiratory status
REASON FOR THIS EXAMINATION:
ETT tube placement, progression of pulmonary process

FINAL REPORT
HISTORY: BMT for lymphoma with respiratory status worsening.

FINDINGS: In comparison with study of __date__, the tip of the endotracheal tube now measures approximately 3.2 cm above the carina. Central catheter and nasogastric tube remain in place. There is continued mild enlargement of the cardiac silhouette in a patient with low lung volumes. Indistinctness of engorged pulmonary vessels is consistent with elevated pulmonary venous pressure. The possibility of supervening consolidation cannot be excluded if there is appropriate clinical symptomatology.

Figure 4-1: An example clinical note. The age, gender, and admitting diagnosis have been highlighted. Also note, that descriptions such as "status worsening" suggest deterioration and possible in-hospital mortality.

notes (or fewer if the stay has fewer notes). These criteria result in 29,979 unique ICU stays, an equivalent number of patients, and 320,855 notes. The dataset is randomly divided into a 7:2 train/test split.

As "easy" prediction tasks, we extract several coded variables for each patient that remain constant throughout the stay, including: age, gender, ethnicity, and admission type. In addition, we also retrieve "common" clinical outcomes and findings during the stay, such as: diagnosis, length of stay, and in-hospital mortality. We then try to predict these characteristics and outcomes from different representations of the text notes.

As observed in replication studies, one of the central obstacles in replicability—even for work done on public datasets—is that descriptions of data cleaning and preprocessing are often inadvertently underspecified [46]. Therefore, we make our code publicly available.[1]

[1]Code available at `http://www.github.com/wboag/wian`.

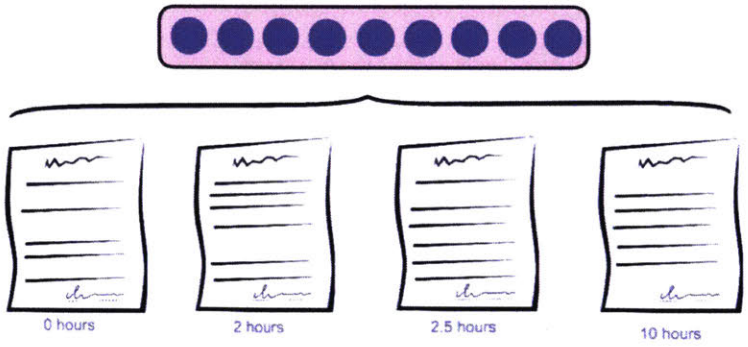


Figure 4-2: A patient's time in the ICU generates a sequence of timestamped notes. Each of the methods described transforms the sequence of notes into a fixed-length vector representing the ICU stay.

4.4 Methods

MIMIC-III v1.4 contains de-identified clinical notes. In preprocessing these notes, tags indicating de-identified protected health information are removed. Phrases written entirely in capital characters are then replaced by a single token, effectively coalescing common structural elements; for example, the section heading "RADIOLOGIC STUDIES" would be replaced with a single token. Additionally, regular expressions for common age patterns are used to replace all specified ages with symbols binned by decade to ensure that relevant age information is not lost. Finally, we remove all non-alphanumeric tokens, and normalize all remaining numbers to a single number token.

For each word, we compute the number of unique patients who have a note containing that word—this is the "document" frequency. For each note, we compute the term frequency-inverse document frequency, or *tf-idf* of each word and keep the top-20 words of that document. Thus a patient's stay is represented as an ordered list of filtered bags-of-words.

The following subsections describe several approaches to aggregate each patient's multiple note vectors into one fixed-size *patient vector* that summarizes their stay in the ICU, as illustrated in Figure 4-2.

4.4.1 Bag of Words

Bag of words (BoW) is one the simplest methods for creating vector representations of documents. Using the top-20 tf-idf words from each note produces a vocabulary of size $|V| = 17,025$ words. In this representation, the *patient vector* is a $|V|$-dimensional sparse, multi-hot vector. If a word appears in any of the notes for a given patient, then the corresponding dimension for that word is "on" in the resulting *patient vector.*

Bag of words presents a strong baseline representation for downstream predictive tasks. In this work, its strength is a result of its high dimensionality relative to other models: by reducing the representations into a smaller, denser space, other models may inadvertently throw out information with predictive value. More specifically, we expect that bag of words will perform well on tasks that involve the prediction of categories which may be directly represented by single words in their notes. For example, we would expect a note which frequently contains the word "male" to correctly identify the patient as male.

4.4.2 Word Embeddings

Because of the success of word2vec in recent years, we embed words and documents into a dense space in order to accommodate soft similarities. We train clinical word vectors using the publicly-available word2vec tool[2] on 129 million words from 500,000 notes taken from MIMIC-III. Hyperparameters were specified using Levy et al. [65] as a reference: 300-dimensional SkipGram with negative sampling (SGNS) method with 10 negative samples, a min-count of 10, a subsampling rate of 1e-5, and a 10-word window. These clinical embeddings are available for public use on the MIMIC-III Derived Data Repository.[3]

As shown in Figure 4-3, we create a note representation by aggregating the top tf-idf words in the document. With these top words, we look up each of their word2vec embeddings (blue) and collapse them into a final vector using elementwise -max, -min, and -average. We apply the same aggregation scheme (max, min, and average) to collapse the patient's list of

[2]Code available at `https://github.com/tmikolov/word2vec`.

[3]Data available at `https://physionet.org/works/MIMICIIIDerivedDataRepository/`.

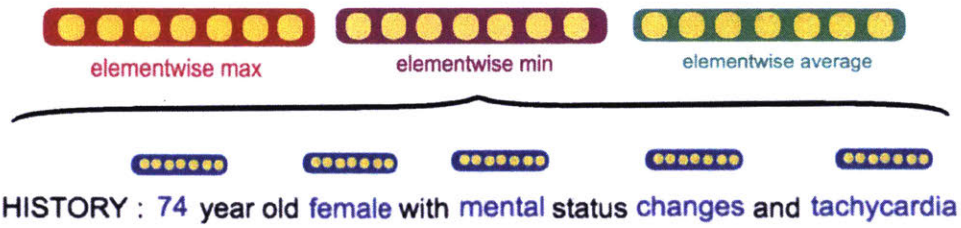


Figure 4-3: How the embedding for a single document is built by combining constituent word embeddings.

document vectors into one fixed-length *patient vector*.

4.4.3 Recurrent Neural Network

One problem with the approaches described above is that they all ignore temporal ordering of the documents. That is to say, they fail to track the progression of a patient's state over time during the ICU stay. One solution to this limitation is to use a sequence-based model. We use a Bidirectional LSTM network, which has proven to be effective at modeling temporal sequences [42, 57]. In order to provide a fair comparison, we build the list of document vectors for each patient in the same way that was done for word embeddings. These document vectors are then fed into the LSTM one document per timestep.

Our LSTM was implemented in Keras [18] using a Bidirectional LSTM with 256 hidden units, a dropout rate of 0.5, and a 128-unit fully connected layer immediately before the output label softmax. Models were trained for 100 epochs.

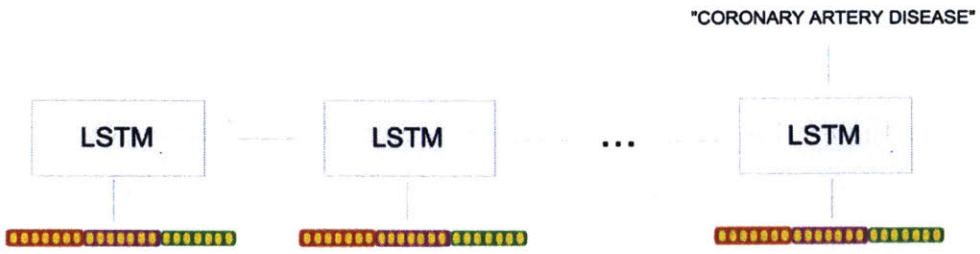


Figure 4-4: The many-to-one prediction task for the LSTM, in which a document representation is fed in at each timestep, and it makes a prediction (e.g., diagnosis) at the end of the sequence.

4.5 Experimental Setup

The principal aim of this work is to better understand what information is captured by various representations of clinical notes. Because most of the derived representations have non-interpretable dimensions from the embedding process, we cannot look for correlations between individual dimensions and our queries. Instead, we use downstream predictive performance to asses whether a particular representation has encoded the necessary information.

We consider the following prediction tasks modeling clinical states and outcomes:

1. **Diagnosis.** We filter down to patients with one of the 5 most common primary diagnoses and predict: *Coronary Artery Disease*, *Pneumonia*, *Sepsis*, *Intracranial Hemorrhage*, and *Gastrointestinal Bleed.*[4]

2. **In-Hospital Mortality.** Binary classification of whether the patient died during their hospital stay.

3. **Admission Type.** Binary classification of *Urgent* or *Elective.*

4. **Length-of-Stay.** Three-way classification of whether patients stayed in the ICU for *Less than 1.5 days*, *Between 1.5 and 3.5 days*, and *longer than 3.5 days.*

We are also interested in whether the notes are able to capture basic demographic information:

1. **Gender.** Binary classification of *Male* or *Female.*

2. **Ethnicity.** Binary classification of *White* or *Non-White.*

3. **Age.** Three-way classification of age as *less than 50 years old*, *between 50 and 80 years old*, or *older than 80 years old.*

While these tasks reflect those commonly found in research, we use them to evaluate our representations rather than as clinically-actionable targets. For example, it might be noted

[4]While these are the top 5 most common diagnoses in our cohort, we note that they correspond to very different conditions. We expect that this will make the task easier than, for example, discriminating among diagnoses that are similar.

that a single patient can suffer from multiple conditions, but here we consider only their primary diagnosis. Similarly, the ranges for age and length-of-stay are reasonable, but would need to be tailored in other conceivable applications. In both cases, however, these choices serve to highlight the types of information each representation is capturing.

Binary classification tasks are evaluated using AUC, while multi-way classification tasks are evaluated using the macro-average F1-score of the different labels. Predictions are made for bag-of-words and word embedding representations using a scikit-learn [76] support vector classifier with linear kernel. Predictions are made for the LSTM using a softmax layer.

4.6 Results

Performance for the 7 classification tasks using the 3 representation models are shown in Table 4.1 (binary classifications) and Table 4.2 (multi-way classifications). In general, our findings match our expectations: while a complex model tends to do well for "downstream" tasks involving reasoning, such as diagnosis and length-of-stay, it struggles to compete with a simpler model in token-matching tasks like age and gender.

Specifically, the bag-of-words (BoW) model performs best at predicting so-called "common-sense" tasks: age, gender, and (less significantly) ethnicity, for which there are words which almost directly predict the labels. In contrast, the LSTM model outperforms BoW on tasks more related to clinical reasoning: diagnosis and length of stay, for which we expect the temporal information to be important in predictions. Embeddings serve as a halfway point between BoW and LSTM; while the method does not make use of a temporal sequence, this experiment allows us to untie the pre-trained word vectors from the temporal dynamics of the LSTM. In doing so, we see that the embeddings typically perform competitively against BoW, but the LSTM is able to further use them.

Table 4.1: AUCs for the binary classification tasks.

	in-hospital mortality	admission type	gender	ethnicity
BoW	**0.821**	**0.883**	**0.914**	**0.619**
Embeddings	0.814	0.873	0.836	0.580
LSTM	0.777	0.870	0.837	0.533

Table 4.2: Macro-average F1 scores for the multi-way classification tasks.

	diagnosis	length of stay	age
BoW	0.828	0.724	**0.635**
Embeddings	0.828	0.730	0.544
LSTM	**0.836**	**0.758**	0.450

4.7 Discussion

As shown in Table 4.1 and Table 4.2, the different models exhibit varied performance across tasks with no consistent winner. Bag-of-words tends to do well on tasks where a single word, or a few words, are strongly associated with prediction categories. Notably, bag-of-words is much better at predicting age. This is likely because the normalized, per-decade age tokens created during preprocessing are, of course, strongly associated with predicting age. The LSTM, on the other hand, had a difficult time distinguishing between the age token embeddings since all age tokens fall nearby one another within the embedding space, as shown in Figure 4-5.

For these tasks, bag-of-words provide a strong baseline because some standard demographic information, such as age and gender, are typically specified in the notes. However, it is precisely because of their frequency of occurrence that information retrieval methods, such as tf-idf, underestimate their importance. Recall that tf-idf reduces the score of exceedingly common words. While this step is clearly important in the treatment of "stopwords"—words that are so common they provide no additional value—here it inadvertently removes commonly recorded information. This presents a challenge for aggregating the word embeddings of a note into one single document embedding because including too many words in the aggregate statistical values (i.e., averages, maximums, and minimums) drives down the "informativeness" of the representation by adding noise to these aggregate statistics.

Table 4.3: Most predictive words for gender: (a) Male, (b) Female.

(a) Male

man	1.4012
he	1.0589
wife	0.9953
male	0.7956
his	0.6772
prostate	0.2435
prop	0.1965
ofm	0.1850
hematuria	0.1816
esophagectomy	0.1812
distention	0.1756
trauma	0.1748

(b) Female

she	1.0176
woman	0.9051
her	0.7561
husband	0.7004
breast	0.3206
daughter	0.2656
nausea	0.2309
female	0.2246
commode	0.2183
responded	0.2052
fick	0.2009
cco	0.1975

Table 4.4: Most predictive words for admission types: (a) 'Urgent' admissions, and (b) 'Elective' admissions.

(a) 'Urgent'

ew	0.2639
er	0.2495
fracture	0.2258
fx	0.2248
osh	0.2235
b	0.2194
disease	0.2138
vertebral	0.2061
cabg	0.2029
fractures	0.1971
fall	0.1893
arteriogram	0.1877

(b) 'Elective'

sda	0.8048
flap	0.4646
esophagectomy	0.4617
artery	0.4435
epidural	0.4415
valve	0.3845
lobectomy	0.3838
resection	0.3644
avr	0.3527
replacement	0.3324
nephrectomy	0.2812
whipple	0.2740

Further, all methods achieve high AUCs for mortality, admission type, and gender; similarly, each performs poorly for ethnicity. The highest ethnicity AUC is still 20 points lower than the worst reported AUC for the other tasks. This suggests that predicting ethnicity from notes is an inherently difficult challenge. This is largely because race, while commonly coded elsewhere, is not typically specified in the notes. Additionally, 71% of patients are white in our dataset. This class imbalance may be large enough that a "default" value may be assumed and not recorded. When ethnicity *is* mentioned, it is usually to denote a language barrier, for example, "Spanish-speaking" or "required translator."

In general, interpretability is seen as a desirable feature for machine learning, particularly in the clinical setting: doctors care not only about what decision is made, but what information is used to inform that decision. Here, BoW seems to have a natural advantage over other embedding models, as it is very easy to examine what words have the most predictive power for given tasks.

Indeed, Table 4.3 clearly demonstrates the interpretability of the features for predicting gender. Words such as 'man', 'male', 'wife', and 'he' directly suggest a male patient, and these are shown to have high predictive power for gender. More interestingly, we see words corresponding to gender-correlated conditions and body parts, such as 'prostate' for men

Figure 4-5: PCA 2-D projection of the word embeddings. Vectors of the special age tokens are colored red. Note that these tokens cluster close together in the embedding.

Table 4.5: Most predictive words for length-of-stay: (a) Short stay (0 - 1.5 days), (b) Medium stay (1.5 - 3.5 days), (c) Long stay (> 3.5 days)

(a) Short stay (0 - 1.5 days)

ml	0.5295
pt	0.5086
to	0.3570
b	0.3403
sensitivity	0.2489
meq	0.2090
atrial	0.1934
tamponade	0.1784
valuables	0.1770
vomited	0.1738
s	0.1708
weaning	0.1676

(b) Medium stay (1.5 - 3.5 days)

followed	0.2014
aps	0.1888
lifting	0.1811
of	0.1796
device	0.1790
trunk	0.1747
available	0.1644
metastatic	0.1610
the	0.1603
holes	0.1576
this	0.1520
decubitus	0.1509

(c) Long stay (> 3.5 days)

amio	0.2470
mn	0.2393
brain	0.2172
decreasing	0.2002
fentanyl	0.1971
withdrawal	0.1933
vasospasm	0.1900
previously	0.1890
coiling	0.1811
exercises	0.1799
dobbhoff	0.1779
frequently	0.1776

and 'breast' for women. Unsurprisingly, BoW performs better than other methods on this task.

Admission type, with features shown in Table 4.4 is less-easily interpreted, but still provides understandable features. Words such as 'er' and 'ew' refer to the emergency room or ward, and 'fracture' or 'fall' refer to traumatic injuries, all of which reasonably suggest an urgent-care admission. Conversely, many of the predictive words for elective admissions suggest chronic conditions or planned surgical procedures ('artery', 'valve', 'replacement'). We see that BoW also performs quite well on this task.

However, we see some differences when we examine the predictive features for the length-of-stay task in table 4.5. In contrast to gender or admission type, the features for length-of-stay are much more generic, seeming to have little interpretable relation to the prediction task. At the same time, we see that the LSTM achieves a higher F1 score as compared to the BoW model for this task. This suggests that BoW is interpretable for the simple token-matching tasks, but not the harder reasoning tasks. Therefore, more complex and performant models should be used for these harder tasks.

4.8 Conclusions

In this work we consider both demographic and clinical prediction tasks in order to "stress test" a variety of common note representations. We show that different representations have different strengths: while complex models can outperform simple ones on reasoning tasks, they struggle to capture seemingly "easy" information. On the other hand, simple word-matching models prove to be very effective and interpretable for tasks that are so simple that complex models tend to overlook their differences. In doing so, we motivate the need for considering multiple representations rather than adopting a one-size-fits-all approach. Finally, to promote open and reproducible research, our code is publicly available, alongside word vectors trained on a very large corpus of clinical notes.

Chapter 5

Representations for Predicting Outcomes Across Changing EHR Systems

Existing machine learning methods typically assume consistency in how semantically equivalent information is encoded. However, the way information is recorded in databases differs across institutions and over time, often rendering potentially useful data obsolescent. As machine learning methods are more widely adopted in healthcare, mitigating this erroneous assumption will become critical.

The following chapter presents work that appeared at KDD and was done in collaboration with Jen Gong, Peter Szolovits, and John Guttag. In it, we map database-specific representations of information to a shared set of semantic concepts, thus allowing models to be built from or transition across different databases.

We demonstrate our method on machine learning models developed in a healthcare setting. In particular, we evaluate our method using two different intensive care unit (ICU) databases and on two clinically relevant tasks, in-hospital mortality and prolonged length of stay. For both outcomes, a feature representation mapping EHR-specific events to a shared set of clinical concepts yields better results than using EHR-specific events alone.

5.1 Overview

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. Mining EHRs enables the development of risk models on retrospective data and their application in real-time for clinical decision support. Such models facilitate improving outcomes while lowering costs. However, this task is complicated by the fact that EHRs are constantly changing—utilizing new variables, definitions, and methods of data entry. Furthermore, EHR versions often differ across institutions, and even in different departments within the same institution.

While specification changes can appear minor, each difference means that a risk model developed on a prior version may depend on variables that no longer exist or are defined differently in the current version. For example, the Society for Thoracic Surgeons' Adult Cardiac Surgery Database has undergone many transitions since its introduction in 1989 [73]. 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 [86].

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 [25, 3]. 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 [98]. The alternative, building version-specific models, is prohibitively labor intensive and creates a

problem during transition periods, when there are insufficient data 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 enables building models that can leverage data from another database without restricting the data to a subset or requiring database integration, a difficult problem [23, 34].

We present a case study on the structured data in the Medical Information Mart for Intensive Care (MIMIC-III) [48], which is described in Section 2.2.2. This data spans two EHR versions, CareVue (2001–2008) and MetaVision (2008–2012). There are approximately 9,000 items specific to CareVue and approximately 3,000 items specific to the MetaVision data.

In this case study, we relate EHR-specific data to clinical concepts from the Unified Medical Language System (UMLS) [9], a collection of medical ontologies. An ontology consists of a set of concepts (*entities*), and *relations* between entities. Although general domain ontologies (e.g., [10]) and tools for identifying equivalent semantic concepts (e.g., [28]) 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 two epochs of EHR systems in the publicly available MIMIC-

III dataset.

5.2 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 [78, 68, 89].

Sun [89] 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, Carroll et al. [12] 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 from the EHR.

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, Wiens et al. [99] 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 [99], 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 [75], we utilize a domain-specific vocabulary encoded through expert knowledge.

5.3 Method

In this section, we describe a feature representation that captures the EHR encodings (Section 5.3.1). Next, we describe the EHR-specific feature representation for each patient (Section 5.3.2), and then the conversion of this representation to the UMLS concepts, called *concept unique identifiers* (*CUIs*) (Section 5.3.3).

5.3.1 Bag-of-Events Feature Representation

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.

To that end, we consider a feature space that 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 [96, 94, 17].

The BOE features capture the different kinds of events encoded in the EHR systems. While using the values of lab tests or vital signs would certainly lead to improved predictive performance [62, 63, 31], 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.

5.3.2 EHR Item ID Feature Construction

We construct features from the EHRs to reflect the clinical events that occurred. In the MIMIC-III database, events are defined by an ID, an associated description, and a text or numerical value. While numerical values capture measurements of patient state, text values often add to the semantic meaning of the events. Because of this, we assign new identifiers for each unique (ID, text value) pair. These new unique identifiers are referred to as *Item IDs* in the rest of the chapter.

Figure 5-1 shows an example. In MIMIC-III, the ID 229 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. Events recorded in the chart contain many unique values associated with this ID, indicating the sites where the line could have been placed. For example, the text "PA Line" indicates a pulmonary arterial line, which has very different clinical implications than a "peripherally inserted central catheter" invasive line.

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.

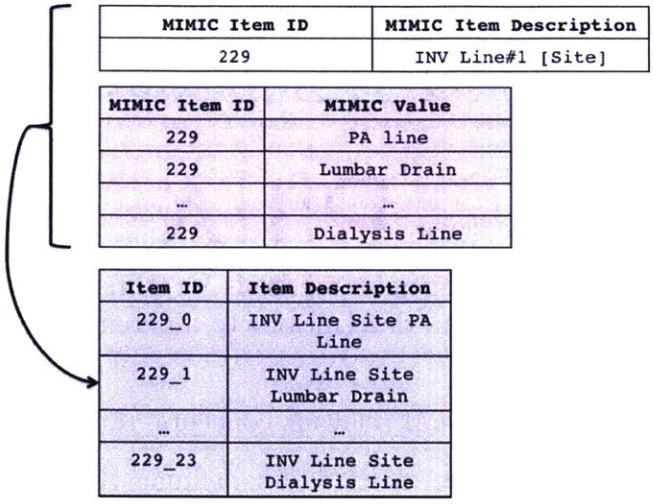


Figure 5-1: Text values often modify the semantic meaning of the corresponding items. We assign new unique item IDs with item descriptions that append these values to the initial item description. In this example, ID 229 in MIMIC is associated with a number of distinct text values in patients' charts that modify its semantic meaning.

5.3.3 Mapping EHR Item ID to UMLS Concept Unique Identifier

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 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 [84]. 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

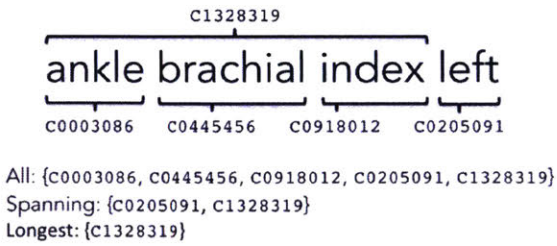


Figure 5-2: **All**, **Spanning**, and **Longest** methods for annotating "ankle brachial index left." These approaches relate the item descriptions to different sets of CUIs.

Item IDs to CUIs often identifies several relevant concepts, rather than a single one.

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 [22]. 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 5-2). 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 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 5-3). 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 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 5-3 shows the distribution of number of CUIs per Item ID for the different mapping

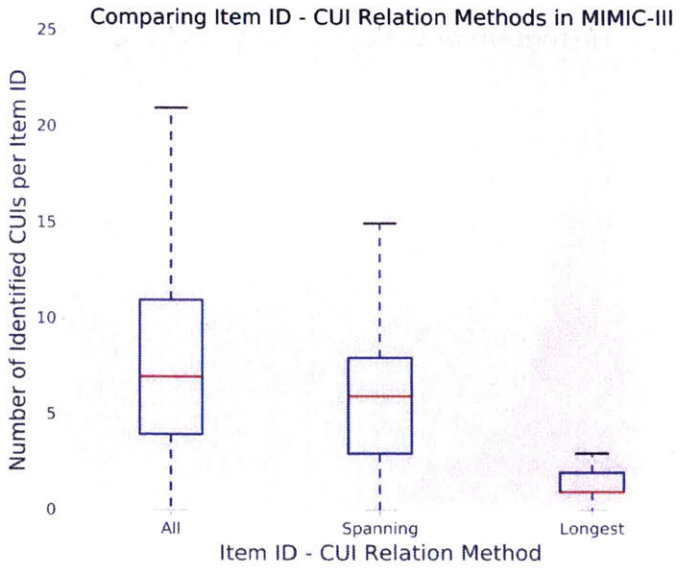


Figure 5-3: Distribution of number of identified CUIs per Item ID: Comparing *All*, *Spanning*, and *Longest* relation methods.

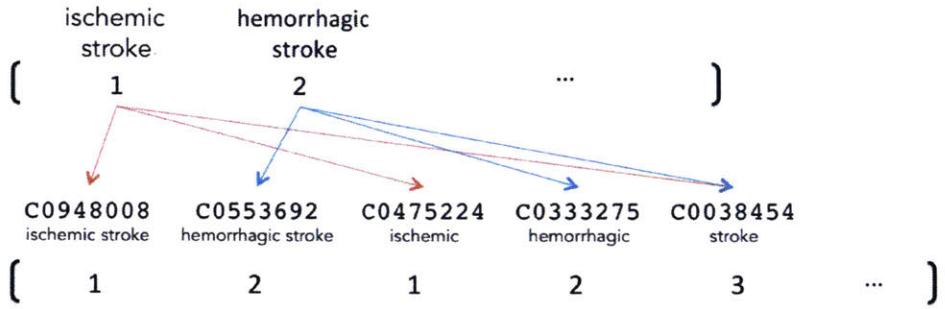


Figure 5-4: Transformation of Item IDs BOE representation to CUIs BOE representation using the *all* method.

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 5.5.2, 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 5-4 depicts an example of this conversion using all CUIs.

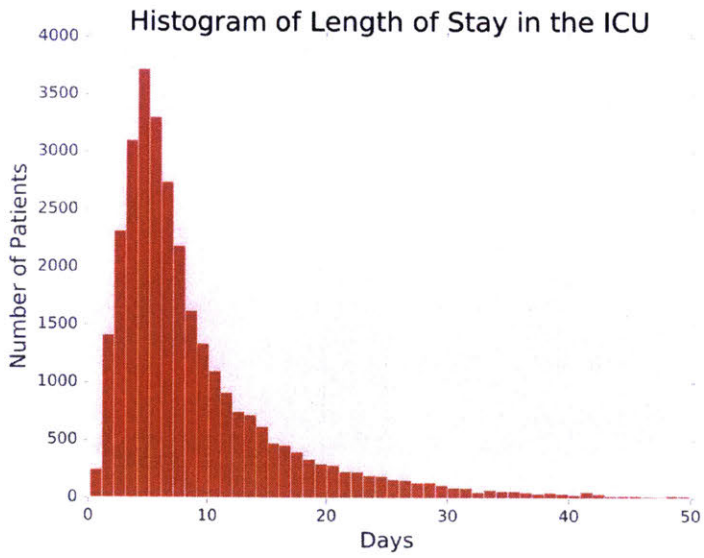


Figure 5-5: Length of stay in the ICU in MIMIC-III. Outliers (LOS > 50 days) truncated for clarity of visualization.

5.4 Experimental Setup

In these experiments,[1] 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.

5.4.1 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

[1] Code available at https://github.com/mit-ddig/event-cui-transfer.

Table 5.1: Number of patients and clinical outcomes (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%)

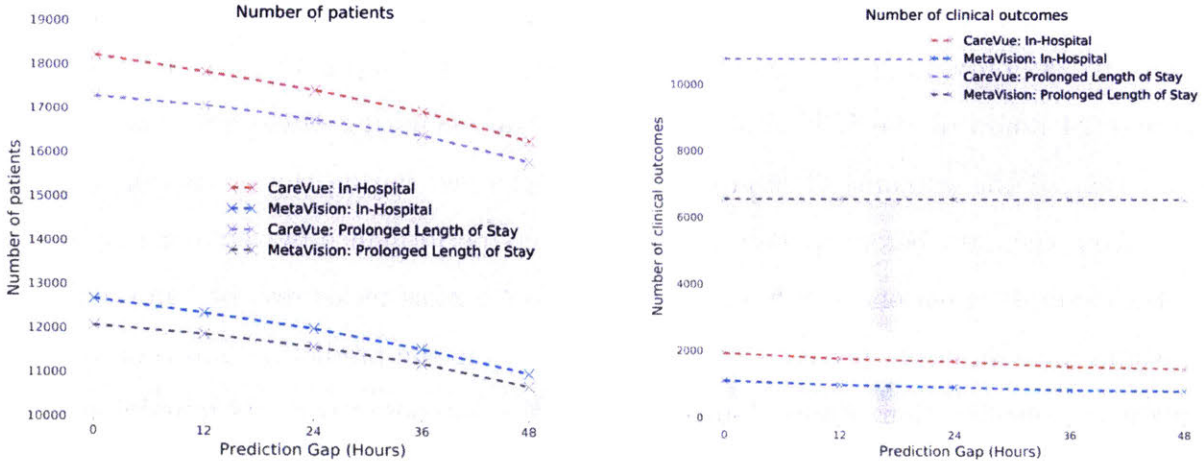


Figure 5-6: Number of patients remaining in the ICU (left) and clinical outcomes (right) with prediction gap 0–48 hours.

information from this time period [62].

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 the hospital. We define prolonged LOS in the ICU as a stay exceeding the upper quartile (> 11.3 days). Figure 5-5 shows the distribution of length of stay across the patients in the ICU. Table 5.1 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

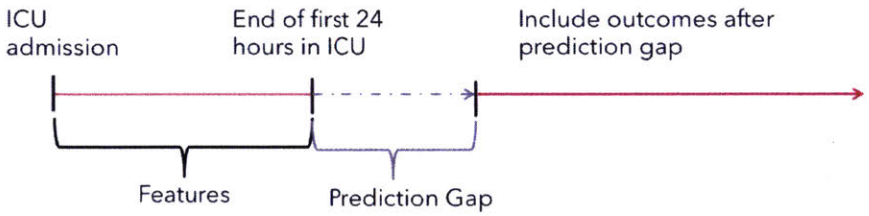


Figure 5-7: Diagram of relationship between information used to construct feature vector (first 24 hours in the ICU) and prediction gap between information used and outcomes.

the 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 5-6 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 5-7) for both CareVue and MetaVision.

5.4.2 Model Definition

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

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_+ \sum_{i:y_i=+1} \log\left(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}\right) + C_- \sum_{i:y_i=-1} \log\left(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}\right) \tag{5.1}$$

We used the scikit-learn LIBLINEAR implementation to train and test all models [76, 27]. 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 fea-

tures. We employed L2-regularization to reduce the risk of overfitting, since our data are small relative to the data dimensionality (see Table 5.1).

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.

5.5 Experimental Results

5.5.1 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) [62], a well-established acuity score that is commonly used as a baseline when developing risk models for mortality in the ICU [31, 32, 44] and also uses information 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 5-8). We used the Wilcoxon signed-rank test [100] to evaluate significance of the differences between the Item IDs-only results and the SAPS II + Item IDs results. 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

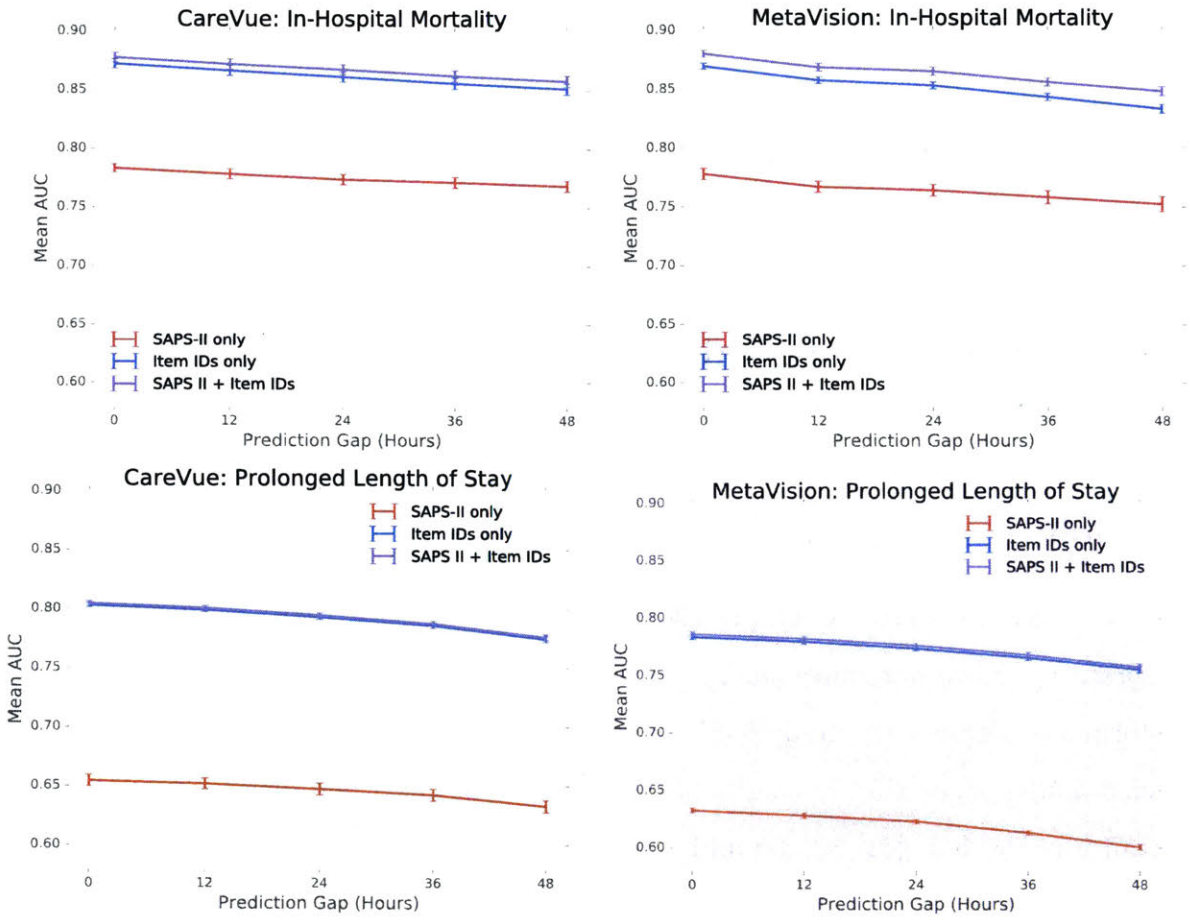


Figure 5-8: Mean AUC across 10 2:1 stratified holdout sets and 95% confidence interval shown for each database and outcome considered. Item IDs + SAPS II (purple) significantly outperforms Item IDs-only (blue) or SAPS II only (red) in predicting in-hospital mortality (top) and prolonged LOS (bottom) in CareVue (left) and MetaVision (right).

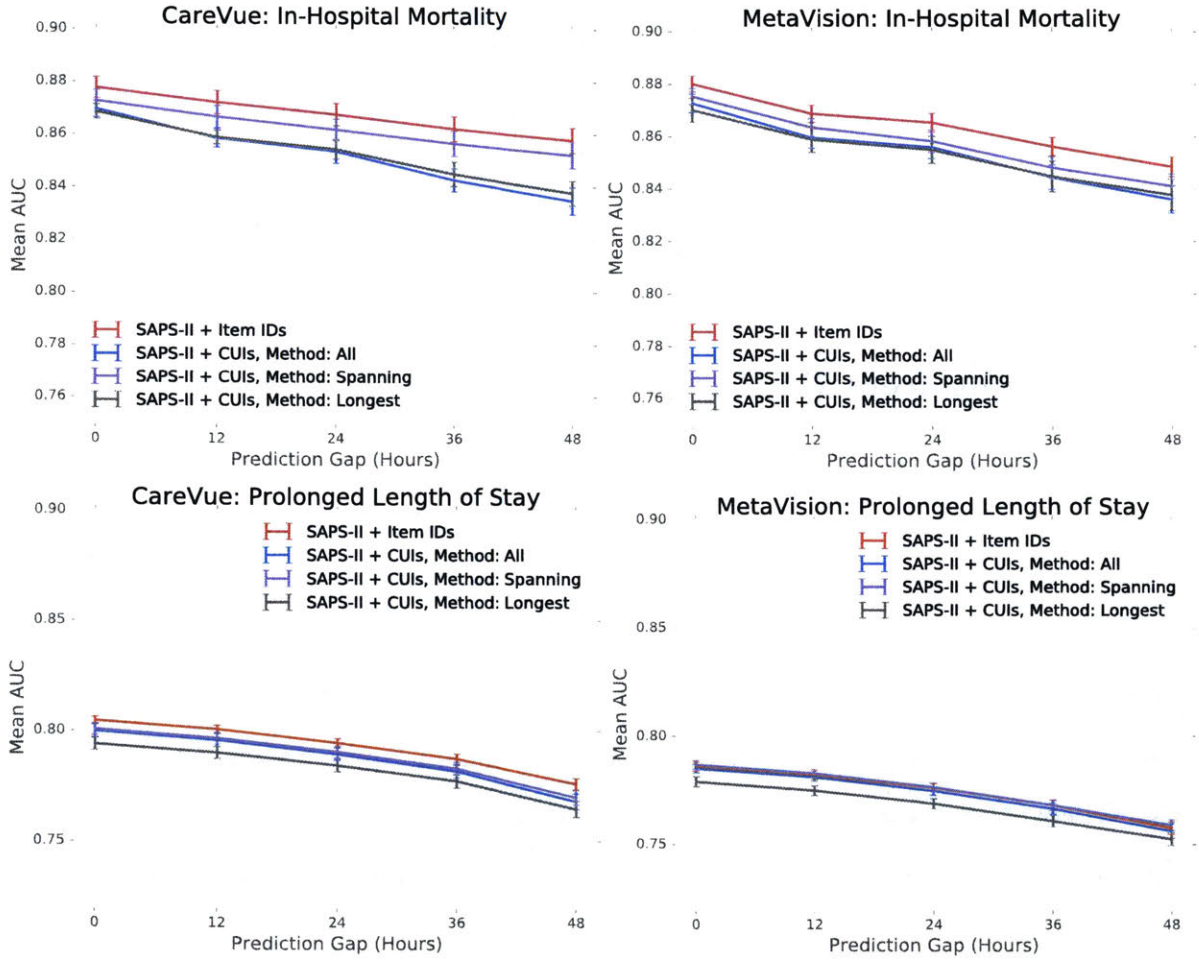


Figure 5-9: Mean AUC across 10 2:1 stratified holdout sets and 95% confidence interval shown for each database and outcome considered. Converting to CUIs from Item IDs results in small, but statistically significant differences in performance in 3 out of the 4 tasks considered. Mean AUC across prediction gaps shown for the outcomes of in-hospital mortality (top) and prolonged LOS (bottom) in CareVue (left) and MetaVision (right).

Table 5.2: 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**

+ BOE (Item IDs or CUIs) feature space.

5.5.2 Mapping Item IDs to CUIs

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 5.3.3.

The mean AUCs across 10 2:1 stratified training:holdout splits are shown in Figure 5-9, and the Wilcoxon sign-rank test p-values for in-hospital mortality and prolonged length of stay are shown in Table 5.2 and Table 5.3, respectively. The mean differences in AUCs across all the prediction gaps were statistically significant for the outcome of in-hospital mortality in CareVue and MetaVision, as well as the outcome of prolonged length of stay in CareVue (p-value = 0.0051). However, they are small in magnitude (Δ AUC $\leq$ 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 very small in magnitude. This small difference shows that representing clinical events using CUIs can still achieve high predictive performance on predicting mortality in the ICU within a *single* EHR system.

As Figure 5-9 shows, the *spanning* method appears to have improved or comparable

Table 5.3: 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

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 5.4 shows the number of item IDs in each EHR version and the resulting number of CUIs from the cTAKES mapping using the *spanning* approach.

5.5.3 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 5-10(a)), and 2) training on the *shared* set of Item IDs between Train DB and Test DB (Figure 5-10(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 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 5-11. 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.

Using the UMLS CUIs, we increase the AUC on in-hospital mortality by at least 0.01

Table 5.4: 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

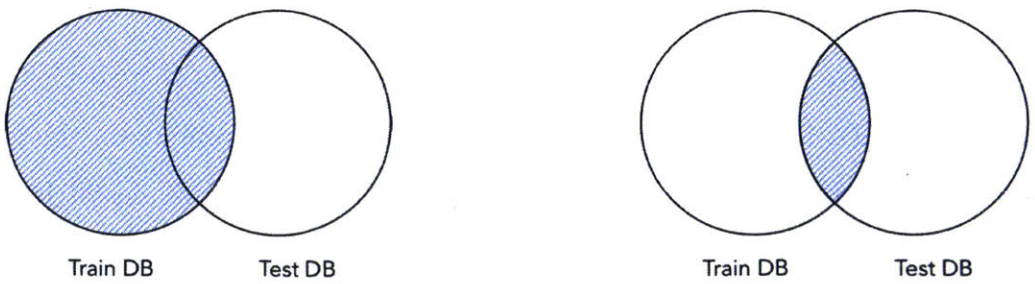


Figure 5-10: Baseline approaches: (a) Train a model on *all* items in the training database (Train DB) (left), and (b) Train a model only on *shared* items that appear in both the training and test databases (right).

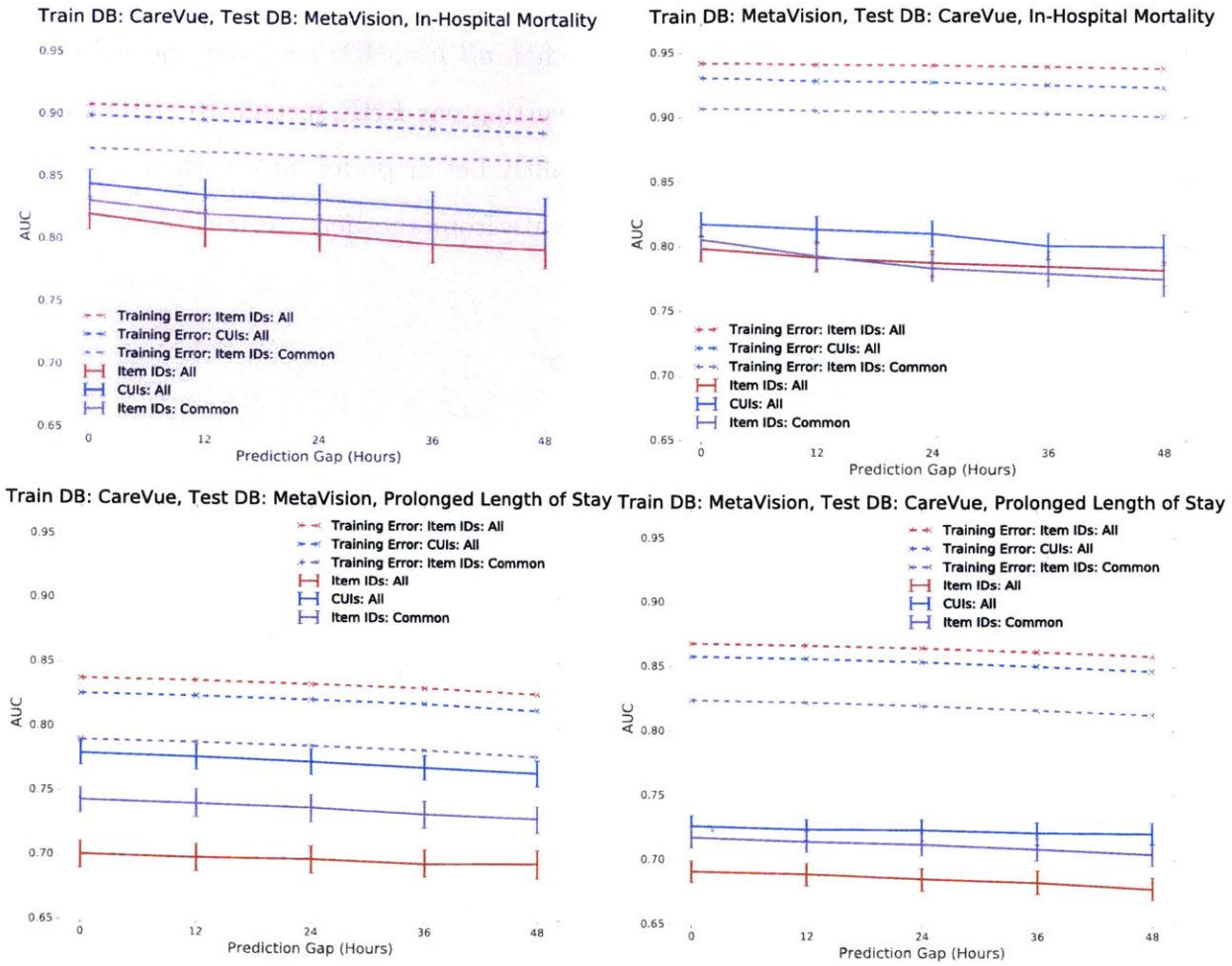


Figure 5-11: AUC when training on TrainDB and testing on TestDB using EHR-specific Item IDs (all), Item IDs (shared), and CUIs. 95% confidence intervals are shown for each database and outcome considered. The dashed lines show the training AUC of each model on Train DB, while the solid lines show the 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).

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 even 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. Thus, 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.

5.6 Conclusion and Discussion

We introduce 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 demonstrate the utility of our approach for constructing risk models for patients in the intensive care unit. We leverage 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.

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.

Second, 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.

Although we demonstrate the utility of this method in a clinical setting, entity resolution for database matching is a common problem. As databases in finance, government, and other sectors evolve and data analytics gains traction, resolving changes in information recording over time is an important task.

Chapter 6

Conclusion

This thesis provides several case studies in which leveraging representations of text is important for clinical predictive tasks. Although not exhaustive, these case studies reflect contributions toward the broader goal of using machine learning to improve healthcare; specifically, our results motivate the future use of clinical text, and suggest that no single representation will suffice. Further, we affirm our belief that free-text data is critical to delivering on the full potential offered by EHR data.

In Chapter 3, we saw that as EHRs contain an increasingly large amount of data, standard approaches to hospital mortality prediction can be improved by incorporating information from clinical narratives. Specifically, we demonstrated that standard acuity features tend to become less predictive over time due to decreasing data that can be used for support, and that this effect can be offset by including topic-based features derived from clinical notes. Paramount to this observation is that while patient support decreases over time, the volume of notes about each of the remaining patients increases over time. While this is true about other patient-collected data as well, the notes become increasingly informative about patient state.

Further, the case study presented in Chapter 3 suggests that notes contain complementary information; thus motivating the need for their inclusion in order to obtain the highest performing predictive models. This notion appeals to intuition since we would expect that

clinical notes both provide a summary of important events recorded by other instrumented signals and contain observations that cannot be otherwise recorded (i.e., subjective observations).

In Chapter 4, we saw that many simple NLP models do well on a variety of tasks, but notably that no one model performs best across all tasks. The first of these findings is unsurprising, since simple machine learning methods often perform better than their more complex counterparts when sufficient data are available [38, 2]. Instead, a more satisfying realization can be discovered in this observation; namely, we may be approaching a state where sufficient clinical data are available that simple methods perform well. This is important because, historically, barriers to accessing clinical data have often impeded machine learning methods, particularly when data include clinical notes that are laden with protected health information.

The second finding, that no single machine learning method always performs best, is initially surprising, but ultimately appeals to our intuition. In essence we considered the prediction of intrinsic note information in order to "stress test" several common note representations, showing that complex models outperform simples ones on reasoning tasks, but struggle to capture seemingly "easy" information. Conversely, the simplest word-matching models prove to be very effective and interpretable for tasks where complex models tend to overlook their differences. In this finding we can see a likeness to the differences between extractive and abstractive summarization. Specifically, our "easy" tasks benefit from methods that show extractive properties since the information we are predicting is able to be found directly in the text; whereas the more difficult tasks require methods that are abstractive insofar as they are able to reconcile and synthesize information that may not appear directly in the text.

In Chapter 5, we introduced an approach to constructing machine learning models that are portable across different representations of semantically similar information. Our method leveraged domain knowledge from the UMLS in order in order to transfer a model from one EHR to another. We expect this to occur increasingly frequently as EHRs are replaced,

resulting in periods of time during which there are insufficient data to learn useful predictive models. As such, our method may grow in importance since it facilitates the use of models built using the previous EHR system.

When considering the impact of our work, we must also consider its limitations. Rather than doing so comprehensively, we highlight the two considerations that are most important. First, we chose to work with MIMIC [81, 48] because it is an open, publicly-accessible database, allowing us to conduct case studies that are reproducible. Applying our methods to other institutions could lend insight to how well our approaches perform in the presence of different care staff, practices, and patient population characteristics, as well as differences in EHR systems. We believe the reproduction of results at multiple sites is crucial in establishing the validity of methods, and has proven difficult to do [46].

Second, our work presents methodologies that serve as steps in the direction of using machine learning to improve healthcare. However, improving healthcare will require changes to the provision of care; thus, the impact of our work is indirect. Indeed, change will require not just this work, but many others to incrementally improve the state of healthcare. As a community, we should embrace the many research opportunities this affords. Healthcare demands improvements in replacing mundane clinical tasks, facilitating better decision making, and providing tools to advance clinical practice.

Bibliography

[1] C.W. Arnold et al. Clinical case-based retrieval using latent topic analysis. In *AMIA Annual Symposium Proceedings*, volume 2010, page 26. AMIA, 2010.

[2] Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 26–33. Association for Computational Linguistics, 2001.

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

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3(5):993–1022, 2003.

[5] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[6] W. Boag, K. Wacome, T. Naumann, and A. Rumshisky. CliNER: A lightweight tool for clinical named entity recognition. In *AMIA Joint Summits on Clinical Research Informatics*, 2015.

[7] William Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. WhatâĂŹs in a note? unpacking predictive value in clinical note representations. 2018.

[8] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[9] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[10] 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.

[11] 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.

[12] 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. *J Am Med Inform Assoc*, 19(e1):e162–e169, 2012.

[13] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

[14] Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. Bidirectional LSTM-CRF for clinical concept extraction. *arXiv preprint arXiv:1611.08373*, 2016.

[15] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2:27:1–27:27, 2011.

[16] Dustin Charles, Meghan Gabriel, and Michael F Furukawa. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2012. *ONC data brief*, 9:1–9, 2013.

[17] 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.

[18] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[19] Raphael Cohen, Iddo Aviram, Michael Elhadad, and Noémie Elhadad. Redundancy-aware topic modeling for patient record notes. *PloS one*, 9(2):e87555, 2014.

[20] B. deBruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18:557âĂŞ562, 2011.

[21] Frank Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association : JAMIA*, 24 3:596–606, 2017.

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

[23] X. L. Dong and D. Srivastava. Big data integration. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1245–1248, April 2013. doi: 10.1109/ICDE.2013.6544914.

[24] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133 (1):e54–e63, 2014.

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

[26] Anne Elixhauser, Claudia Steiner, D Robert Harris, and Rosanna M Coffey. Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27, 1998.

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

[28] 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.

[29] Xiao Fu and Sophia Ananiadou. Improving the extraction of clinical concepts from clinical records. *Proceedings of BioTxtM14*, 2014.

[30] Marzyeh Ghassemi, Tristan Naumann, Rohit Joshi, and Anna Rumshisky. Topic models for mortality modeling in intensive care units. In *Proceedings of ICML 2012 (Machine Learning for Clinical Data Analysis Workshop), Poster Presentation*, Edinburgh, UK, June 2012.

[31] 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.

[32] Marzyeh Ghassemi, Marco A F 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 *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, page 446. NIH Public Access, 2015.

[33] Jeffrey Alan Golden. Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen. *Jama*, 318(22):2184–2186, 2017.

[34] 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.

[35] 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.

[36] T. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*, volume 101, pages 5228–5235, 2004.

[37] Paulina Grnarova, Florian Schmidt, Stephanie Hyland, and Carsten Eickhoff. Neural document embeddings for intensive care patient mortality prediction. In *NIPS 2016 Workshop on Machine Learning for Health Workshop*, 2016.

[38] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[39] Neil A Halpern and Stephen M Pastores. Critical care medicine in the united states 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical care medicine*, 38(1):65–71, 2010.

[40] Neil A Halpern, Stephen M Pastores, and Robert J Greenstein. Critical care medicine in the united states 1985–2000: an analysis of bed numbers, use, and costs. *Critical care medicine*, 32(6):1254–1259, 2004.

[41] Robert E Hirschtick. Copy-and-paste. *Jama*, 295(20):2335–2336, 2006.

[42] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural computation*, 8:1735–1780, 1997.

[43] Benjamin D Horne, Heidi T May, Joseph B Muhlestein, Brianna S Ronnow, Donald L Lappé, Dale G Renlund, Abdallah G Kfoury, John F Carlquist, Patrick W Fisher, Robert R Pearson, et al. Exceptional mortality prediction by risk scores from common laboratory tests. *The American journal of medicine*, 122(6):550–558, 2009.

[44] Caleb W Hug and Peter Szolovits. Icu acuity: real-time models versus daily models. In *AMIA*, 2009.

[45] Thanh N Huynh, Eric C Kleerup, Joshua F Wiley, Terrance D Savitsky, Diana Guse, Bryan J Garber, and Neil S Wenger. The frequency and cost of treatment perceived to be futile in critical care. *JAMA internal medicine*, 173(20):1887–1894, 2013.

[46] Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. Reproducibility in critical care: a mortality prediction case study. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 361–376, Boston, Massachusetts, 18–19 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v68/johnson17a.html`.

[47] Alistair EW Johnson, Andrew A Kramer, and Gari D Clifford. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy*. *Critical care medicine*, 41(7):1711–1718, 2013.

[48] 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.

[49] Cathy Jones. Glasgow coma scale., 1979.

[50] Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics*, 45(1):129–140, 2012.

[51] R L Kane, T A Shamliyan, C Mueller, S Duval, and T J Wilt. The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. *Medical Care*, 45(12):1195–1204, December 2007.

[52] Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, page ocx019, 2017.

[53] Mark T Keegan, Ognjen Gajic, and Bekele Afessa. Severity of illness scoring systems in the intensive care unit. *Critical care medicine*, 39(1):163–169, 2011.

[54] W A Knaus, J E Zimmerman, D P Wagner, E A Draper, and D E Lawrence. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical Care Medicine*, 9:591–597, 1981.

[55] William A Knaus. Apache 1978-2001: the development of a quality assurance system based on prognosis: milestones and personal reflections. *Archives of Surgery*, 137(1): 37–41, 2002.

[56] William A Knaus, DP Wagner, EA e a1 Draper, JE Zimmerman, Marilyn Bergner, P Gl Bastos, CA Sirio, DJ Murphy, T Lotring, and A Damiano. The apache iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. *CHEST Journal*, 100(6):1619–1636, 1991.

[57] Bart Kosko. Bidirectional associative memories. *IEEE Trans. Syst. Man Cybern.*, 18 (1):49–60, January 1988.

[58] Giovanni Landoni, Marco Comis, Massimiliano Conte, Gabriele Finco, Marta Mucchetti, Gianluca Paternoster, Antonio Pisano, Laura Ruggeri, Gabriele Alvaro, Manuela Angelone, et al. Mortality in multicenter critical care trials: an analysis of interventions with a significant effect. *Critical care medicine*, 43(8):1559–1568, 2015.

[59] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

[60] J R Le Gall, P Loirat, F Nicolas, C Granthil, F Wattel, R Thomas, P Glaser, P Mercier, J Latournerie, P Candau, et al. Use of a severity index in 8 multidisciplinary resuscitation centers. *Presse medicale (Paris, France: 1983)*, 12(28):1757–1761, 1983.

[61] Jean Roger Le Gall, Janelle Klar, Stanley Lemeshow, Fabienne Saulnier, Corinne Alberti, Antonio Artigas, and Daniel Teres. The logistic organ dysfunction system: a new way to assess organ dysfunction in the intensive care unit. *Jama*, 276(10):802–810, 1996.

[62] J.R. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, 270(24): 2957–2963, 1993.

[63] 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.

[64] Li-Wei H Lehman, Mohammed Saeed, William J Long, Joon Lee, and Roger G Mark. Risk stratification of icu patients using topic models inferred from unstructured progress notes. In *AMIA*. Citeseer, 2012.

[65] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. In *Transactions of the Association for Computational Linguistics*, pages 211–225, 2015.

[66] Yen-Fu Luo and Anna Rumshisky. Interpretable topic features for post-icu mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 827. American Medical Informatics Association, 2016.

[67] 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.

[68] 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.

[69] J Michael McGinnis, Leigh Stuckhardt, Robert Saunders, Mark Smith, et al. *Best care at lower cost: the path to continuously learning health care in America.* National Academies Press, 2013.

[70] Wang MD, Khanna R, and Najafi N. Characterizing the source of text in electronic health record progress notes. *JAMA Internal Medicine*, 177(8):1212–1213, 2017. doi: 10.1001/jamainternmed.2017.1548. URL `+http://dx.doi.org/10.1001/jamainternmed.2017.1548`.

[71] Edward J Mills, Kristian Thorlund, and John PA Ioannidis. Demystifying trial networks and network meta-analysis. *Bmj*, 346:f2914, 2013.

[72] Office of the National Coordinator for Health Information Technology. Office-based physician electronic health record adoption. *Health IT Quick-Stat 50*, 2016.

[73] Society of Thoracic Surgeons. Society of Thoracic Surgeons National Database, 2016. URL `https://www.sts.org/national-database`.

[74] Gustavo A Ospina-Tascón, Gustavo Luiz Büchele, and Jean-Louis Vincent. Multicenter, randomized, controlled trials evaluating mortality in intensive care: Doomed to fail? *Critical care medicine*, 36(4):1311–1322, 2008.

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

[76] 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.

[77] 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.

[78] 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. *J Biomed Inform*, 45(4):689–696, 2012.

[79] K. Roberts and S. Harabagiu. A flexible framework for deriving assertions from electronic medical records. *J Am Med Inform Assoc*, 18:568âĂŞ573, 2011.

[80] A Rumshisky, M Ghassemi, T Naumann, P Szolovits, VM Castro, TH McCoy, and RH Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, 6(10):e921, 2016.

[81] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[82] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

[83] Suchi Saria, Gayle McElvain, Anand K Rajani, Anna A Penn, and Daphne L Koller. Combining structured and free-text data for automatic coding of patient outcomes. In *AMIA Annual Symposium Proceedings*, volume 2010, page 712. American Medical Informatics Association, 2010.

[84] 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.

[85] George Siontis, Ioanna Tzoulaki, and John Ioannidis. Predicting death: an empirical evaluation of predictive tools for mortality. *Archives of internal medicine*, pages archinternmed–2011, 2011.

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

[87] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. Clamp âĂŞ a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, page ocx132, 2017. doi: 10.1093/jamia/ocx132. URL `+http://dx.doi.org/10.1093/jamia/ocx132`.

[88] K Strand and H Flaatten. Severity scoring in the icu: a review. *Acta Anaesthesiologica Scandinavica*, 52(4):467–478, 2008.

[89] Yao Sun. Methods for automated concept mapping between medical databases. *J Biomed Inform*, 37(3):162–178, 2004.

[90] Justin Travers, Suzanne Marsh, Mathew Williams, Mark Weatherall, Brent Caldwell, Philippa Shirtcliffe, Sarah Aldington, and Richard Beasley. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax*, 62(3):219–223, 2007.

[91] Inigo Jauregi Unanue, Ehsan Zare Borzeshi, and Massimo Piccardi. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *CoRR*, abs/1706.09569, 2017. URL `http://arxiv.org/abs/1706.09569`.

[92] US Centers for Disease Control and Prevention. Health disparities in hiv/aids, viral hepatitis, stds, and tb. https://www.cdc.gov/nchhstp/healthdisparities/africanamericans.html. Accessed September 26, 2017.

[93] O. Uzuner, B. South, S. Shen, and S. DuVal. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. volume 18, page 552âĂŞ6, 2011.

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

[95] J-L Vincent, Rui Moreno, Jukka Takala, S Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PM Suter, and LG Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, 22 (7):707–710, 1996.

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

[97] Griffin M Weber, Kenneth D Mandl, and Isaac S Kohane. Finding the missing link for big biomedical data. *Jama*, 311(24):2479–2480, 2014.

[98] 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.

[99] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J Am Med Inform Assoc*, 0:1–8, 2014.

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

[101] Jesse O Wrenn, Daniel M Stein, Suzanne Bakken, and Peter D Stetson. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53, 2010.

[102] Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. A study of neural word embeddings for named entity recognition in clinical text. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2015:1326âĂŤ1333, 2015. ISSN 1942-597X. URL `http://europepmc.org/articles/PMC4765694`.

[103] Thomas R Yackel and Peter J Embi. Copy-and-paste-and-paste. *JAMA*, 296(19): 2315–2316, 2006.

[104] J E Zimmerman and A A Kramer. Outcome prediction in critical care: the Acute Physiology and Chronic Health Evaluation models. *Current Opinion in Critical Care*, 14:491–497, 2008.