Learning Clinical Data Representations for Machine Learning

By

Lina Sulieman

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

December 15, 2018

Nashville, Tennessee

Approved:

Daniel Fabbri, Ph.D

Bradley Malin, Ph.D

Tom Lasko, M.D., Ph.D

Colin Walsh, M.D., MS

Christopher Fonnesbeck, Ph.D

DEDICATION

To my Mother, my eternal source of inspiration and strength

ACKNOWLEDGMENT

The Ph.D. journey has been a life-changing experience. It is not only the keyboard strokes that made it possible, but it is also a journey that cannot be completed without the help of others. First and foremost, I cannot begin to express my thanks to advisor and mentor Dr. Daniel Fabbri, who without his support and nurturing, I would not be here. Over the past four years, he guided and taught me the skills that will be in the toolkit in my future career. He encouraged and fed my curiosity for learning and trying new ideas regardless of how many time I crashed his servers or drained his pen on my papers. I would be not be writing this dissertation without his financial and educational support.

I would also like to extend my deepest gratitude to Dr. Bradley Malin, my first mentor, and my Master's thesis advisor. He has been and will be the most enthusiastic teacher. He always finds time to teach and steer me in the right direction. I would like to express my most profound appreciation to my committee: Dr. Tom Lasko who taught me indirectly the machine learning cornerstones, Dr. Colin Walsh who always reshapes my thinking about the clinical aspect of any problem or idea that I pitch for him, and Chris Fonnesbeck who I was very fortunate to have on my committee.

I would like to thank others who were invaluable sources of assistance. I would like to thank Joseph Coco for his tremendous help in my experiments, the core of my dissertation. I would like to thank Kevin Johnson and Mark Frisse for creating an excellent training environment that shaped my perspective on biomedical informatics in formal and informal discussions.

I cannot begin to express my gratitude to the Department of Biomedical Informatics, who's endless support and unconditional belief in me made the completion of my dissertation possible. I especially want to thank Cindy Gadd for creating the opportunity to become a trainee six years ago, and Gretchen Jackson for taking every chance she has to mentor and explore possible options in my career. Special thanks to Rischelle Jenkins who makes everything possible even the hardest and effortful milestones in my Ph.D. training.

I would like to thank my fellow trainees, past and present for their incredible friendship, fun coffee breaks, and insightful discussion. I would like to extend my sincere thanks to my friends

outside DBMI. Our hikes, road trips, and long hours of studying and working made my journey full of joy and fun.

Finally, I am deeply indebted to my family. My mother, whose love and support made me who I am today. I am the luckiest person in the world that you are my role model, my inspiration, and my mother. I would like to thank my sisters and brother for being there for me, no matter what. I would like to thank my father. Although he is no longer with us, he always believed in my educational achievements. If he were among us today, he would be proud of this achievement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EMR	Electronic medical records
CDS	Clinical decision support
ICU	Intensive care units
CHF	Congestive heart failure
AMI	Acute myocardial infarction
NLP	Natural language processing
NLM	National library of medicine
UMLS	Unified medical language system
TF-IDF	Term frequency-inverse document frequency
AMIA	American medical informatics association
JAMIA	Journal of American medical informatics association
ICD	International classification of diseases
MDL	Minimum descriptive length-based
ADE	Adverse drug events
PCA	Principal component analysis
ADR	Adenoma detection rate
SVM	Support vector machine
HMM	Hidden Markov model
CRF	Conditional random fields
POS	Part of the speech
VUMC	Vanderbilt University Medical Center
NN	Neural network
ReLU	Rectified linear unit
Tanh	Hyperbolic tangent function
CNN	Convolutional neural networks
RNN	Recurrent Neural Networks
LSTM	Long short-term memory
CBOW	Continuous Bag-of-Words model
OW	Observation window
PW	Prediction window
PP	Prediction point
ADM	Before-admission model
DAM	During-admission model
ADM	At-discharge model
BDAM	Before, during admission, and after discharge model
LOS	Length of stays
CO_2	Carbon dioxide
PCV	Packed cell volume
PTT	Partial thromboplastin time
LDA	Latent Dirichlet Allocation
AUC	Area under the receiver operator characteristic curves
CUI	Concept unique identifiers

STY	Semantic types
MHAV	My Health At Vanderbilt
SD	Synthetic derivative
NER	Name entity recognition
DNN	Deep neural network
RF	Random forest
LR	Logistic regression
BoW	Bag of words
CPT	Current procedural terminology
PPV	Positive predictive values
CDF	Cumulative distribution function

CHAPTER 1 INTRODUCTION

Machine learning in healthcare

The adoption of electronic medical records (EMR) increases secondary usage of clinical data and encourages researchers to mine EMR data to extract facts and relations. The discovered medical knowledge can improve the care experience for patients, enhance knowledge about a disease and a treatment, and expand our capacity for analyzing the effectiveness and efficiency of healthcare systems [1,2]. On the population level, aggregated data facilitates bio-surveillance of epidemics and diseases, tracks health disparities, and identifies societal factors affecting population health [3]. On the patient level, EMR-based clinical research enhances phenotyping methodologies, discovers the effects of treatments, performs risk stratification, and personalizes treatment based on the patient genetics.

EMR contain invaluable information that can be leveraged to improve patient care. For example, clinical decision support (CDS) systems are designed to aid health professionals in decision-making. CDS systems retrieve patients' data and match them to a computerized knowledge base to actuate efficient and effective recommendations [4–6]. CDS systems are activated when clinical facts and observations are fed into them. For example, quick medical reference (QMR), developed by Miller, applies an algorithm to find a differential diagnosis using historical and physical findings and symptoms [7]. Researchers at Vanderbilt University Medical Center (VUMC) developed Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT), which is an advanced CDS system that identifies treatment options for patients using their genotypes [8].

In the last decade, machine learning and statistical models have been applied to recognize patterns and build predictive models in healthcare data. Risk stratification models have been developed that predict the risk of developing acute disease or encountering a negative outcome such as of stroke, mortality, or readmission [9–13]. In intensive care units (ICUs), machine learning models can identify deteriorating and critically ill patients, and predict the length of stay using vital signs, laboratory values, and medications [14,15]. In cancer, genotypes, treatments, and images are used to predict progression, recurrence, treatment response, and outcome [16–18].

Clinical data can be classified into two main categories based on their storage format: 1) structured data stored in a pre-defined format such as laboratory tests, medications, etc. 2) unstructured data that narrates the patient's treatment in the form of clinical documents. Clinical documents assemble clinical events that patients go through and communicate the patient story to the creator of the document and other providers in the same, or a different healthcare organization [19]. Clinical documents have been leveraged to predict outcomes [20]; extract information, assertions, relations, and risk factors [21–28]; and for classification [29–31]; deciphering sentiments and behaviors [20,32]; identifying phenotypes [33–35]; assigning diagnostic billing codes [36]; and monitoring clinical events such as drug adverse events and infections [36,37].

Challenges in secondary usage of EMR data

Although using EMR data in translational and clinical research can be cost-effective, the data are rarely suitable for ideal research usage. The variability in quality, messiness, and incompleteness of clinical data complicate the curation and implementation of data in translational research [38].

Clinical data, by nature, are messy and noisy. Noise in structured clinical data can imply a high variance or an irrelevant measurement. For instance, international classification of diseases (ICD) codes can be noisy because they are assigned as an inference to patients before the final definitive diagnosis, and are not revise later [39]. Similarly, noise in unstructured clinical data can be duplicated entries, irrelevant text, or poor grammar [40,41]. Noise and messiness can impair the data’s ability to provide influences in research [42]. Variations in data collection, patients, and data sources (healthcare providers, physiological resources, consumers, and patients) contribute to the aforementioned challenging characteristics of clinical data. Training models on noisy data could lead to overfitting, detecting incorrect patterns, and may produce inconsequential results [43,44]. Noisy text hinders pre-processing and effective semantic analysis [45].

Machine learning researchers handle noise in data by learning latent representations (e.g. Gaussian regression models or topic models), applying active learning, using rule-based models, polishing and correcting noisy data, training robust learners such as random forests, and eliminating possible noisy elements [42,44–48]. Training models with the noise assumption can reduce the effect of noise in the dataset. Holding the assumption of noisy data, Lasko inferred phenotypes from noisy sparse uric acid sequences by applying non-parametric Gaussian Process Regression [42]. Miotto *et al*. implemented a noise-masking algorithm to corrupt the input and introduce noise to data when training deep learning model [49].

Missing data is another challenge that faces machine learning researchers in the medical informatics field. Multiple factors contribute to the prevalence of this problem including: financial burdens that restrict some treatment plans, the health conditions of the patient, and variability in clinical practice that influences the order and documentation of medications and laboratory tests [50,51]. Missing data could bias the analysis and undermine the credibility of the trained model [52,53]. Omitting the entries with missing data is one approach, but this could also bias the analysis. Imputation is a statistical approach that infers missing values using methods such as mean, and multiple imputation using chained equations [51–53]. Multiple imputation has been used to estimate missing medications in [54]. Walsh *et al*. implemented additive regression, bootstrapping, and predictive mean matching to impute missing data before predicting the risk of suicide attempts [55].

Healthcare providers communicate and review plans of care in unstructured clinical documents [56]. The variability and flexibility in writing unstructured clinical documents increases the abstractions of clinical components for quality measurements, research, and assessing patients' health [19,56]. However, those two aspects of documentations encumber the extraction of clinical events from documents and creating text features and representation for machine learning models [56]. Text representations that account for semantics and syntax in the text improve the performance of text-mining algorithms and yield results similar to humans [36,57,58]. Phrases and n-gram features have been implemented to represent the context of the words, but did not outperform other approaches to text representation, such as bag of words [57]. Hu *et al*. improved the performance of traditional text clustering methods by building a concept thesaurus, using the semantic relations extracted from Wikipedia to impute missing semantics in the text representation [59].

At the patient level, clinical data are inserted into EMR and interpreted in a longitudinal format. Medications, laboratory tests, billing codes, and clinical documents are associated with a timestamp. Understanding the trajectories of some diseases is a challenge, as it requires performing an analysis across the dimension of time. Longitudinal analysis has been implemented in risk stratification, predicting diagnosis, and discovering phenotyping [42,60,61]. Using longitudinal measurements in prediction models remains an active research field [62]. The absence of longitudinal data could impact the accuracy of machine learning algorithms. Finding a data representation to summarize and retain sequential information can ease the challenge of building and training prediction models [63]. Using methods that account temporality can lead to a substantial improvement in the performance of the prediction model [63].

Clinical features representation in machine learning

The performance of machine learning methods depends on features representations [64,65]. Feature engineering combines human ingenuity and knowledge in creating representations [64]. Training machine learning models on mediocre data representations or outdated features can affect the model performance and its clinical validity. Failure to include pertinent information, informative representations, or changes in the patient's health (e.g. changes in hypertension status, complications after hip replacement) can under- or overestimate the risk of undesirable outcomes such as readmission or poor disease prognosis. For example, one study investigated patient hypertension deterioration over time, and found that the most predictive features of the hypertension status (i.e. normal versus out of control) were changes in hypertension status patterns before the prediction point [66]. Creating representations for hypertensive patients that embed the changes in features such as the hypertension status improved the performance of prediction model.

Problem statement and objectives

Clinical data are heterogeneous and multi-dimensional [67]. Challenges and shortcomings in training predictive models using structured and unstructured data still exist. To harness the power embodied in EMR data, clinical machine learning should be trained on informative feature representations. The objective of this dissertation is structuring complex and heterogeneous clinical data to develop machine learning models that could improve patients' care. The work addresses the temporality and semantics of clinical data and integrates those concepts in prediction models.

This dissertation presents methods to identify appropriate representations for structured and unstructured data that improve risk prediction, classification, and information extraction. The rest of the dissertation discusses: 1) the temporal aspect of structured data, and 2) the semantic representations for unstructured data.

The temporality of data and its effect on prediction model performance

The EMR is a rich repository for the longitudinal encounters and clinical pathways of patients and providers. The volume and variety of patient data are increasing; hence, manual analysis for a long sequence of data is impractical. Most risky and critical events, including complications, readmission, or mortality, are often preceded by warning signs or patterns. For instance, the diagnosis of congestive heart failure (CHF) is usually preceded by dyslipidemia, angina, and/or diabetes [68]. Evaluating the risk of developing CHF should incorporate the latest clinical information about the patient to obtain an accurate assessment.

Most clinical events and entities, such as medications, admissions, emergent visits, and diagnoses, follow a sequential, irregular, and auto-regressive nature. A clinical event often stimulates the occurrence of the next. A main clinical event causes the incidence of other clinical events in a treatment pathway. For instance, when a patient is admitted due to acute myocardial infarction (AMI), laboratory tests are ordered, procedures are performed, and medications are prescribed to treat the patient.

Irregularity in time between clinical events makes it challenging to bin the events and create a representation for machine learning models. Depending on the type of disease or the event of interest (mortality, disease progression, readmission, diagnosis), the time between clinical entities varies. For chronic conditions such as hypertension and CHF, clinical manifestations and symptoms might happen over years. For acute events such as AMI, stroke, or readmission, the leading clinical events and symptoms happen over hours or days.

To predict the next possible acute events, most prediction models aggregate information before one "major event" (e.g. hospitalization due to a broken hip). Any information that happened after the major event might be excluded from the prediction models. For instance, after hospitalization, clinical data that are entered during hospitalization are used to predict the risk of readmission. Moreover, some models aggregate all values for a clinical event into one value using the statistical mean or median. Hence, all changes will be normalized to one value regardless of their temporality.

For dynamic or time-changing features and output, the time dimension enhances the ability to understand dynamic clinical phenomena such as the disease trajectory, medications' effect, and potential outcomes [63]. Time is integrated into representing clinical data, querying medical records, and discovering the relations between longitudinal features [69]. Creating features that represent temporal changes in the clinical data can improve prediction models that have a time-oriented output [69,70]. Temporal data have different representations such as primitives (point or interval), time series projection (linear or non-linear), associations of clinical events, and dichotomized values in time intervals [71–73].

Static prediction models collect information at a static or baseline timepoint and predict the clinical output [74,75]. Collected data do not include the changes in patterns or clinical transitions after the baseline time point [74,75]. Static prediction models may fail to predict the patient's health status that manifests in a dynamic way [76]. For instance, readmission models are trained on data collected until discharge day regardless of any clinical events encountered after the discharge that might change the risk. Training a dynamic model that incorporates changes of time-dependent predictors can improve discrimination in the models [75]. Figure 1.1 depicts the pipeline of building objective 1. The temporality of clinical data is important in predicting negative outcome: mortality or readmission. Learning a dynamic representation for the time-varying features and accurately identify patients who are at high risk of suffering from one of the two negative outcomes: readmission or death.

Text mining and natural language processing for clinical documents

Some clinical events and facts, that activate CDS, are locked in the free text format inside a clinical document, such as pathology reports, and discharge summaries [4]. Due to variability in documenting the same clinical events, simple extraction and rule-based methods may not extract those facts and events. Natural language processing (NLP) algorithm can locate the important information and all its possible phrasing variations to actuate CDS system or enhance model performance is a necessity. For example, a complete screening for tuberculosis and treatment requires reading the patient's medical history, physical examination, chest radiography (if required), tuberculin skin test (if required), and laboratory testing for human immunodeficiency virus infection, and M. tuberculosis (when required) [77,78]. Hence, incorporating NLP methods to extract clinical data from text could improve the performance of CDS systems.

NLP has been used for decades to structure information in clinical documents. Nevertheless, there are still several challenges and barriers that limit NLP and machine learning algorithms' application in clinical text. Traditional NLP algorithms that implement syntax or linguistic rules require, mostly, an annotated dataset in the training and validation phase. Obtaining annotated data has substantial barriers such as time, money, and finding experts.

In text mining, engineering and creating features to represent the clinical documents has many challenges. Simple features such as words or phrases have limitations such as lacking the context and the semantics of the words, and the rigid dependency between words and phrases. Moreover, traditional machine learning methods such as random forest and logistic regression do not account for word context.

Medical ontologies can be used to extract features from clinical documents. Some medical terminologies include semantics in their structure. For example, unified medical language system, a medical terminology maintained by national library of medicine (NLM), assign AMI, congestive heart failure, and colorectal cancer to "disease" type. However, structuring documents using medical terminologies has some challenges. These semantics only consider if a connection between medical concepts exists, not how close they are. For instance, AMI is closer to congestive heart failure disease than to colorectal cancer. The unified medical language system (UMLS) does not quantify the relationship between disease. Prediction models trained on text represented by medical concepts yield a similar performance to models trained standard features such as bag of words [79]. Moreover, medical ontologies require ongoing curation and maintenance, lack context, and miss abbreviations and misspellings common in clinical notes.

Semantics, word meaning, syntactic representation, and relevant text can improve performance of the machine learning model [38,56,80]. Domain knowledge experts usually curate and maintain semantic features and clinical terminologies' concepts. Manually curated clinical semantic and representations have some issues such as affordability, ongoing maintenance, variation clinical annotation standard, and scalability [36]. All aforementioned barriers hinder the development of scalable clinical NLP tools.

In traditional NLP machine learning algorithms, the context of the words is rarely used in extracting or classifying clinical documents. Moreover, traditional text representations such as term frequency-inverse document frequency (TF-IDF) learn simple predictors that accentuate a few variables or apply linear transformations. These representations cannot be stacked to learn deeper and more complex representations [64,81]. For instance, bag of words does not encode the words' meaning in the representation, while latent semantic indexing learns only a linear representation and de-correlated vectors. A representation that incorporates context and mimics the human way of comprehending text could improve the performance of NLP algorithms to answer questions in chat bots, identify sentiments in movie reviews, or classify the phenotypes in clinical notes.

Combining the semantics and context of clinical text to train NLP models to create text representations can enhance classification of clinical text. Traditional semantic methods map words to equal distances, in terms relevant to each other, regardless of their similarities. For unstructured clinical data, learning the semantic relationships between words to create a text representation can map words more accurately in the textual feature space. The context around the word in the text representation can influence the documents' classification, patients' portal messages for example, as depicted in Figure 1.1.

Information extraction and de-noising clinical documents

With the rise of the EMR, providers document more information in patients' charts compared to paper-based systems [82]. While acquiring more data about patients is an invaluable data source, information overload can make it difficult to differentiate between pertinent information and noise [83]. Presenting a mix of new important information along with redundant and/or old information may interfere with the decision-making process [82,84]. Researchers reviewing and synthesizing information from clinical notes, especially unfamiliar notes, could encounter barriers such as searching difficulty, redundancy, and poor readability [85].

With the rapid rise in data volume, tools to highlight and present relevant information could reduce cognitive burden on researchers, especially those reviewing complex patient charts. Information extraction models are trained to understand the semantics of phrases and sentences and select important text. In the past, to train models that extract relevant information, some would argue that a gold-standard or manual annotation is required to train the machine learning model [86]. Annotating clinical notes manually is expensive and complex in terms of time, money, and depth of knowledge required. Annotating clinical documents requires recruiting knowledgeable annotators, such as clinicians and nurses, which could be difficult due to their time constraints [37]. Scaling and reproducing annotation, whether in another clinical organization or at a different time within the same organization, is another formidable barrier facing annotation and training NLP algorithms [37]. In past years, the NLP community has worked toward overcoming these barriers by creating and releasing de-identified and annotated datasets [36]. However, the released datasets are not large enough to train a scalable model. Therefore, pre-annotated documents have become a necessity in the era of big data.

Semantics and context can play an important role in extracting the information relevant to a phenotype. Training a model that extract sentences including relevant information can be used to pre-process and de-noise clinical text, or pre-annotate clinical text for active annotation learning models. An extraction model utilizes the semantics of the words in the documents, the context within a sentence to classify sentences into a binary output: sentences that include information about a phenotype, including relevant medical findings, medications, and procedures, and sentences that do not, as depicted in Figure 1.1.

Dissertation aims

Incorporating semantics during feature construction can improve the performance of machine learning models. This dissertation addresses methods for constructing clinical features from structured and unstructured EMR data. The developed feature representations are evaluated on three biomedical informatics problems: predicting patients at high risk of experiencing an adverse outcome (readmission or death), classifying the needs of patients in their messages to healthcare providers, and retrieving information about a phenotype or disease from clinical documents. Each problem is formulated around the three goals depicted in Figure 1.1. The machine learning pipeline discussed in this dissertation can be applied to other datasets in different organizations regardless of the underlying EMR structure.

AIM 1: Dynamic representation for structured data: The first aim describes the construction of features for structured data that capture changes in values overtime. Chapter 3 presents a method to construct dynamic post-discharge features and train a risk prediction model. We identify important major events that could lead to readmission or death after discharge. The goal of this aim is leveraging the information collected after discharge, in addition to data collected before and during admission. We evaluate the model to predict the outcome of patients hospitalized for hip fracture and CHF. We published the work in that chapter in American medical informatics association (AMIA) proceedings, in a 2016 paper entitled “Predicting negative events: using post-discharge data to detect high-risk patients” [87]. AMIA is an American professional non-profit organization that lead the initiative in “transforming health care through trusted science, education, and the practice of informatics“ [87].

AIM 2: Semantic and context representation in classification: The second aim evaluates the effectiveness of utilizing semantics and context of words in classification. It describes a deep learning method to create an abstract representation for clinical text. Chapter 4 details the design of different mechanisms for learning semantic and context of words in unstructured clinical notes. We published this work in a paper entitled "Classifying patient portal messages using convolutional neural networks" in the Journal of Biomedical Informatics (JBI) [88].

AIM 3: Semantic and context representation in information retrieval: Chapter 5 details the challenges and opportunities in retrieving information from clinical documents. It describes a deep learning information retrieval algorithm that uses the semantics of words, and the sequential context of words in sentences to learn a text representation. The method leverages big data to find a correlation between structured phenotypes (e.g. and the association between sentences and billing codes) and sentence representations to extract sentences related to a phenotype without using an annotated dataset in the training phase. The model assigns a score or a probability for each sentence. The scores can be used to extract relevant information to lessen the cognitive burden of reading long documents. The work in this chapter will be submitted to the journal of American medical informatics association (JAMIA).

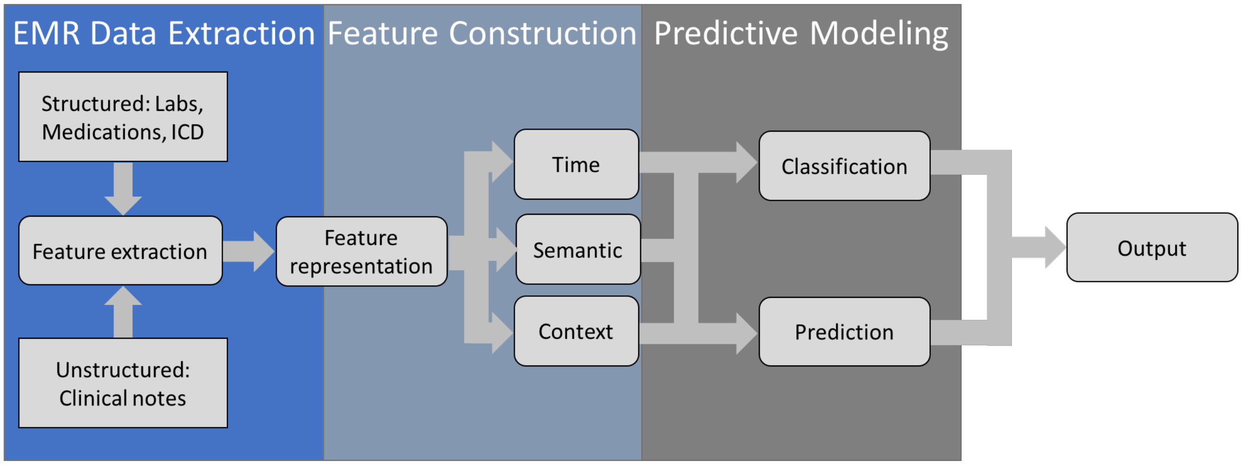


Figure 1.1 Training machine learning model on clinical data: structured and unstructured. Objective 1: creating dynamic features for structured longitudinal data by focusing on time dimension. Objective 2: creating informative text representation for clinical documents by embedding semantics and context of words for classification. Objective 3: learning text representation for information extraction from clinical documents by embedding semantics and context of words

CHAPTER 2 BACKGROUND

The dissertation explores the importance of clinical features engineering and learning informative representations to train a practical clinical machine learning model. To recognize the importance of these steps in training machine learning, one must understand the importance of each step in developing the clinical models focusing on different clinical feature representations developed in previous research. The chapter describes clinical feature engineering in clinical machine learning: data extraction, feature construction, and representation learning.

Feature engineering

Features are the main characteristics that describe a sample such as a patient or an image. For instance, a patient can be represented as the set of the diseases diagnosed with, the medications prescribed, the results of the withdrawn laboratory tests, and the personal or electronic encounters with healthcare providers. Feature engineering creates a numerical representation by extracting data and transforming descriptors into features that improve the predictive power of the machine learning model. Feature engineering may involve one or more of the following steps: 1) Feature extraction; 2) Feature construction; and/or 3) Feature learning.

Feature extraction from the EMR

Clinical informatics problems range from predicting critical patient events (readmission, disease diagnosis, unexplained access to patient records), to clustering groups of similar patients (cancer patients with similar gene expression), to extracting patients' information from clinical text (medications, diagnosis) [13,55,89–96]. Patient data are multivariate where multiple clinical events and values contribute to the next clinical decision [97]. Clinical machine learning models combine different patients' events to predict, discover, or explain a clinical problem in unseen cases, which can be time-consuming or slightly challenging to comprehend by a human.

As mentioned previously, some clinical events are stored in a structured format where each clinical event has a pre-defined set of description fields. For instance, diagnosis codes or ICD record has a patient identification number, the ICD code, and the date and time when the ICD was assigned to the patient. Other clinical events are documented and stored in a narrative or unstructured format. Structured data has been used to predict outcomes [20], extract information, assertions, relations, and risk factors [21–28], classification [29–31], deciphering sentiments and behaviors [20,32], identifying phenotypes [33–35], assigning diagnostic billing codes [36], and monitoring clinical events such as drug adverse events and infections [36,37].

Depending on the problem of interest, researchers extract data from one of the two sources. Unstructured data holds more information than structured data; however, it is harder to extract clinical information for chart review or features for clinical models [98]. Therefore, structured data are the typical choice in training clinical models. Hybrid models that implement data from both structured and unstructured resources might outperform the models trained on data from one of those two sources [99].

Documents and feature extraction

Different methods have been studied to extract patients' data from various EMR data sources. ICD codes are the most common data source used in identifying patient phenotypes [100]. A systematic review investigated the data sources implemented in identifying top 10 phenotypes in 80 studies, and diagnosis codes were the main source in identifying phenotypes in 40 studies [100]. Diagnosis codes can be an easy and fast source to extract records of acute phenotype; despite, previous studies concluded that diagnosis code alone might not be sufficient [100]. Clinical documents can be employed or/and combined to extract relevant patient information? such as demographics, procedures, vital signs, and laboratory values. NLP rules are usually applied to mine and assign phenotypes' labels to the notes.

Feature construction

Feature construction methods transform data to obtain more discriminative patterns which improve the performance of the model [101,102]. Sometimes, feature transformation methods are applied to reduce the dimensionality of the data [103,104]. Feature transformation methods can be simple such as scaling or averaging. Other methods apply a more complex mathematical transformation on the original features [105,106]. The following paragraphs describe some traditional data construction and transformation and their application in biomedical informatics.

Scaling, ratios, discretization, normalizing, and averaging are conventional methods of transformation [101]. For example, laboratory values can be represented by the average of each laboratory test over a range of time. Feature scales and ranges vary. Features with large values might dominate the training, and their weights might get updated faster. As a result, they become the main predictors regardless of their importance in the prediction [107]. If the same features have different scales, the model results might differ [108]. Hence, scaling and normalizing features are applied before training some machine learning such as logistic regression.

Discretization converts continuous values to discrete features. Discretization may improve the performance of machine learning algorithms [109,110]. It can homogenize the attribute in the dataset, clarify the non-linear relationships, and enable the derivation of count variables for non-continuous data [110]. Discretization can be supervised where the target or output is used in the discretization, or unsupervised where output information is not available or not used in dichotomizing the variables.

Minimum descriptive length-based (MDL) and ChiMerge are two supervised discretization methods. MDL divides the range of continuous values by splitting the values into bins and calculating the entropy for the output classes. The algorithm stops when it reaches a minimum entropy value [111]. ChiMerge implements the bottom-up approach. It starts from the individual values as individuals bins and merges the close values. ChiMerge evaluates similarities between bins using Chi-square tests and stops merging the bins when the significance level of the test is 0.05 or lower [112].

Reference range, equal width, equal frequency, and clustering binning are unsupervised discretization methods. Reference range categorizes values into clinically-relevant ranges (or bins) such as low, normal, and high for laboratory values, and chapters for ICD codes. Equal width divides the values' range into k bins of equal width, while, equal frequency divides the range into k bins with an equal number of observations [113]. In clustering binning, a clustering algorithm (e.g. k-means, finite mixture) creates non-overlapping bins that minimize the distance between the values in bins [114]. Maslove *et al*. evaluated different methods of data discretization to classify the arterial blood gas into 13 categories and cardiac output into four circularity shock types using clinical features [110]. The authors reported that supervised discretization methods were more accurate that unsupervised methods. Among the unsupervised discretization methods, equal frequency and k-means performed well in classifying the ABG and cardiac output [110].

Time in clinical data and machine learning

Patients encounter unexpected events that quickly change their health status, such as having a stroke or breaking a hip. Clinical interventions are applied with respect to other clinical events to move the patients to stable health status [115]. Clinicians prescribe medications, order laboratory tests, and request procedures based on the patient condition and the severity of the illness. Hence, the distribution of clinical events changes over time leading to an irregular sampling time [72].

Clinical environments are dynamic where timing of the same clinical event varies for two patients with the same health condition based on the urgency, available resources, and the day of the clinical events. A patient's health and clinical environment contribute to this variability. The non-stationary nature of the care pathway increases the challenge of sampling and creating features [115]. Hence, incorporating features' changes, and accounting for the non-stationary nature of clinical data in constructing features and training machine learning models can improve their performance [70,115,116]. Without proper feature extraction, patients who have mild and severe episodes of illness during two years might be grouped with patients who have an average severity of illness during the same period. For instance, using temporal features improved the detection of adverse drug events (ADE) compared to a cruder model that create features regardless of their time [117].

In longitudinal clinical data, exact values and intervals are two typical representations of temporal data [118]. In temporal dichotomization, clinical features are extracted by dividing the clinical timeline into bins with a start and end time. The dichotomized clinical bins outline the extraction boundaries for the clinical data. Longitudinal clinical data can be dichotomized by clock time, the events sequence (e.g. second event, fifth event), time measures in the series (e.g. three months after the first diagnosis code), or significant clinical events (e.g. time between admission and discharge). Optimum time boundaries can improve the performance of the model and create stationary representation [115]. The mean, median, or rate of change can aggregate continuous values, while counts or binary values can represent categorical variables in the created bins.

Some clinical variables are more important in prediction than other variables [97]. For example, the last values of a laboratory test, or the order time of a specific laboratory test are highly predictive for the next laboratory test or medication order [97]. Batal *et al*. implemented temporal pattern mining algorithms to predict clinical events such as adverse medical events [72,73]. These algorithms discovered more informative patterns by mining the events backward starting from the event of interest. The authors evaluated backward mined patterns on predicting adverse medical conditions associated with diabetes. Their analysis shows that patterns mined backward are more efficient than patterns mined toward the event of interest [72].

Feature representation learning

Feature representation learning generates features by transforming the input or predictors into a more informative representation for prediction or classification [64,119]. Representation learning can improve the performance of clinical NLP models by learning meaningful features and latent factors from clinical documents. Feature representation learning gained a lot of attention in fields that have complex data such as image classification, speech recognition, and text mining [64,120,121]. As previously mentioned, creating features manually for clinical text that have complex dependencies can be time-consuming, non-generalizable, and non-scalable [37]. Moreover, creating and curating simple features for complex data may underrepresent the relationship between input and output [122].

The curse of dimensionality in clinical data, such as images and text, can reduce the generalization of machine learning [64]. Moreover raw data representation can grow exponentially and can reduce generalizability especially in datasets that have few observations or complex input-to-output relationships [64,123]. Some techniques project the original features into a new space with a lower dimension while retaining most of the original information or improving the encoded information to reflect the relationship between the features.

Principal component analysis (PCA) is a popular unsupervised dimensionality reduction technique. It applies linear transformations on dependent inter-correlated variables or features to create a smaller number of orthogonal less correlated features that capture most of the relevant information [124,125]. PCA has been successfully used in biomedical informatics to investigate malignant melanoma treatment, measure speech behaviors of stroke patients, and detect walking gaits of patients with knee osteoarthritis [126–128].

Text mining using clinical documents

Unstructured clinical documents are a comprehensive source of clinical events. Some clinical data does not exist in a structured format due to the lack of self-reporting, using general coding, and using codes designed for billing [129]. Some clinical events required to activate CDS are locked in a free text format inside a clinical document such as pathology reports, discharge summaries, or progress notes [4]. For instance, colorectal cancer screening is captured by combining data from self-reporting, clinical reports, and charts, which all exist in a free format [130]. Adding features from clinical text can improve prediction models and phenotype algorithms [131,132]. Adenoma detection rate (ADR), a quality metric for colonoscopy, is another clinical event that is rarely reported in the structured format [133]. Calculating ADR mandates a careful review of EMR charts and pathology reports [133]. NLP methods extract clinical data that could improve this and other CDS systems.

Structured data within clinical documents should be carefully combined to achieve the desired results. One study combined structured and unstructured data to predict billing codes that should be assigned to patients [99]. The authors found training a model on the combined features can lead to lower performance. More accurate results were achieved by combining the prediction of two models: one trained on structured data, and another trained on unstructured data [99].

Text mining extracts patterns from EMR and structures the text into features. The resulting features can be fed into machine learning models such as clustering, or classification [134,135]. NLP analyzes and represents text by applying ranges of techniques borrowed from linguistics, computer science, and artificial intelligence [136,137]. The NLP algorithms convert unstructured data into machine-readable, structured data by processing the lexical, semantical, or syntactic levels of the text [138]. They apply statistical models to 1) identify spelling and grammatical errors, 2) categorize words, phrases, or entities (also known as name entity recognition, NER), 3) disambiguate words, 4) identify negotiation and uncertainty, and 5) extract relationships between text entities [139].

NLP has been implemented to extract medical knowledge from clinical documents. In the simplest implementation, NLP applies rule-based methods, n-gram, and regular expressions to structure the text [139]. MedEx creates a structured representation for medication from clinical documents using sequential semantic tagger and a chart parser [95]. On an advanced level, researchers train logistic regression, support vector machine (SVM), hidden Markov model (HMM), and conditional random fields (CRF) to perform NLP tasks [139]. One group developed a polarity module to detect negated text by training an SVM classifier on bag or words, cue words, negation dependency path, and constituency tree fragments [140]. Another study trained an SVM model to locate body sites [141].

Most clinical NLP studies mine clinical concept strings from medical terminologies such as UMLS, SNOMED, or RxNorm [142–144]. Clinical NLP tools such as KnowledgeMap, cTakes, MedLEE, RegEX can retrieve the medical concepts in unstructured data [145–148]. Those tools apply NLP methods to identify terms and map them to a pre-defined set of medical concept strings and semantic categories to extract information from clinical text [142–144,149]. A normalization step links the clinical term to a unique concept identifier and produces the concepts along with their text indicators (location and string) [149] KnowledgeMap maps clinical text to concepts in UMLS [142]. cTake combines SNOMED CT and RXNorm concepts, with sentence boundary detection, negation detection tool (NegEx), and part of the speech (POS) parser to extract clinical entities from clinical notes [143,144,147,150].

Researchers have been training machine learning models on various text representations to mine and extract clinical information such as medication, diagnosis, and tumor information [151,152]. Vector space models map or embed each term, word, or document as a point or a vector in space. In a few spaces, the distance between the mapped vectors of words correspond to similarities between words and smaller distance implies higher similarities [153].

Text mining and NLP have various applications in the biomedical informatics field. Latent Dirichlet Allocation (LDA) models have been applied to discover psychiatric symptoms (suicide, severe depression) and depressive disorder comorbidities (postpartum, brain tumor) in psychiatric documents [154]. Sentiments scores were extracted from discharge summaries in an i2b2 dataset and fed into a Cox regression model to predict the risk of readmission and mortality after discharge [20,155]. Another group trained an elastic net model to select essential features from nursing notes represented by TF-IDF vectors [156]. Liao *et al*. used ICD9 billing codes and concepts extracted from unstructured notes to develop a phenotype algorithm [157]. Combining both structured and unstructured data improved the performance of phenotyping algorithms for multiple sclerosis, Crohn's disease, ulcerative colitis, and rheumatoid arthritis [157]. NLP has been applied to identify surgical patients with pancreatic cysts, patients who can be discharged from a neonatal intensive care unit, and children with asthma [158–160].

Challenges in applying language processing to clinical text

The rich vocabulary in clinical notes can increase the complexity and the variation in clinical narratives [19]. Healthcare providers use sub-languages and abbreviations tailored for sub-domains and express the same information in various words and writing styles, as demonstrated in Table 2.1 [4]. The examples in Table 2.1 were extracted, inferred, or inspired from clinical documents in synthetic derivative, a de-identified version of Vanderbilt University Medical Center EMR.

Expressivity of documentation, which conveys both patient and provider impressions, lead to variations in the length and the content of the document [161]. Hence, documents vary in length depending on the patient condition and the document creator. The atypical grammar and various writing styles are other challenges that researchers face in clinical NLP [19]. Formatting and structuring the clinical document in a pre-defined format could ease text mining; however, restricting the style of documentation could increase the cognitive burden, decrease efficiency, and introduce a delay in clinical workflow [161–163].

Table 2.1 Examples of similar phrases written and communicated differently in clinical notes.

Clinical event	Sentence variation 1	Sentence variation 2	Sentence variation 3
Describing occlusion in heart	Diffuse distal LAD occlusion	Showed 90% RCA occlusion	65% ostial SVG-LAD lesion
Treating patient with Heparin	She was treated with TPA and heparin	As well as bridging with heparin	Pt was treated aggressively with intravenous heparin
Medication prescription	Atorvastatin 20 mg tablet daily	Toprol 100 mg once a day	Levothyroxine 30 milligram daily
Describing patient with cancer diagnosis	He is a gentleman with T3 rectal cancer	Patient was diagnosed at age 65 with colon cancer	Pt is a pleasant male with presenting to discuss therapy for stage III pancreatic cancer

Multimodal algorithms and ontologies (e.g., LOINC for laboratory tests, UMLS for medical concepts, and RxNorm for medication) must be deployed to extract clinical entities from unstructured data [142,144,164]. Rule-based extraction methods including the aforementioned clinical NLP tools are subject to challenges including misspelling, structural ambiguity, lexical coverage in dictionaries, acronyms, and abbreviations [165]. For example, UMLS has some problems including missing concepts and ambiguous terms such as "other location of complaint" [4]. One study reported that using UMLS, as a source of lexical knowledge to extract medical concepts in discharge summaries and chest x-ray reports, did not outperform a local or customized lexicon [148]. Engineering text features using terminologies may not improve the metrics of a machine learning model. Another study evaluated whether prediction models trained on features extracted via medical dictionaries will outperform prediction models trained on the actual terms and words in the documents [79]. The difference between the two models was not statistically significant [79].

Information extraction from clinical notes

Information extraction can enhance a providers' access to patients' data. Extraction models can locate sentences about a phenotype, identify indicators for infectious diseases, cancer, and diabetes, and provide data for research [166]. Effective extraction models can increase the flexibility of documentation without worrying about the information accessibility or restricting providers to write only in a specific format [167]. Adapting and developing extraction algorithms has been difficult due to the nuances and noise in the medical text [168]. This dissertation proposes a method that extracts sentences contain clinical events and facts about a specific phenotype in clinical documents.

Challenges in manual annotations

Most clinical text mining and information extraction methods are supervised models. Those NLP algorithms require annotated datasets to train them. Nevertheless, manual annotation is laborious, time-consuming, incomplete, inconsistence, and prone to human error [145,165]. For instance, annotator needs, on average, 87.2 ± 61 seconds to manually de-identify narrative in one note [169]. At Vanderbilt University Medical Center (VUMC), a locally hosted crowdsourcing system was implemented to recruit annotators to perform tasks such as annotating a document or an image. Most annotators are medical students, and a few of them are nurses and clinicians. The cost of annotating a document is $20 per hour. The annotated datasets are usually small. For example, the number of annotated documents in each i2b2 dataset, a well-known series of annotated datasets that were released as part of NLP challenges, ranges between 400 and 2600 documents [24,155].

The complexity of annotation is another challenge that might affect the results and scalability of NLP algorithms. The complexity of annotation varies depending on the target task, the delimitation of annotation in the text, and the inclusion criteria [170]. As a result, annotated sections could differ among annotators. Even with the availability of resources to perform annotation, it is non-scalable due to the high variability and the small size of generated datasets [171]. The quality of the annotated training dataset could impact the trained models, which affect the scalability and reliability of the trained models [172].

Machine learning: definition and approaches

Machine learning includes methods that learn patterns from data and examples, and use the detected patterns to predict unseen data, or suggest a decision for new examples under uncertainty [173]. Each example or input consists of numerical independent variables or features. Machine learning algorithms can be supervised or unsupervised. Supervised learning methods learn the mapping between the input and output or dependent variable. The outputs' format determines the problem and learning method. Real values require regression models while categorical values require classification learning models. Unsupervised methods discover patterns in the data using the input values only.

Different machine learning techniques have been developed and implemented over the years. Random forest technique has been used for regression and classification. Random forest trains a group of weak learners or weak decision trees [174]. Each decision tree is trained on a sampled subset of the training dataset and only on a subset of the features [174]. The decisions of all trees are aggregated to generate one predicted output [174].

Regression is a common machine learning algorithm that has multiple learners depending on the distribution of the output. Linear regression asserts a linear relationship between the input and the output, and learns weights for the input features that map the linear relationship while accounting for a normally-distributed error [173]. Logistic regression generalizes the linear regression for binary classification by replacing the normal or Gaussian distribution for the output with Bernoulli, and passing the output of the model through a sigmoid function that limit the output to the [0-1] range [173]. Researchers prefer regression because they can explain and extract the learned relationship between the input and the output using the learned weights, especially in medicine where explanation is preferred over prediction [175,176].

Deep learning

Data representation influences the performance of machine learning models [177]. Conventional machine learning models have a limited ability to process complex data in the raw format [178]. Representation learning discovers useful representations for classification or prediction from raw input data [64,178]. Representation learning is a deep learning model that generates complex concepts using simpler ones, and multiple levels of non-linear modules or layers [177,178]. Each level builds a more complex representation and slightly more abstract representation compared to the previous level and feeds it to the next level. For instance, a deep learning model constructs a representation for a house image using lines and edges. Figure 2.2 depicts an example of a general deep neural network.

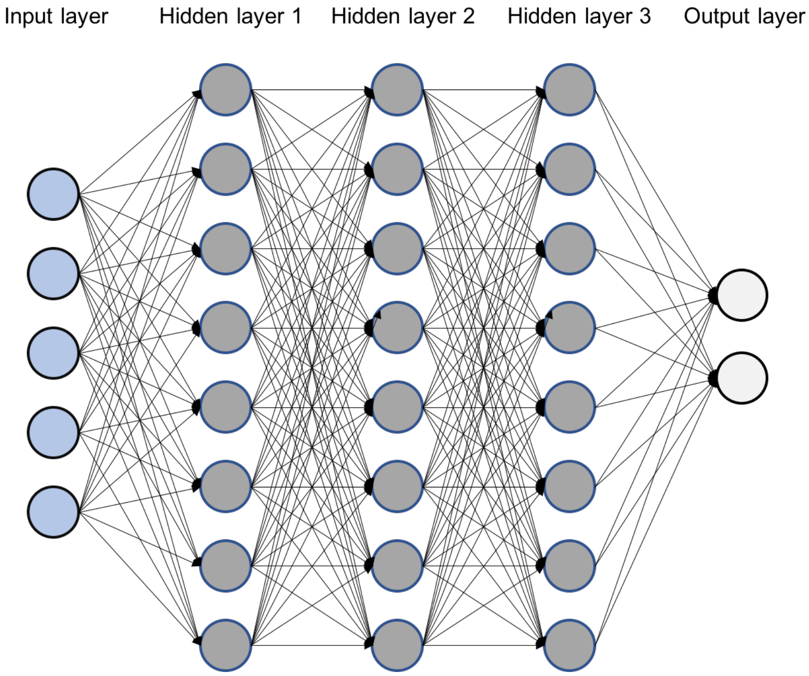


Figure 2.1 An example of a deep neural network

Deep learning models have different variations in their architecture, depending on the target task. Feed forward neural network (NN), one of the conventional deep learning methods, consists of:

1) An input layer with fixed-size that holds the observed values. In our house example, the house image consists of pixels, where each pixel holds a float value.

2) One or more layers, known as learning layers, neural layers, or hidden layers (Figure 2.2). The layers transform the input into higher and more abstract representations that are easier to model

[179–181]. Each hidden layer consists of hidden units (or neurons) that apply non-linear functions on its input, calculate the weighted sum of the inputs from the previous layer, and pass it to the next one [177–179,182]. The rectified linear unit (ReLU), hyperbolic tangent function (Tanh) are common non-linear functions. The trainable weights in each layer hold real values [182]. The output of the neural network layer activates the neurons in the next layer [182,183]. The number of hidden layers and the number of hidden units can vary depending on the complexity of the model and the purpose of learning. A shallow neural network has one hidden layer with a high number of hidden units while the deep network consists of two or more hidden layers. For the house image, each layer learns higher abstracted detail about the house. The first hidden layer learns the small lines that constructs the house. The second hidden layer learns the individual components such as windows and doors. The third hidden layer learns the house's front view, the roof, etc. The last layer learns the final shape of the house.

3) A fixed-size output layer that holds the value to be learned. This could be a probability, a class, or a numerical value. In our house example, the output will be if there is a house in the picture or not.

Several different deep learning models that have been implemented with two specific types applied to NLP. Convolutional neural networks (CNN) is a deep neural network method that models temporal and spatial correlations between features. It processes data and takes into account the proximity of the data in two-dimensional space, such as images. A CNN applies layers of functions (i.e., convolutional layers) on all possible regions in an input matrix to compute its output. CNN maps features using multiple convolutional functions and selects most important representations from a pool [178]. The main two advantages of CNN are detecting patches of features regardless of their location in an input matrix (i.e., location invariance), and composing patches of features into higher-level representations. CNN is mainly applied to images classification, computer vision, signal processing, and face recognition [184–187]. Chapter 4 provides a detailed description of CNNs.

Recurrent Neural Networks (RNN) are another variation of neural networks. RNNs process a sequence of elements, one at a time, and include information about all the preceded elements. RNNs are suitable for sequence analysis such as speech and language [178]. Long short-term memory (LSTM) is an RNN variation that controls the proportions of current element information and the proportions of previous elements to include in the training. Chapter 5 describes the architecture of the RNN and LSTM.

The popularity of deep learning in healthcare has increased due to the rise in the volume of EMR data and its successful application in other domains. Neural networks have been used to create patient representations to predict readmission, disease, and medication [42,49,61,188–190]. Miotto *et al.* successfully inferred representations for patients and evaluated them by predicting the probability of developing various diseases using data from Mount Sinai [49]. Rajkomar *et al.* trained a deep learning model to predict medical events such as unplanned readmission, in-hospital mortality, and prolonged length of stay [93]. Choi *et al.* trained an RNN model on longitudinal diagnosis codes, medications, and clinic visits to infer patient representations that predict future diagnosis codes and medications [61]. Lasko *et al*. implemented stacked sparse auto-encoder to identify multiple population subtypes using longitudinal sequences of serum uric acid measurements [42]. Bajor *et al*. trained LSTM on billing codes to predict the classes of medications prescribed for patients [189]. All of these studies investigated different applications of deep learning methods on EMR data to solve a health care informatics problem.

Deep learning and text representation

A variety of deep learning models have been employed in NLP to translate text, answer questions, and classify text [184,191,192]. The trend of implementing deep learning in NLP has sparked after the successful implementation of word embedding. Traditional machine learning algorithms are trained on hand-crafted or traditional features that have some drawbacks, as discussed earlier. Deep learning methods can leverage a humongous amount of unlabeled text to learn a useful representation. Deep learning has demonstrated promising results using non-clinical text [193,194]. It can overcome some of the aforementioned challenges when applied to clinical text. Implementing a deep learning pipeline that learns underlying context and semantics and creates an informative representation of the clinical text can have various implementations in the clinical environment.

Recently, researchers have trained unsupervised vector space models to learn word embeddings from large, unlabeled datasets [195,196]. Mikolov *et al*. trained two neural networks to create the word embedding that considers words similarity [195]. The first neural network is Continuous Bag-of-Words model (CBOW) that predicts a target word using surrounding words [197]. The second neural network, skip-Gram model, predicts surrounding words using the target word [197]. For instance, CBOW model predicts the kidney embedding using the words "I scheduled the" and "transplant annual appointment," while skip-Gram model uses kidney to predict "I scheduled the" and "transplant annual appointment," as Figure 2.3 depicts. Pennington *et al*. learned word representation using word-word co-occurrence counts, and used a log-bilinear regression model to create a word embedding [196]. Trask *et al.* encoded both word- and character-level representations to learn word embeddings [198]. These learned word representations preserve the meaning or analogy of word [195,196,198]. Hence, similar words are mapped to close proximity in the vector space. Word embeddings have been implemented in text classification, analyzing social media posts (e.g. the mention of adverse drug reaction in tweets), text summarization, and question answering [192,199–201].

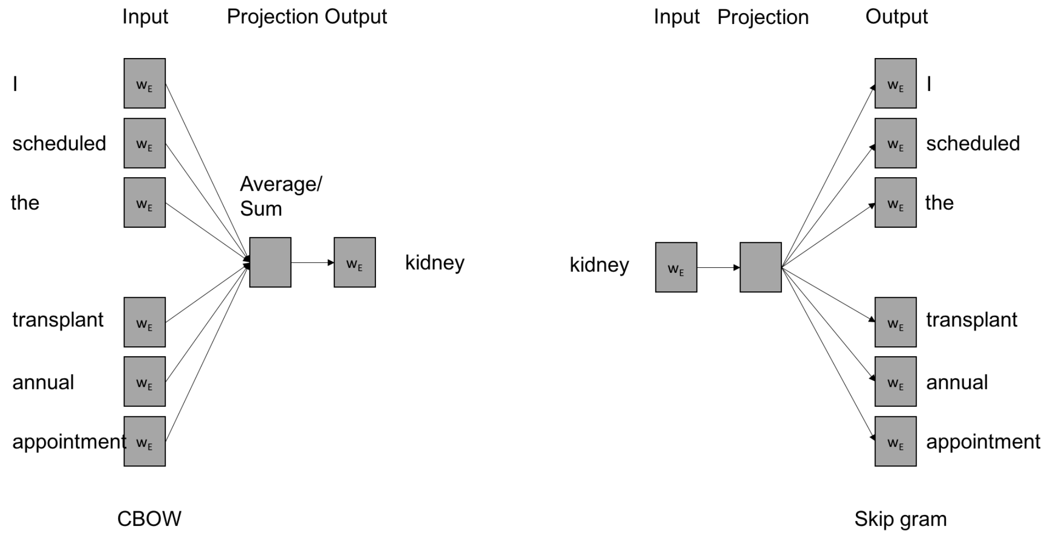


Figure 2.2 A visual illustration for the continuous bag of words (CBOW) and Skip-Gram model to learn word embedding, inspired by figures in [197]

Deep learning models have been used to analyze text [193], including CNN [184,202]. Recently, CNNs have been used in NLP to extract features for sentence classification [184], sentence modeling [203], sentiment analysis [204], and sentence matching [205]. Grnarova successfully combined Word2Vec and CNN to predict intensive care unit mortality rates [206]. Their results demonstrate that a CNN outperforms bag of words and paragraph vector inferred by Paragraph2Vec model.

The impact of EMR data representation

Training clinical machine learning models start after identifying the problem or need that is time-consuming or/and requires a series of computations. Defining the problems and the hypothesis of our experiments and modeling inform the informaticians of EMR data required to train the machine learning model. Extracting and creating features is the first step in developing machine learning models. An informative representation captures the essence of patients' trajectories, the changes over time, or the content of EMR data, in a structured or unstructured format. Learning or constructing an efficient representation can improve the accuracy of the model. In structured data time and temporality can improve the performance of the machine learning model. In unstructured data, semantics and context of words can enhance the classification of documents and the extractions or information. Some machine learning models are trained on features that lack time dimension, semantic, or context. Moreover, most text features and NLP algorithms require human knowledge which can be expensive and hard to obtain.

This dissertation builds on and expands previous work in biomedical informatics that to address creating representations for structured and unstructured EMR data that enhance the performance of clinical machine learning. The proposed methods extract and construct the features based on the dimensions that reflect the changes in patient status in structured data (Chapter 3). Moreover, the proposed methods to address the learning and the creation of features that incorporate semantics and context without human knowledge. The learned representation for the clinical text can be used for classification (Chapter 4) and information extraction (Chapter 5).

CHAPTER 3 DYNAMIC FEATURES: TIME AND POST DISCHARGE DATA

Introduction

After confirming a diagnosis, health care providers create a treatment plan to transfer the patient into a stable state and prevent future relapses such as unavoidable readmission, death, or disease prognosis. Most prediction models collect data or predict output at one baseline point such as the discharge day. Static prediction model may provide misleading results during consecutive follow-ups due to multiple reasons [207]. Events that happen between the baseline point and the prediction time might change the outcome. For instance, premature discontinuation of treatment might change the survival probability of a breast cancer patient [207]. On the other hand, the values and the importance of some predictors change after data collection, which leads to time-varying effects on the outcome. For example, the increase in red blood cell units in a trauma patient during the treatment is associated with worse outcomes [208].

Time plays a vital role in feature construction and clinical model implementation. This work evaluates the importance of time in constructing dynamic features and deploying machine learning model for one of the challenging problems in biomedical informatics: predicting patients at high risk of being re-admitted or die after being diagnosed with a disease for the first time.

Care providers and the administrators of healthcare organizations are keen on determining which patients might experience complications that lead to readmission or death. Healthcare providers may allocate additional resources to the patients and intervene before an adverse outcome transpires [209]. Unfortunately, resources are limited. Decision makers within health organizations (e.g., physicians and care coordinators) specify a subset of high-risk patients who can receive special attention while other high-risk patients do not have access to such resources [210,211]. Identifying and prioritizing, which patients should be assigned assistance, is a vital informatics problem.

Traditionally, negative outcome prediction systems are executed at the time of discharge to identify high-risk patients [212]. However, such systems are limited in their applicability because patient status often changes after their discharge. Risk prediction models are not amended to incorporate such variation. This lack of dynamic post-discharge knowledge can cause risk assessment errors and potential readmission penalties, under meaningful use regulation, calculated via a payment adjustment factor [209].

Objectives

To investigate the effect of time in constructing features, such as pre-discharge and post-discharge features, this aim focuses on several core questions in this chapter.

1. How can the patient timeline be divided to predict the outcome after being hospitalized and diagnosed with a disease?
2. What are the post-discharge data that are collected after a discharge to construct post-discharge features and train a dynamic post-discharge model?
3. Does the performance of risk prediction model improve if the post-discharge features are added to the static features in training?
4. How do different time periods of post-discharge information impact the predictions?

This work evaluated standard static and post-discharge prediction models using three years of data from VUMC's EMR system for two phenotypes: 1) an acute condition in the form of a hip fracture and 2) and a chronic progressive disorder in the form of congestive heart failure (CHF). The results in our work published in AMIA proceeding demonstrate that [87]:

- Predicting at successive post-discharge days, and including dynamic post-discharge data in the prediction model, outperforms state-of-the-art static "at discharge" models, such as LACE [213,214];

- The importance of post-discharge clinical features grows as the prediction horizon for adverse events is pushed further into the future;

- Higher utilization of clinical resources (e.g., appointments and medications) after discharge are correlated with a negative outcome; and

- Combining the structured data with the content of clinical documents such as discharge summaries or patients' electronic messages did not improve the performance of prediction.

Clinical significance

The proposed model highlights the main components for building decision support tools that identify high risk patients using the data after discharge. The approach introduces techniques that help developers to train and build a CDS tool that identifies high risk patients using dynamic and recent post-discharge data by:

1. Identifying the time of collecting clinical features after discharge to train high-risk patients prediction model.
2. Measuring the improvement of including dynamic data about the patient after discharge in training clinical risk prediction models.
3. Applying the concept of dynamic features to clinical informatics problems that exhibit dynamic nature such as phenotype diagnosis or hospitalization.

Background

The number of proposed risk prediction models has increased over the last decade. Kansagara and colleagues performed a comprehensive systemic review to evaluate the performance of risk prediction models and their suitability for clinical use [215]. Studies usually compare their models to an established one for evaluation purposes. LACE is an established index that provides a risk score to predict the readmission or death, specifically for CHF patients, using length of stay (L), the acuity of admission (A), comorbidity score (C), and the number of the emergency department visits (E) [213,214].

Several studies have shown that post-discharge data can assist in the prediction of negative events in special circumstances. For instance, certain studies investigated the surgical quality assessment at the point of discharge and observed that over a quarter of the complications are diagnosed post-discharge [26,216,217]. Another study found that post-discharge data could improve the prediction of the presence and the severity of the spasticity in upper limbs in the year following a stroke [218].

While post-discharge data has rarely been used in readmission prediction, Hersh and colleagues performed a systemic review about the post-discharge environment and its relation to readmission after heart failure [219]. They reviewed 26 studies published between 1985 and 2011 to evaluate the importance of integrating post-discharge environment in heart failure readmission models. In the review, only seven studies included post-discharge data and focused mainly on whether the patient had a primary care provider. They concluded that the socio-economics of the post-discharge environment are a key indicator that affects readmission probability. Another correlated factor with readmission is the number of follow-ups after discharge. Specifically, patients with a more significant number of early follow-ups tend to have a lower likelihood of unplanned readmission, especially for patients with a higher number of comorbidities [220–223].

Some studies created bins and time windows to divide the clinical timeline and construct features. Creating temporal bins can improve prediction models such as readmission model. Zhao et al. improved the ADE prediction by dividing the timeline before the ADE into temporal bins and creating temporal features [117]. Wang et al. developed a one-sided convolutional non-negative matrix factorization to extract temporal patterns for diabetic patients from segments of clinical sequences[224].

Dynamic models incorporate changes in predictors based on the patient trajectory which simulate the clinical approach [75,208]. Clinical dynamic models include new patient's data such as new laboratory values, new medications during hospitalization or after discharge. Tangri et al. applied dynamic modeling to predict whether chronic kidney disease patient will need a kidney transplant [75]. Retraining the model with latest-available-measurement for labs improved the discrimination and goodness of fit of risk prediction [75]. Hubbard developed a time-dependent prediction for mortality within discrete time intervals in trauma dataset [208]. The results show the features' importance in predicting future mortality changes over time, and dynamic modeling increases the precision and the accuracy of prediction [208]. Caballero proposed a framework to re-estimate the probability of readmission to the ICU after being discharged to a lower care level when a new feature is observed [225]. Dynamic estimation of the readmission to ICU had higher AUC, sensitivity, and specificity compared to static models [225].

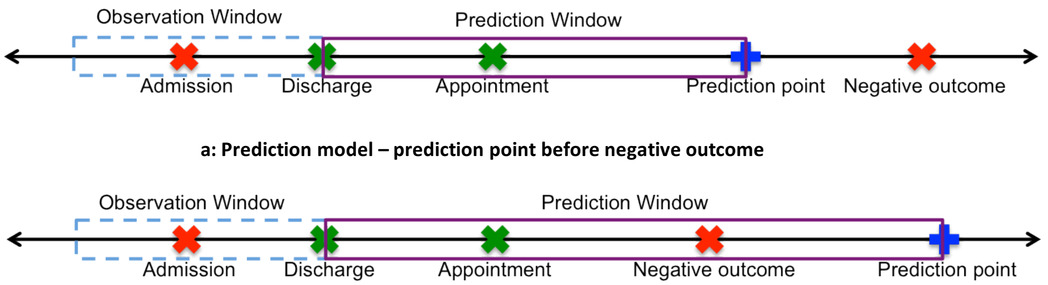
Methods

This section describes the dynamic risk prediction models. First, we construct temporal features from different time bins. Second, we propose a method to construct post-discharge dynamic features extracted from different time windows. Third, we evaluate these features by predicting patients who might encounter an undesirable outcome (readmission or death) within a prediction window.

Feature construction

We represent longitudinal EMR data with an $\boldsymbol{M}^{NxT}$ matrix, where T is the length of time for extracted EMR data and N is the number of patients in the cohort. All patients in the cohort were hospitalized and diagnosed with a phenotype for the first time. Each row in the matrix represents one patient vector denoted by ***t*.** For each patient *p*, we divide the longitudinal medical record vector ***t*** into three temporal bins: **before admission**, **during admission**, and **after discharge**. An observation window (OW) specifies the time from which to extract features. Using a combination of features, we predict whether the patients would have a negative outcome within a prediction window (PW), where the window starts at the discharge day and ends at the prediction point (PP). Static features have a fixed value regardless of the time of the prediction. Dynamic features values change based on the prediction point. For instance, the number of appointments that patients scheduled before admission is the same, while the number of scheduled appointments after discharge may vary for 30 days versus 60 days PW.

For the purposes of this investigation, we represent outcomes as a dichotomous variable. A patient has a negative outcome if he/she experiences a negative event (-1) in the prediction window and has a non-negative outcome (+1) otherwise. Figure 3.1 visualizes model settings in which both patients experience negative outcomes, but only the patient in Figure 3.1b has a negative outcome in the prediction window.



a: Prediction model – prediction point before negative outcome

b: Prediction model – prediction point after negative outcome

Figure 3.1 Prediction models based on temporal features (a) before and (b) after a negative outcome

Prediction Model and Features evaluation

We constructed five different patients' representation extracted from different times or observation windows to evaluate the importance of temporal features and different representations. We quantfied the effecitvness of representations by predicting the outcome of the patients.

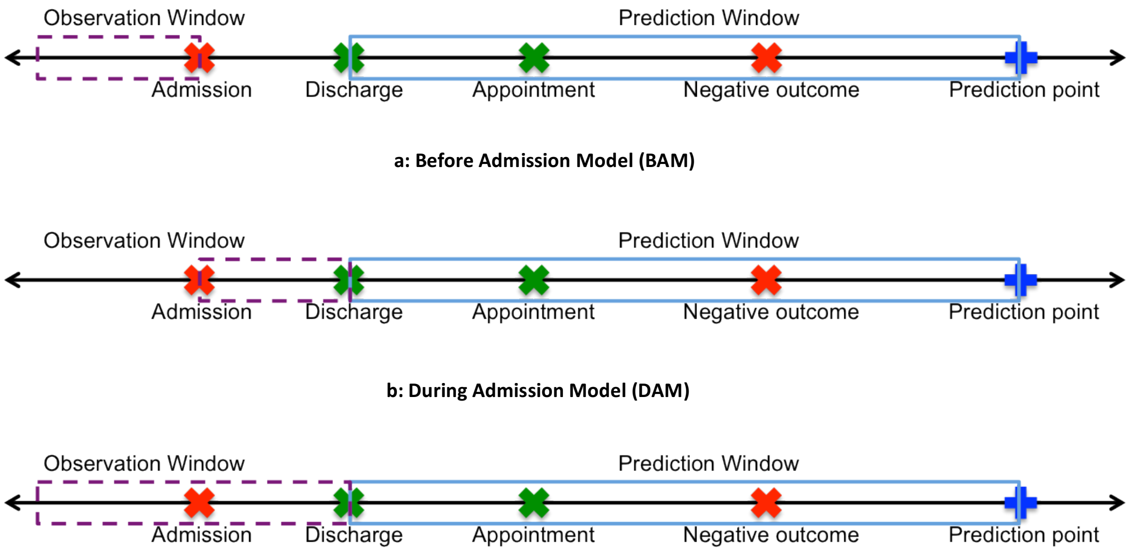
Model 1: LACE: Starting with the most common method in the literature, we build a static prediction model using predefined LACE features, where E is the number of the emergency department visits in the past six months. LACE assigns points to each variable based on its value and calculates probabilities using regression models [214]. Patients with a score higher than ten are considered to be high risk. We retrieved the four LACE features from the EMR and calculated the risk score, which we fed as a feature into the prediction model. LACE was the baseline in the models' comparison.

Model 2: Before-Admission Model (BAM): To learn whether prior health status can forecast the future health status of patients, we represented the patients using before the admission features. Each patient's entry is assigned to a pre-admission feature vector denoted by ***b***. The **BAM** matrix is visualized in Figure 3.2a.

Model 3: During-Admission Model (DAM): We investigated whether the representation learned from data collected about a phenotype is sufficient to predict the outcome. The model represented patients using admission data as visualized in Figure 3.2b where the observation window begins at admission and ends at discharge.

Model 4: At-Discharge Model (ADM): This model represents the patient by before and during admission features as visualized in Figure 3.2c. This static representation is the conventional approach in training risk prediction model.

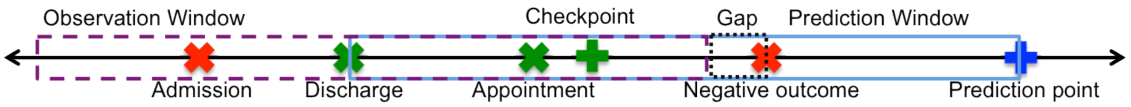
Model 5: Before, During admission, and After discharge Model (BDAM): All previous models predict the outcome by creating a static representation for the patients. This dynamic model incorporates post-discharge data to predict the patient's outcome, which Figure 3.2d visualizes. This model mimics the actual clinical practice in re-estimating the possible outcome based on updated information about the patient. We defined three key parameters to extract and construct dynamic post-discharge features: 1) checkpoint: time point at which we collected post-discharge data, 2) prediction point: time point to identify the length of the prediction window, and 3) gap: an unbiased parameter that excludes data prior to negative events that might introduce biased knowledge about an upcoming negative event. Post-discharge representation integrates all changes after discharge except the ones that happened immediately before the end of the observation window (i.e., events in the gap window). Thus, a checkpoint C, a gap G, a prediction point PP and prediction window PW variables determine the dynamic information in post-discharge features. The BDAM model combines the three temporal representations to predict the outcome.



a: Before Admission Model (BAM)

b: During Admission Model (DAM)

c: At Discharge Model (ADM)



d: Before, During and After admission Model (BDAM)

Figure 3.2 Prediction models based on varying temporal features

Overview of Features

For each patient, we extracted the demographics (age, gender) and data inserted from one year before to one year after the first documented incidence of the phenotype under investigation (i.e., hip fracture and CHF). The extracted data were the number of resources that were allocated for treatment, including: 1) medications, 2) laboratory tests, 3) appointments, 4) previous admissions, 5) the average of previous the length of stays (LOS), 6) days since the last admission and 7) the count of the ICD, Ninth Revision (ICD-9) in each of the 20 chapters. Table 3.1 summarizes the features and their temporal period (e.g., before admission, during admission, or after discharge). In addition, we extracted the number of documents, grouped by their types, which were created and stored in the EMR during hospitalization.

I retrieved the average values of the most common laboratory tests that were ordered for 80% of the patients during admission and after discharge, including carbon dioxide levels (CO_2), creatinine, glucose, hematocrit or packed cell volume (PCV), partial thromboplastin time (PTT), potassium (K), sodium (Na). Clinicians order these laboratory tests to evaluate heart and kidney functionality, electrolyte balances, and blood clotting timing. Table 3.2 lists the extracted laboratory tests, their normal ranges, and the diagnostic purpose of the test.

Unstructured data representation

Clinical communication and discharge summaries are other sources of patient information. We implemented the following NLP methods to convert the documents or communication messages into a structured format:

1- Latent Dirichlet Allocation (LDA): we trained LDA on clinical communications and discharge summaries and extracted 20, 30, and 50 topics. A binary vector for the LDA topics was created where the corresponding topic is set to one if the document included that topic.

2- Average of Word2Vec vectors: we trained a Word2Vec model on discharge summaries and clinical communication. For each document, we tokenized the document into words and retrieved the Word2Vec embedding for each word. The mean of the words' vectors represented the text.

For each patient, we created two document vectors: discharge summary vector to incorporate at-discharge unstructured information, and post-discharge clinical communication to incorporate information communicated after discharge. we concatenated the discharge summary and clinical communication vectors to the structured data vector.

Table 3.1 Summary of the features included in the models with the observation window taken from. The symbols *, +, and - represents extracted from before, during and after time periods respectively

Feature Type	Feature Values	Feature Bin		
Demographics	Age and gender.			
Laboratory tests	Number of laboratory tests.	*	+	–
	Average values of: glucose, creatinine, partial thromboplastin time (PTT), hematocrit or packed cell volume (PCV), Carbon Dioxide levels (CO_2), potassium (K), and sodium (Na).		+	–
Medication	Number of medications prescribed for the patient.	*	+	–
ICD	The count of ICD9 in each of 20 chapters.	*	+	–
	ICD deviation post-fracture (the ratio of ICD chapters number in an appointment after discharge to the average number of ICD chapters before hip fracture incidence).			–
Routine care	The average of Braden score, the number of ECG tests, the number of times a patient received respiratory care.		+	
Admission	Length of Stay (LOS).		+	
	Last day of previous admission.	*		
	Average LOS	*		
Appointment	The number of appointments.			–
Documents and communication	The number of communication message			–
	The number of documents initiated for per document type.		+	
Post-discharge time	Number of days since discharge.			–
	Number of days until prediction point			–

Table 3.2 Common laboratory tests, their normal values, and the diagnostic purpose

Laboratory test name	Normal values	Purpose	Abnormal values reasons
Creatinine	Male: 1.3 mg/dL Female: 1.1 mg/dL	Test kidneys functionality	Higher than normal level is an indicator of kidney malfunction

			such as kidney failure, blocked urinary tract, and kidney damage.
Partial thromboplastin time (PTT)	25-35 seconds	Measuring the time that the blood takes to clot	Abnormal or long PTT time indicate bleeding disorder or disorder in clotting process
Hematocrit or packed cell volume (PCV)	Male: 55% Female: 42%	Measuring the percentage of Red Blood Cells (RBC) in blood	Low PCV: indicator of anemia, over-hydration, and destruction of RBC High PCV indicator of dehydration, congenital heart disease, or abnormal increase in RBC
Carbon Dioxide levels (CO_2)	23 to 29 mEq/L	Detecting the body's electrolytes imbalance	Low levels: indicator of acidosis, Kidney disease High level: indicator of breathing disorders, hyperaldosteronism
Potassium (K)	3.7 to 5.2 mEq/L	Assessing the kidney and heart functions.	Low levels: Chronic diarrhea, renal artery stenosis, diuretics High levels: blood transfusion, Kidney failure, acidosis
Sodium (Na)	135 -145 mEq/L	Measuring balance between sodium and water in consumed foods and drinks	Hyponatremia (Na < 135 mEq/L): kidney disease, heart failure, or ketones in blood from starvation Hypernatremia (Na > 145 mEq/L): dehydration, severe vomiting, or diarrhea

Negative Outcome Prediction Over Time

The model predicted the outcome for different prediction points and prediction windows ranging from seven days to one year. The variations in the prediction window assess the utility of patients' representations, specifically the changes in post-discharge data importance.

Model Implementation and data extraction

This section describes the data extraction and provides high-level overview of the prediction algorithms.

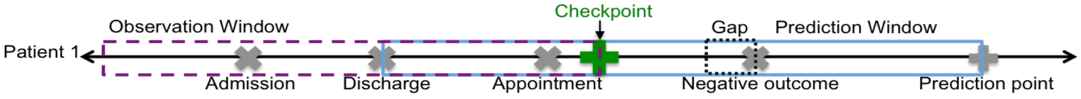
We extracted patients who were diagnosed with a hip fracture or CHF, using the ranges of 820.* and 428.* ICD9 codes, respectively. We excluded patients who had no encounters before admission and after discharge, and visited VUMC for only that admission. We analyzed only the first admission for that phenotype and excluded repeated the admissions.

Input matrix**:** We construct the BAM, DAM, ADM, and BDAM input matrices. For the BDAM matrix, different checkpoint days can influence the prediction's performance based on the changes consolidated in dynamic post-discharge features. We uniformly sampled 100 checkpoints at random day between the discharge and prediction points to account for the variability in post-discharge data. The results report the average of 100 predictions. The random day sampling approach provides a viable option for testing the model since the output changes as different amounts of post-discharge data are included.

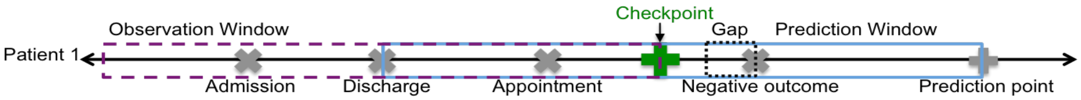
Changing the prediction window length: The prediction window length varied from 7 to 365 days to study the changes in risk and the effects of the phenotype over time.

Model implementation: We used a random forest classifier from Scikit-learn and conducted five-fold cross-validation [226]. We trained the random forest model using the parameters: 500 trees with 15 maximum depth and 15 as the minimum split.

In the BDAM matrix, the gap value was five days for all prediction points except for seven days' prediction. For one week prediction, the gap value was assigned to three days to yields four days to sample post-discharge data. The values in the post-discharge data depend on the checkpoint location. Different checkpoint locations lead to different BDAM entries for the same dataset as shown in Figure 3.3 (Figures 3.3a and 3.3b exhibit the vectors constructed with different checkpoints for the same patient). To minimize the effect of randomness, we built 100 matrices at each prediction point by randomly sampling checkpoints and averaged the area under the receiver operator characteristic curves (AUCs).



a: An example of a random checkpoint



b: An alternative random checkpoint

Figure 3.3 Building BDAM vector using two checkpoints (a) an example of a randomly sampled checkpoint and (b) an alternative sampled checkpoint

Important features and their relationship to outcomes

We extracted the features that, on average, have the highest importance across all folds and all 100 samples in the post-discharge model. The features importance will demonstrate the static and dynamic features importance in predicting the outcome. We analyzed the importance of the features that comprised around 50% of the total feature importance. Additionally, we analyzed the non-linearity of features with partial dependence plots. The partial dependence shows the dependence of prediction on a subset of the input variables. It finds the marginal average of the prediction to identify the effect of chosen subsets of features on the prediction probability after accounting for the rest of the input features. For a given predictor, the *y* values in the partial plot show the average of prediction probability across all trees in the forest.

Results

This section begins by summarizing the patient populations and their negative outcome rates. The following subsections present the performance of the models, with LACE as the baseline. Finally, the result section concludes by reporting the importance of the features incorporated in the model and the outcome.

Patient Population

This study was conducted on patients who were admitted to VUMC between 2010 and 2013. We included patients who were 65 years and older and were diagnosed with either a hip fracture or CHF. We excluded patients who did not have an admission on the onset date of the phenotype. This selection criteria yielded 704 hip fracture patients and 5250 CHF patients.

Around 25%, and 21% of hip fracture and CHF patients, respectively, had a negative outcome within 90 days after discharge. The one-year survival rates exhibited a similar trend, as shown in Figure 3.4. For patients who exhibited a negative outcome within the first 7 days, more than half were admitted to the emergency department or died within the first three days after discharge.

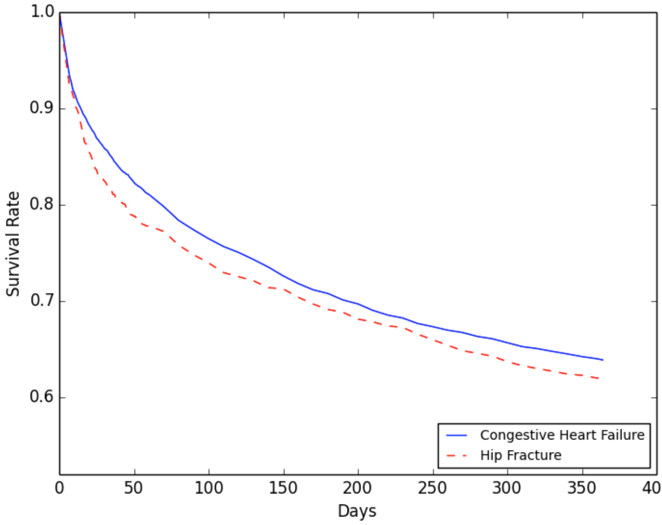


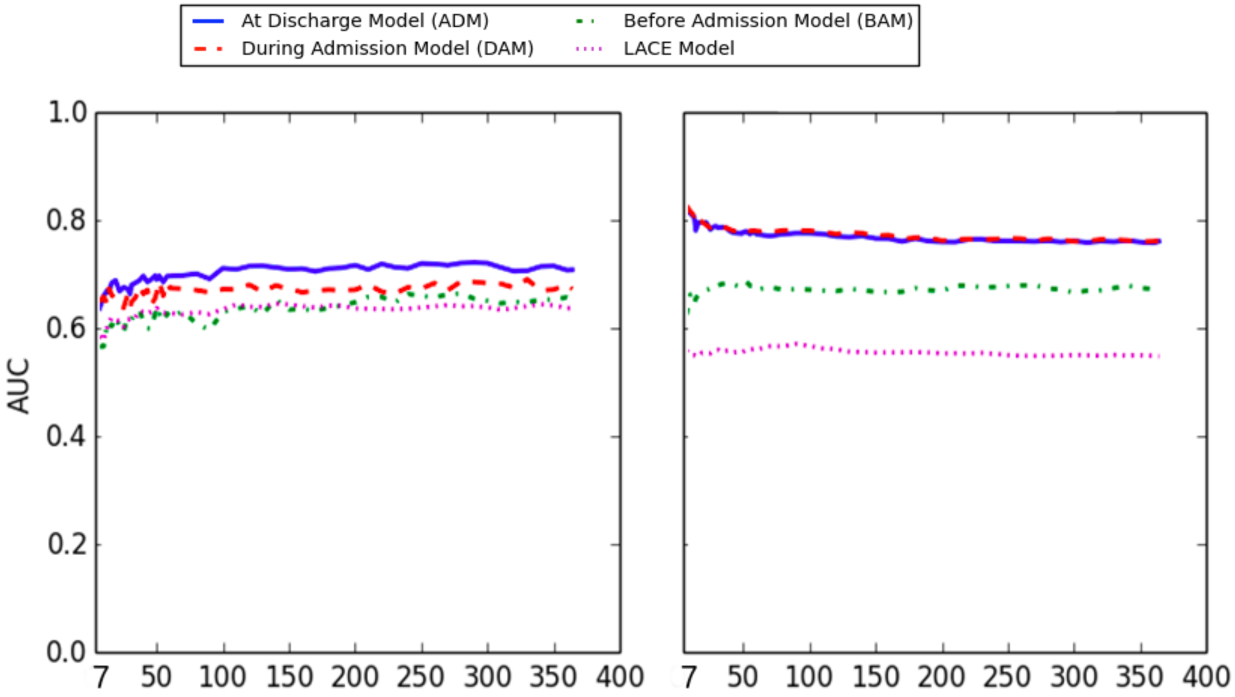
Figure 3.4 Survival rate for patients within one year from the diagnosis

Before Discharge Model Results

Figure 3.5 reports on the AUC for predicting the negative outcome within one year using LACE, before admission, during admission, and at discharge) models for hip fracture and CHF patients.

Across all prediction points, LACE exhibited the lowest AUC values for both cohorts. The before model has similar AUC values as LACE for hip fracture patients, while it has higher AUC than LACE when applied on the CHF cohort. However, the during and at-discharge both outperform the LACE and before models. The during model exhibits almost the same AUC as the at-discharge model when predicting the outcome for the CHF patients. By contrast, at-discharge model has a higher AUC than the during model for the hip fracture patients.

In both cohorts, the before model had the lowest AUC within the first 20 days. After 20 days, the AUC values increased slightly. The during and at-discharge models exhibited the same trend in performance over time. In CHF, the during and at-discharge models had the highest AUC during the first two weeks of prediction. Afterwards, the AUC values decline slightly and smoothly until they reach their lowest point at the one-year prediction. By contrast, the AUC values of the during model and at-discharge were the lowest within the first two weeks for the hip fracture patients and increased slightly until they reached their highest AUC values at the one-year prediction.



Days
(a) Hip Fracture

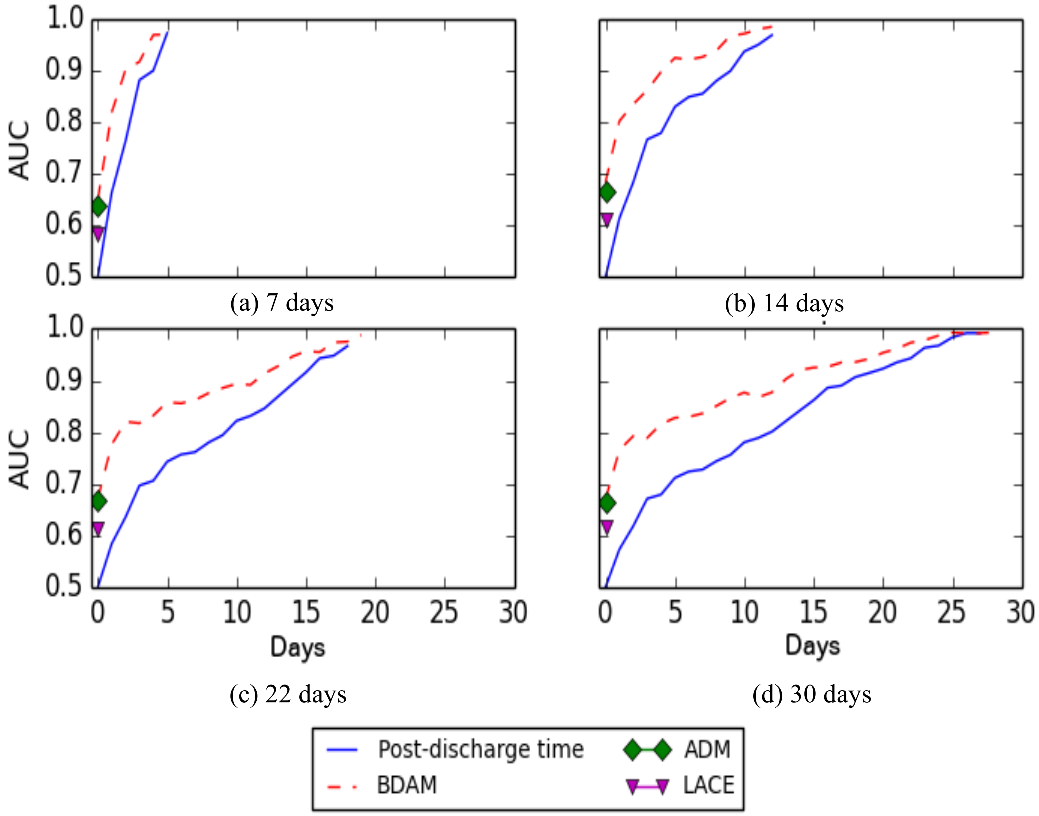
Days
(b) Congestive Heart Failure

Figure 3.5 AUC values for BAM, DAM, and ADM models for outcome prediction within one year. The x-axis shows the predictio at which the model was run. The y-axis corresponds to the average AUC

Post-Discharge Model Results

I analyzed the post-discharge model in two ways: (i) using a single feature representing the days since discharge (post-discharge time), and (ii) all post-discharge data including days since discharge. The results in these two models highlight the importance of time in the post-discharge model (i.e., the longer the patient is out of the hospital, the less likely the patient will have a negative outcome).

Figure 3.6 depicts the AUC values for specific checkpoints for the 7, 14, 22, and 30 days prediction points. The *x*-axis corresponds to the sampled checkpoint. The red dashed-line represents the average AUC for post-discharge model at a given checkpoint, while the diamond and triangle shapes correspond to the AUC of at-discharge model and the traditional LACE model. The solid blue line represents the AUC for the post-discharge time only model. The BDAM model outperformed all models as shown in four figures. Moreover, the AUC values increased as additional post-discharge data are included.



(a) 7 days

(b) 14 days

(c) 22 days

(d) 30 days

Figure 3.6 AUC values for BDAM model applied at different successive checkpoints for hip fracture patients at 7, 14, 22 and 30 days negative outcome

Figure 3.7 depicts the performance of the LACE, ADM (at discharge) and BDAM (post-discharge) models for different prediction points (where the post-discharge data are averaged across the checkpoints). The upper grey line, lower grey line, and solid black line represent the minimum, maximum, and average AUC of 100 sampled post-discharge matrices, respectively. Both the BDAM and the ADM models performed better than the LACE model by 20 - 30% AUC. In this setting, BDAM had a higher AUC than ADM at all prediction points by 15.8 - 26.5% for hip fracture and 9.7 - 12.1% for CHF patients. Using BDAM, predicting the negative outcome within 7 days had a higher AUC compared to predictions within 30 days.

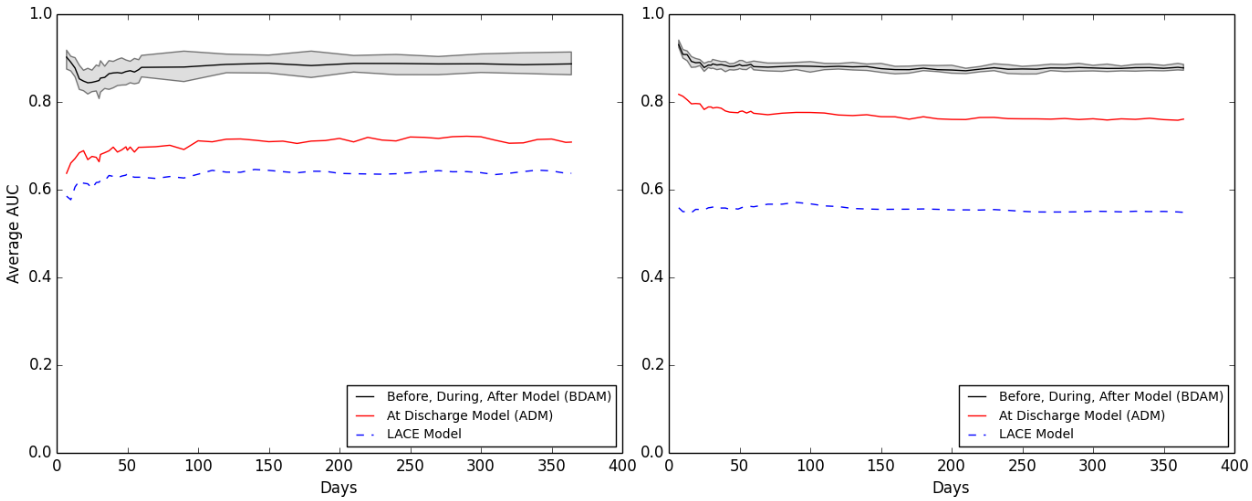
BDAM results trended differently for the two phenotypes. For hip fracture, the AUC decreased until it reached the lowest values at 21 days, then increased between 21 and 60 days, staying relatively constant afterwards. For CHF, the AUC decreased until 21 days.

Feature Importance

For each prediction point, we retreived the features that, on average, exhibited the highest importance. At all prediction points in both cohorts, days from discharge and days until prediction point displayed the highest importance. The number of appointments after discharge was the third most important predictor for a negative event within 30 days while clinical communication was the forth or fifth important feature within the same time range.

In hip fracture patients, age is one of the top 10 predictors of a negative outcome within 30 days. Diagnosing hip fracture patients with infectious, blood stream, genitouribnary, or circulatory diseases during admission were among the top predictors for readmission within 30 days. In CHF patients, the average number of labs during admission (e.g., creatinine, K, Na, and CO_2) were strong predictors for negative events within 60 days.

Figure 3.8 depicts the partial dependency plots for the most important features for 7, 21, 30, and 60 days for hip fracture patients. In the partial dependence plot, the *x*-axis shows the values of the important variable. For a given *x* value, the *y* value indicates the probability of the negative outcome after accounting for the values of the other input variables. A small *y* value indicates a low probability of positive label, while a large *y* indicates a high probability of a positive label.



(a) Hip fracture

(b) Congestive Heart Failure

Figure 3.7 Outcome prediction within one year for a) hip fracture and b) CHF

For example, patients who scheduled a small number of appointments after discharge had a higher probability of experiencing a negative outcome. A low and high number appointments scheduled before discharge were associated with a high probability of negative outcome. In addition, a low and high quantity of post-discharge communications had a high probability of a negative outcome. During admission, patients who were prescribed a larger number of medications and laboratory tests, in comparison to other patients in the cohort, had a higher probability of having a negative outcome. The partial dependency plots of laboratory values such as K, PVC, and creatinine depict that abnormal values were associated with a high probability of a negative outcome.

Unstructured Data Features

Appending the discharge summaries and clinical communications representation did not improve any of the models' AUC significantly. Implementing clinical communication vector in the post-discharge model improved the AUC slightly. Hence, we removed the unstructured data vectors. All the results reported above were generated by models trained on structured data only.

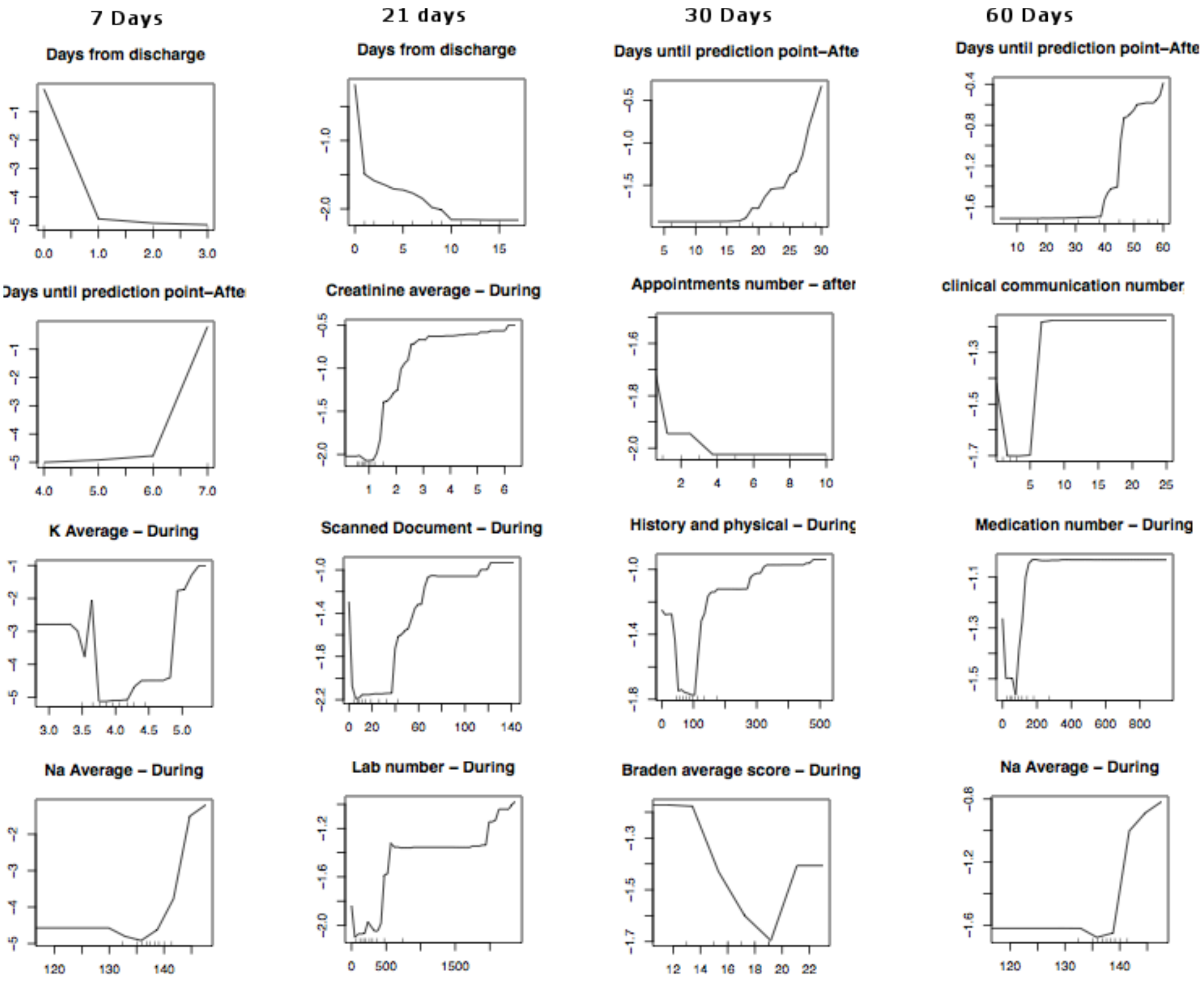


Figure 3.8 Partial dependency plots for outcome predictions for hip fracture patients

Discussion

Predicting negative outcomes and detecting high-risk patients are challenging problems. The findings demonstrate that representing changes about a patient's post-discharge status may increase the performance of such predictions. This finding is further supported by the observation that the LACE model, while simple to implement, is likely to neglect many key features that can enhance predictions.

One of the critical discoveries made in this study is that the post-discharge time is a strong predictor on its own for a negative outcome. This result affirms the observation that the longer a patient remains out of the hospital, the less likely they will be readmitted or die after discharge. Same observation about time was concluded in studies that investigated the breast cancer recurrence, where recurrence rate declined within two years of survival [227–229]. However, the time from discharge in isolation lacks essential clinical information. Including clinical data in the patient's representation can further improve prediction quality. Moreover, time until prediction quantifies the updated post-discharge data missing from the dynamic representation.

One challenge in building the dynamic post-discharge representation is selecting the checking point. Defining the period of collecting post-discharge data controls the amount of updated knowledge that the model gained about the patient. In this work, we proposed a random day model. In practice, a post-discharge model can be executed daily to identify high-risk patients, no matter when they were discharged.

The analysis demonstrates that risk changes over time, especially during the first three months after discharge. In particular, the first seven to ten days are the times when frail patients are at high risk of encountering a negative outcome. A decline of the prediction performance over time suggests a variation in recovery stages for healthy patients and a difference in the risk factors for patients who died or were readmitted within the same time frame. For example, the prediction of the hip fracture outcome had the lowest performance between 14 and 21 days. Several factors could cause this incline such as various pre-existing medical problems, the degree of daily activity, and the ability to attend follow-ups. Further analysis could be done to identify the change in risk factors (e.g., post-discharge complications), and locate the period when those factors are correlated with readmission.

Healthcare providers want to identify the values of the features that are highly associated with the negative outcome. Learning such values may help clinicians understand the reasons leading to a negative outcome and identify early signs of complications. Direct intervention could be applied to lessen risks through outpatient appointments or home visits. In the proposed risk model, the features can be categorized into clinical factors and clinical resources. Both types of features influence the predictions at different levels. The number of clinical resources allocated to treat patients before admission and after discharge could recognize the patients who have a high probability of encountering a negative outcome. The patients who utilized more treatment resources exhibit a higher probability of experiencing a negative outcome except for post-discharge follow-up utilization. A low utilization value implies the existence of barriers preventing the patient from going to see their healthcare provider. Even low utilization values for some features, such as the number of labs ordered during admission, are associated with the negative outcome occurrence. The unexpected utilization could be an indirect measure to identify patients at high risk of encountering the negative outcome.

Summary

The inclusion of dynamic and updated post-discharge data to represent patients can improve the performance of the machine learning model, such as predicting readmission or death. The traditional and static representation such as LACE and at-discharge models focus on constructing features using information available only at the point of discharge or using a small set of pre-defined features. Training a dynamic model on temporal features and post-discharge representation is notable. It shows that static risk prediction methods would benefit from using longitudinal data and updated feature representations. This finding holds true for both acute (i.e., hip fracture) and chronic (i.e., congestive heart failure) patient populations. It is notable that the primary driving factors of our discovery include: 1) time out of the hospital after discharge and 2) the number of the physical (e.g., medication, labs, appointments) and the electronic (communication, documents) resources allocated for a patient.

The proposed method has several limitations. First, this study focused on sampling only one day for post-discharge information. A notable extension is evaluating risk scores on multiple consecutive days to identify changes in risk score and identify patients at-risk to apply early intervention. Second, the feature representation neglected the semantics about the clinical status that might exist in communications between patients and their healthcare providers. For instance, patients who communicate about severe pain or complications from opioid medications may miss their follow-ups. Thus, information intimated to care providers may indicate signs of an impending negative outcome. Using topics or terms frequency in clinical communication did not improve the performance of the model significantly. This result raises the question of whether structuring the clinical communication using standard representations is viable, or a better representation for clinical communication is needed.

CHAPTER 4 SEMANTICS AND CONTEXT IN TEXT CLASSIFICATION

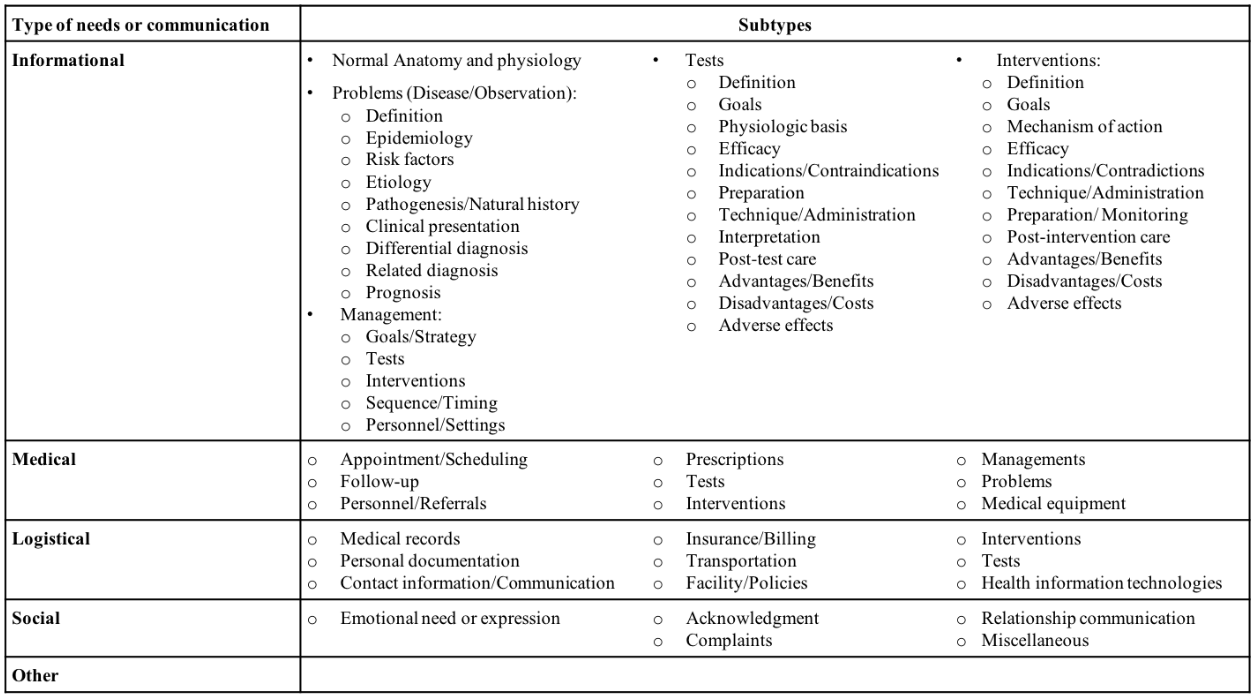
Introduction

EMR systems contain various types of documents which increase the challenge of applying text mining and machine learning algorithms. Applying NLP on clinical notes can unlock and locate information necessary for decision making. Machine learning and NLP algorithms can classify diseases in clinical notes (e.g., suicide, smoking status, colorectal cancer), predict readmission, and extract social factors such as homelessness [230]. In clinical text mining, distinct terms, syntactic, and clinical domain knowledge are ways to represent text. Good feature representations of clinical notes are critical in enabling and enhancing the discovery of the relationship between input and output.

This chapter focuses on evaluating two aspects of clinical text representations: the semantics or the meaning of the words, and the context of surrounding words. Most traditional text representations create distinct features based on their occurrence, grammatical relationships, or rule-based words semantics. Clinical NLP tools such as cTakes and KnowledgeMap combine the variations of medical concepts strings with NLP methods to extract the medical concepts from the unstructured text. Those clinical NLP tools may overlook some concepts due to misspellings, local abbreviations; moreover, those tools do not include the semantic relation between medical and non-medical words, or medical terms that do not exist in the integrated medical terminologies.

Various text mining techniques have been implemented to classify and retrieve information from clinical documents [206,214,231–233]. We selected patient portal messages, an electronic communication between patients and healthcare providers, to evaluate the NLP methods that integrate semantics and contexts. This chapter sought to identify optimal methods to improve the classification of patients' need expressed in portal messages.

Patient portals are secure online systems that enable patients to access personal health information and interact with healthcare systems [231,234–237]. One of most popular features of patient portals is secure patient-provider messaging, a channel through which patients and caregivers can ask questions and receive answers [231,238–243]. Patient portal messages contain diverse content, ranging from important medical questions to social exchanges [231,241,244–247]. One taxonomy of consumer health-related needs divides consumer health communications into informational, medical, logistical, social, and other categories (Figure 4.1) [231,248,249]. Informational communications include questions or answers requiring clinical knowledge, such as the risk factor for a disease. Medical communications request actual medical care including reports of new symptoms to be managed. In logistical communications, patients request pragmatic information such as directions to a facility. Social communications include expressions of gratitude or complaints about a service [231,248,249]. Portal messages often contain multiple categories of consumer health communications. This taxonomy has been applied to questions from patient journals, medical textbooks contents, health-related needs expressed in patients and caregivers interviews, and portal messages [231,248,249].



Type of needs or communication	Subtypes		
Informational	• Normal Anatomy and physiology • Problems (Disease/Observation): ○ Definition ○ Epidemiology ○ Risk factors ○ Etiology ○ Pathogenesis/Natural history ○ Clinical presentation ○ Differential diagnosis ○ Related diagnosis ○ Prognosis • Management: ○ Goals/Strategy ○ Tests ○ Interventions ○ Sequence/Timing ○ Personnel/Settings	• Tests ○ Definition ○ Goals ○ Physiologic basis ○ Efficacy ○ Indications/Contraindications ○ Preparation ○ Technique/Administration ○ Interpretation ○ Post-test care ○ Advantages/Benefits ○ Disadvantages/Costs ○ Adverse effects	• Interventions: ○ Definition ○ Goals ○ Mechanism of action ○ Efficacy ○ Indications/Contradictions ○ Technique/Administration ○ Preparation/ Monitoring ○ Post-intervention care ○ Advantages/Benefits ○ Disadvantages/Costs ○ Adverse effects
Medical	○ Appointment/Scheduling ○ Follow-up ○ Personnel/Referrals	○ Prescriptions ○ Tests ○ Interventions	○ Managements ○ Problems ○ Medical equipment
Logistical	○ Medical records ○ Personal documentation ○ Contact information/Communication	○ Insurance/Billing ○ Transportation ○ Facility/Policies	○ Interventions ○ Tests ○ Health information technologies
Social	○ Emotional need or expression	○ Acknowledgment ○ Complaints	○ Relationship communication ○ Miscellaneous
Other			

Figure 4.1 Dr. Jackson's Taxonomy of consumer health-related needs

Categorizing portal messages into types has many potential benefits, such as prioritizing urgent messages with appropriate actions (e.g., "to reply", or "to do"), and supporting message triage to appropriate personnel or resources [250,251]. For example, a logistical question might be answered by an administrative assistant whereas a medical message would be more appropriately addressed by a clinical provider. Recognizing messages with specific needs could help identify patient problems such as adverse events following a procedure or the need for medication adjustment [252,253].

Given that the number of portal messages is growing, it would be useful to automatically extract a patient's needs to improve care efficiency. Classifying portal messages can be performed in several different ways. Using manual classification, annotators read and assign types to the messages. However, this approach does not scale with growth in message volumes [254–256]. Another approach is asking the patient to select the communication's type from a pre-specified list. However, categories selected by patients to describe content of messages are often inconsistent [241]. A third approach is training a classifier. Cronin and colleagues have previously classified portal messages by representing the messages with bag of words, along with two UMLS values: concept unique identifiers (CUIs), and semantic types (STYs) [257]. They employed rule-based basic to train random forest, logistic regression and naïve Bayes. Their classifiers had an acceptable performance for predicting the category of communications. However, their work has limitations including the use of features lacking semantics and excluding words' context.

Objective

In this chapter, we represented the portal messages using four representations: terms lacking semantics and syntactic, terms and their syntax, context features, semantics and context features. We trained four binary classifiers to identify informational, medical, social, and logistical contents in patients' messages using different features representations. In our evaluation of text representations, we focused on the following questions:

1. Does using text syntax (e.g., verb, noun) improve the classification of portal messages?
2. Does integrating the semantics in the representation enhance the message category identification?
3. Does convolutional neural network learn better features by integrating the words context?

4. Does learning the representation of messages or training more advanced machine learning methods outperform the standard machine learning?

To answer these questions, we trained classifiers on variations of features representation to categorize messages sent via My Health At Vanderbilt (MHAV), a locally-developed patient portal at the VUMC. Our published work demonstrates that creating features that incorporate semantics and context improves the classification and the identification of patients' needs in messages sent to their healthcare provider [88].

Clinical significance

Patients and healthcare providers exchanged, on average, hundreds of messages via MHAV. Although healthcare providers have limited time, they try to respond to patients messages in a timely manner. Training a classifier that identify the types of patients need in the message can help:

1. Reducing the load on healthcare providers who receives patients' messages by prioritizing the important messages that contain medical and clinical requests.
2. Scaling the message prioritization and patients' needs identification to handle the increasing number of patient messages by automating the process and minimizing the manual annotation.

Background

Standard text representation

Converting unstructured text into usable informative structured features is the first and one of the important steps in in NLP pipeline. Text classification has various applications in NLP domains such as information retrieval, sentiment analysis, and web search [194,258–260]. The simplest way to generate structured features from text is bag of words [261]. This method has some drawbacks, including lack of context, treating abbreviations and misspellings as separate entities, splitting multi-word concepts, and ignoring text structure and semantics [262,263]. Although analyzing medical text using standard medical NLP systems such as MetaMap, cTakes, MedLEE considers the medical semantics words, implementing those medical semantics along with standard text representation improves the classification insignificantly [147,231]. Another text representation method is bag of phrases, which uses noun and propositional phrases [264–266]. Bag of phrases appears to provide better representations than bag of words because it conserves the partial ordering of words [266]. Notwithstanding, bag of phrases still has limitations including representing similar phrases as different features, and increasing the feature space [261,266].

Graph-based models extract syntactic and semantic information and convert the text into graphical representations [263,267–271]. The graphs of documents can be converted to vectors but they will be sparse high-dimensional vectors and lack of word context. Luo et al. designed an unsupervised framework to capture relationships between concepts in pathology reports to predict the type of lymphoma [270]. The study demonstrated that SVM classifier that use sentences subgraphs as input outperform SVM classifiers that use n-gram or MetaMap features [270]. In another study, Luo et al. proposed an unsupervised framework that implements graph mining and Non-Negative Tensor Factorization (NTF) to extract features from clinical text [271]. First, the authors converted the clinical text to a graph representation, and identified the important subgraphs (or higher order features) by applying frequent subgraph mining. Then, the authors applied NTF and Tucker factorization schemes to capture similar patients' groups, and similar subgraphs, and identify interactions between groups in different modes (i.e. patients and text subgraphs mode) [271].

Deep learning for text representation

Paragraph2Vec is unsupervised method that learns a fixed-length dense vector for each variable-length paragraph [200]. The model concatenates a paragraph vector with word vectors to predict the next word. Hence, the word's context, as well as the paragraph's context, contribute to word prediction. Paragraph2Vec generates two embeddings: a paragraph embedding and a word embedding. Paragraph2Vec has been implemented to classify movie reviews in the Stanford Sentiment Treebank dataset and Internet Movie Database [200,272].

Miñarro-Giménez used Word2Vec model to mine medical text and identify the relationships across drugs, diseases, and physiological processes [232]. Miñarro-Giménez compared the extracted relations with manually curated gold standard relations from the national drug file-reference terminology [142,232]. Another research team employed Word2Vec to represent clinical documents and build a readmission prediction model for patients diagnosed with chronic obstructive pulmonary disease [233]. Their proposed model has a similar prediction performance to LACE, a widely-used risk model to predict mortality and readmission that has been explained in chapter 3 [214].

Methods

Gold standard

This was an institutional review board-approved study conducted at VUMC. VUMC launched a locally-developed patient portal, MHAV, in 2005. MHAV offers a suite of common patient portal functions including secure patient-provider messaging and access to selected portions of the EMR [273]. All secure messages sent through MHAV are written to the EMR. VUMC maintains a de-identified version of the EMR including over 20 million MHAV messages in a resource called the Synthetic Derivative (SD) [274]. Patient-initiated MHAV messages between 2005 and 2014 were extracted from the SD. We randomly selected 3,000 messages to include in the study dataset. The content of these portal messages was manually classified using a taxonomy of consumer health needs (Figure 4.1). Creation of this gold standard has been described in [246,275].

Feature construction

We used standard methods to convert the messages into structured features including: bag of words, bag of phrases, and graph-based representation. Moreover, we trained Word2Vec and Paragraph2Vec to learn word and paragraph vectors.

Several methods create representation vectors from clinical documents such as bag of words, phrases, tokens, n-gram, and normalized document (e.g. extracting unique UMLS concepts using NLP) [29,94,151,152]. Bag of words is the most straightforward way to generate structured features from the text [261]. This method treats words as discrete entities and assumes that similar documents have the same words and/or have a similar frequency of the words. Bag of words method has some drawbacks, including a lack of context, treating abbreviations and misspellings as separate entities, splitting multi-word concepts, and ignoring text structure and semantics [262,263]. Another text representation method is bag of phrases, which uses nouns and prepositional phrases [264–266]. Bag of phrases appears to provide better representations than bag of words because it conserves the partial ordering of words [266]. Notwithstanding, bag of phrases still has limitations including representing similar phrases as different features, and increasing the feature space [261,266].

Graph-based models utilize syntactic and semantic information to convert text into graphical representations that possess better expressive features [263,267–271]. The graph node, which is a text component, can be one of the following: part of speech, phrase, name entity recognition (NER), token, or semantic node. If two nodes have a relationship (e.g., "appointment" is a "noun"), an edge connects them. Each document is converted into a graph that can be mapped into a feature vector and can be fed to a machine learning model [263,269]. Limitations of graph models include sparse and high-dimensional vectors and lack of word context. For example, the following sentence: “Is it time to schedule a kidney transplant annual appointment?” can be converted into the a depicted in Figure 2.1 (stop words are removed). This graph is then converted into a Boolean feature vector, and a value is set to 1 for all edges that exist in the graph (i.e., transplant -- appointment = 1).

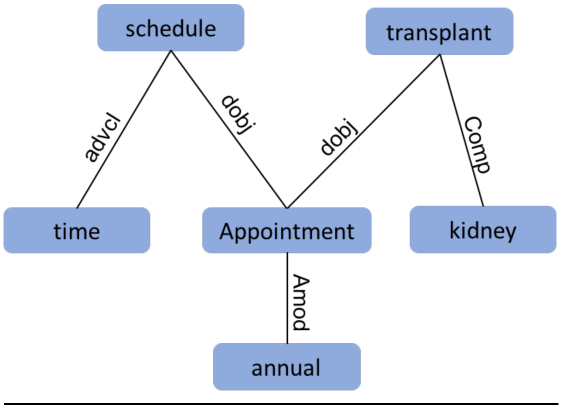


Figure 4.2 A graph representation for the sentence "Is it time to schedule a kidney transplant annual appointment". Advcl: Adverbial clause modifier; Dobj: Direct object; Comp: Compound; Amod: Adjectival modifier

I created four categories of feature representations, as depicted in Figure 4.2. We evaluated the features by training and comparing the classifications metrics for random forest, logistic regression and CNN classifiers:

1- Terms lacking semantics and syntactic: Bag of words, bag of phrases

2- Terms and their syntax: Text graph

3- Context only: CNN trained on random vectors

4- Semantics and context features: CNN trained on Word2Vec vectors

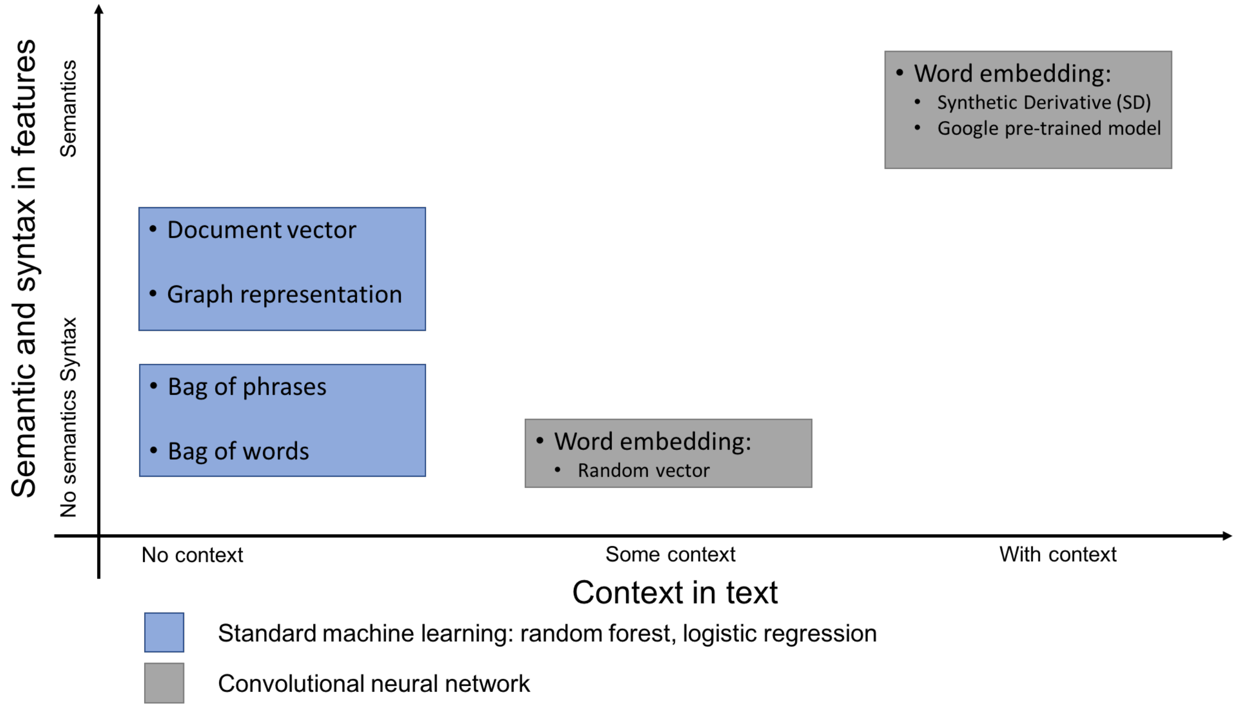


Figure 4.3 Semantics and context in patient portal messages representation

Bag of words (BoW): this feature representation is the simplest method to represent the text but it lacks the semantics, syntax, and context aspects. We extracted the set of words from the portal messages. In the preprocessing step, we excluded stop words retrieved from the natural language toolkit package and words that occurred only once in the entire corpus [276]. We also removed non-alphanumeric characters. We created dictionary that maps a numerical value for each word in the preprocessed messages to create the numerical representation, where the numerical values are not semantically related. We tokenized and represented each preprocessed message with a binary vector, where each the index of the cell is a numerical mapping of the word. The cell was assigned to one if that word existed in the message, else it was assigned to zero.

Bag of phrases: Representing the messages by their phrases preserve small syntactic information restricted by the ones that generate small phrases. Digital Reasoning's Synthesys is an interactive tool that applies supervised and unsupervised machine learning techniques to extract knowledge from unstructured data. The tool implements an entity-centric approach to uncover concepts, events, and relationship between the entities from complex data [277]. Synthesys performs natural language processing to extract entities (e.g., noun, verb, phrases) and facts. Each text is divided into sentences which in turn are divided up into tokens including words and punctuation. The tokens are analyzed for their grammatical roles (i.e., part of the speech). The tool assembles the related entities and tokens, detects synonyms and related concepts, and implements graph analysis techniques, such as associative networks, frequency, and ranking algorithms, to uncover relationships and correlations between entities [278].

Using the Synthesys tool, we tokenized the messages into phrases [278]. We implemented the English engine to extract phrases such as noun and verb phrases. The phrase length varies depending on the number of words forming the phrase. We assigned a numerical value for each phrase. A binary vector was formed for each message, where a cell index maps to the numerical value for a phrase. The cell value was assigned to one if the phrase existed in the message; otherwise, it was assigned to zero.

Graph-based representation: the graph representation accounts for all the syntactic relationships between the words in the message during the feature construction. For each communication, we used the Synthesys to extract the syntactic features from the messages including tokens, POS, and phrases, to form a graph [278]. Synthesys converted the message graph into a binary vector by extracting all relations between nodes. The communications vector in the training dataset had on average 20,868 features.

Word embedding: To construct features considering meaning and semantic relation between the words, we used Word2Vec to represent each word in the communication by a vector. we retrieved the vectors from three models:

Pre-trained model: A Google model that was trained on 100 billion articles from Google News. The Google Word2Vec vectors (https://code.google.com/p/word2vec/) have 300 dimensions. This feature representation includes a general meaning and semantic relations of the words, but not the medical or local semantics of the words.

Word2Vec model trained on a local dataset (SD Word2Vec): To obtain an embedding for the misspelled words and abbreviations in the local dataset, we trained a Word2Vec model using the Gensim package on all SD documents including clinical notes, discharge summaries, portal messages, etc. [279] To train the Word2Vec model, we used 50 as minimum word count, 15 as the window size, and 100 as the number of hidden units, which is also the dimension of the embeddings.

Word2Vec model trained on MHAV messages (MHAV Word2Vec): Different clinical documents have different writing styles [280]. Word2Vec learns word vectors using surrounding words (i.e., context). Hence, Word2Vec model trained on portal messages with the same writing style could provide better word representation compared to models trained on documents with different writing styles. We trained a Word2Vec on 3.3 million MHAV messages (i.e., a subset of the SD) to obtain embedding that consider word context, using the exact parameters of ˝SD Word2Vec˝ model.

Integer Coding vectors: To evaluate the importance of semantics in the Word2Vec embedding for the CNN, we assigned a random integer to each word where the value is unrelated to the meaning of the word. Then, we converted the integer into a binary vector (e.g., 3 is converted to 0011). The generated feature embedding lacks the meaning of the words and the semantic relationship between words.

˝MHAV Para2Vec˝ model: For a paragraph vector, instead of representing the message using features on the word level, Paragraph2Vec infers a representation for the entire message. We trained Paragraph2Vec model on MHAV portal messages only. We assigned the window size to 10, minimum word count to 10, and hidden units to 300. From this model, we used Paragraph2Vec word embedding, that has 300-dimensions for each word, to create word vectors that evaluates the effect of including portal message context in embedding learning. We represented each message by a 300-dimensional vector inferred by ˝MHAV Para2Vec˝ model.

Classification algorithms

We trained four binary classifiers on the four features' representation, one for each communication need, to classify the messages.

Standard machine learning techniques: we trained two standard classifiers: logistic regression and random forest provided by scikit-learn to classify the messages represented by bag of words, bag of phrases, graph vectors, and paragraph vectors [226].

Deep Neural Network (DNN): we investigated whether the complexity of the model or the feature representation would improve the performance of the classification. we trained a DNN to classify the patient messages; however, the length of the message varies, and the input for DNN requires a fixed-length one-dimensional vector for the input layer. To generate one vector for each communication, we aggregated the vectors of all the words in a message using the sum. We fed the vectors into a DNN and a softmax layer that converts the output of the network into two probabilistic output units: positive label and negative label, where the highest probability is the predicted label. We varied the DNN parameters such as the number of layers, number of hidden units, activation functions, to pick the optimal model. We included the details about the ranges of the DNN parameters in section "Model Validation" section.

CNN using Word2Vec embedding: All the previous classifier does not account for the context around the words. Learning a representation for the subsequence of words that include the patient's need would enhance the features. CNN create higher representations for the adjacent input elements within a given window. Applying CNN on text creates higher representations for the subsequence of the words, or the context of adjacent words, as Figure 4.3 demonstrates. Hence, we implemented CNN model as described in [184] and we modified the parameters in the model to fit our analysis. Each message was represented as a matrix-of-words embedding and fed it build a CNN classifier. The CNN model comprises an input layer, a convolution layer, a max pooling layer and a softmax layer. Below, we describe each layer.

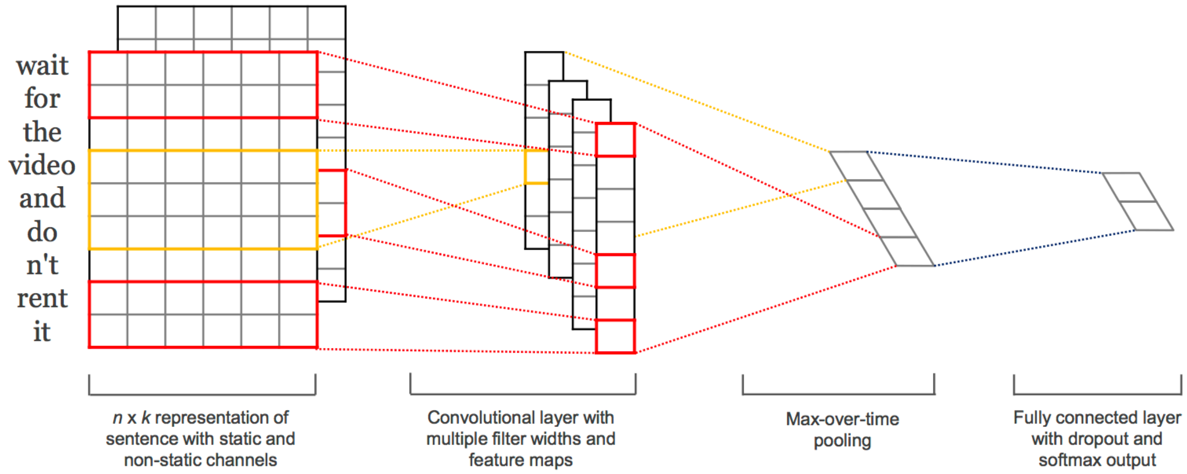

wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 4.4 An example of applying CNN on a text

Let the vector be represented as $\mathbf{w}_j \in \mathbb{R}^{k}$, where $\mathbf{w}_j$ is a *k*-dimensional embedding vector for the *j*-the word in a message. Since CNN requires that all messages have the same number of words, we padded the message with zeros to have the same number of words as the longest communication. Each communication was represented by a two-dimensional *nxk* matrix $\mathbf{c}_i = [\mathbf{w}_1, \mathbf{w}_2, \ldots., \mathbf{w}_n]$, where *n* is the maximum number of words, and *k* is the dimension of the $\mathbf{w}_j$ embedding. A filter **t** with region size *rxk* applies a convolutional operation on the *j:j+r* sub-matrix (i.e. *r* rows between *j* and *j+r* rows) to produce a mapped feature *m* for *r* words. A convolution layer generates mapped feature m_j by applying the non-linear function ƒ:

$m_j = ƒ(\mathbf{t}.\mathbf{c}_{j:j+r} + b)$, where **.** is a dot product.

Where ƒ is a non-linear function such as a sigmoid, **t** is a filter $\in \mathbb{R}^{rk}$, $\mathbf{c}_{j:j+r}$ is a sub-matrix composed of r words, and b is a bias term. Hence, the CNN maps a message with *n* words to *n-r*+1 features $\mathbf{m} = [m_1, m_2, \ldots, m_{n-r+1}]$, where each m is the representation of the words inside the filter. To select the most important mapped features generated by a filter, we used a max pooling layer to pick the feature *m* with the highest value. Filters with different region sizes and convolutional functions can be applied to each communication to obtain multiple representations. The output of max-pool layer for all filters was passed to a fully connected softmax layer with two probabilistic output units: positive and negative labels.

We used the same regularization technique used by Kim [184]. We implemented dropout on the penultimate layer to prevent hidden units from co-adapting and model from overfitting [281,282]. Moreover, L_2 regularization was employed on the weight vectors. If the L_2-norm of the weight vector exceeded a threshold *s*, the vector was rescaled to make L_2-norm equals to *s*. However, we implemented and tested multiple model parameters (listed in Table 4.1 and Table 4.2 below). To evaluate the effect of different semantics of words on the classification accuracy, we represented words using various Word2Vec embeddings trained on clinical documents as well as Google News. We describe the set of parameters used in CNN in the next section.

Model Validation

Searching the parameter space to select the optimal model makes the comparisons of performance between different models on the validation and test sets more reliable. We searched possible parameter values to identify the optimal classifier, optimizing the search process using a Bayesian method. We used Hyperopt package which is a python package that performs Bayesian optimization to tune function parameters and obtain the optimal model [283]. The optimal parameters are selected by finding the parameter combination that minimizes a loss function. The loss value is the prediction error when the model predicts the output of the validation set. To validate models, we randomly divided the dataset into training, validation, and testing with the ratios 0.8, 0.1, and 0.1, respectively. We used Hyperopt to obtain the optimal model applied on the validation set. Table 4.1 summarizes the range of parameters we used in Hyperopt to identify optimal model, and Table 4.2 lists the optimal parameters identified by Hyperopt. We repeated this process five times, and calculated the AUC and accuracy using metrics function in scikit-learn [226].

In the clinical setting, the labels should be assigned to the message. Dichotomizing the probabilities into labels is performed by selecting a threshold to create labels. Precision and recall are metric to evaluate the correctness of the assigned labels. Both metrics are sensitive to the threshold chosen to classify the messages. The standard threshold is 0.5 where any probability that equals or above that threshold has a positive label. Nevertheless, using this static threshold might not suit all the models' output depending on the output probability distribution. Hence, we used the Youden Index to select the thresholds for recall and precision that generated the highest values using the validation set, and we reported both metrics for the test set. The Youden index selects the best threshold by tuning the threshold *t*, dichotomizing probabilities, and calculating the sensitivity and specificity. The threshold that generate the highest value of the equation below is the dichotomizing threshold:

J (Youden Index) = sensitivity(t) + specificity(t) -1.

Where J is function that finds the maximum value of the equation by changing the threshold *t* of calculating sensitivity and specificity.

Table 4.1 Ranges of parameters used in grid search for random forest, logistic regression, deep neural network, and convolutional neural network models

Classifier	**Parameter name**	**Parameter value**
Random Forest	Maximum depth	1 to 30, by 1
	Maximum features	1 to 500, by 1
	Number of estimator	1 to 200, by 1
	Splitting criterion	Gini, entropy
Logistic Regression	Regularization or penalty	L1, L2
	Regulation constant	0.1 to 10, by 0.1
	Fit intercept	True, False
Deep Neural Network	Number of hidden units	100, 150, 200, 250, 300, 350
	Number of layers	1,2,3,4,5
	Drop rate	0.1,0.2,0.3,0.4,0.5
	Learning rate	0.01 to 0.4 with 0.005 increments
	Activation function	Hyperbolic tangent (Tanh), Rectified linear unit (ReLU), sigmoid, linear, softsign
	Initiation function for weights	Uniform, LeCun uniform, normal, Glorot uniform, He uniform
Convolutional Neural Network	Number of hidden units	50, 100, 150
	Filters values	[3,4,5], [4,5,6], [4,5,6,7], [5,6,7], [5,6,7,8], and [7,8,9]
	Learning decay	0.8, 0.85, 0.9, 0.95
	Square normalized limit	5,6,7,8,9,10
	Activation function	Tanh, ReLU

Table 4.2 The parameters of the optimal models identified by Hyperopt.

Classifier	**Parameter name**	**Optimal parameter value**			
		Informational	**Medical**	**Logistical**	**Social**
Random Forest	Maximum depth	16	26	22	29
	Maximum features	235	92	339	343
	Number of estimator	185	133	41	55
	Splitting criterion	Gini	Entropy	Gini	Entropy
Logistic Regression	Regularization	L1	L2	L2	L1
	Regulation constant	0.3	4.5	0.7	0.5
	Fit intercept	True	True	True	True
Deep Neural Network	Number of hidden units	150	350	200	150
	Number of layers	1	2	2	3
	Drop rate	0.1	0.1	0.4	0.1
	Learning rate	0.013	0.025	0.02	0.013
	Activation function	Tanh	Tanh	Tanh	Sigmoid
	Weights initiation function	Glorot uniform	LeCun uniform	He uniform	He Normal
Convolutional Network	Number of hidden units	100	100	150	100
	Filters values	3, 4, 5	3, 4, 5	5, 6, 7, 8	3, 4, 5
	Learning decay	0.9	0.95	0.9	0.95
	Square normalized limit	6	8	6	6
	Activation function	Tanh	Tanh	ReLU	Tanh

Results

Dataset

In the 3,000 MHAV messages corpus, 2,173 (72.4%) contained medical communications; 371 (12.4%) informational communications; 747 (24.9%) logistical communications; and 839 (28%) social communications.

Only 1,867 messages had a singular type, usually medical or social communications. 197 of informational communications contained medical communications as well. Similarly, 359 messages of logistical communications also involved medical communications (see UpSet visualization [284,285], Figure 4.4).

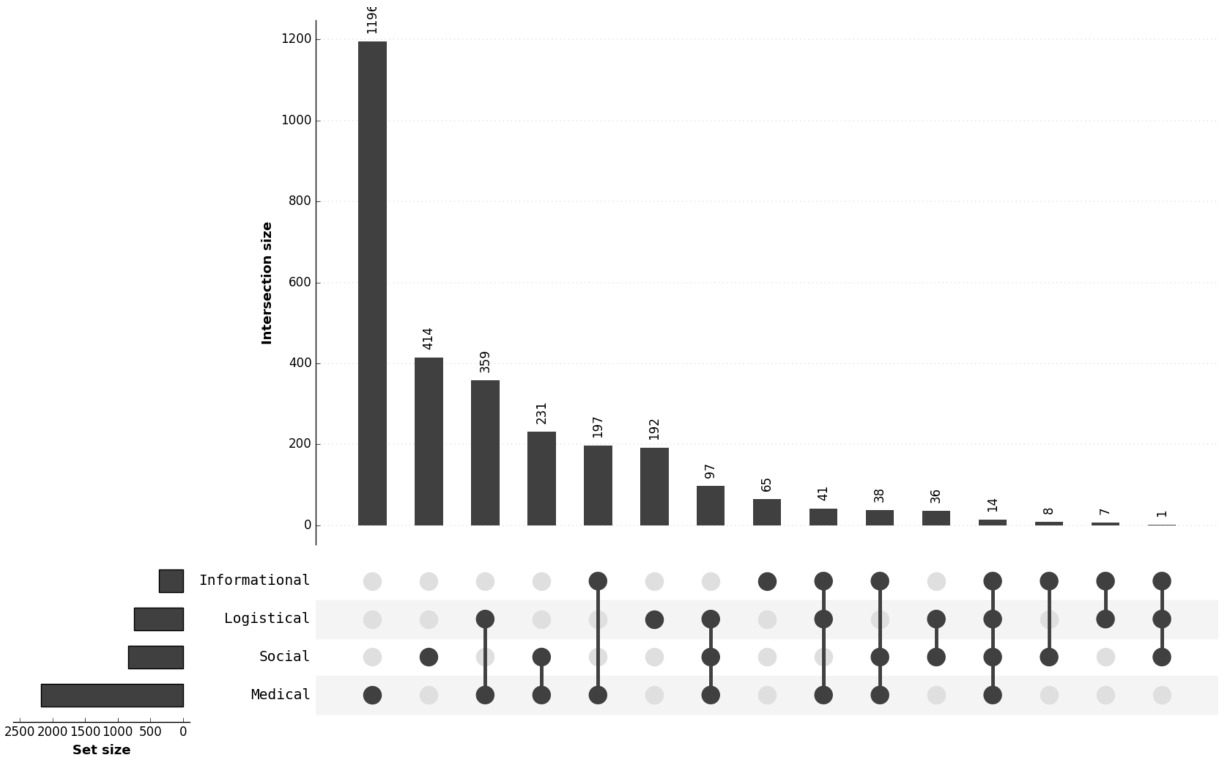


Figure 4.5 UpSet visualization for portal messages grouped by communication types. The small bar graph at the lower left depicts message distribution across categories. Each row in the dot graph corresponds to a category, with columns corresponding to an intersection

The messages consisted of 7,679 distinct words, reduced to 3,371 words after preprocessing. We formed 18,689 phrases, from which 4,362 phrases occurred more than once. The longest communication was a medical message consisting of 823 words. The highest median of words count was 74 for informational messages, while the lowest median of words count was 27 for social (Figure 4.5).

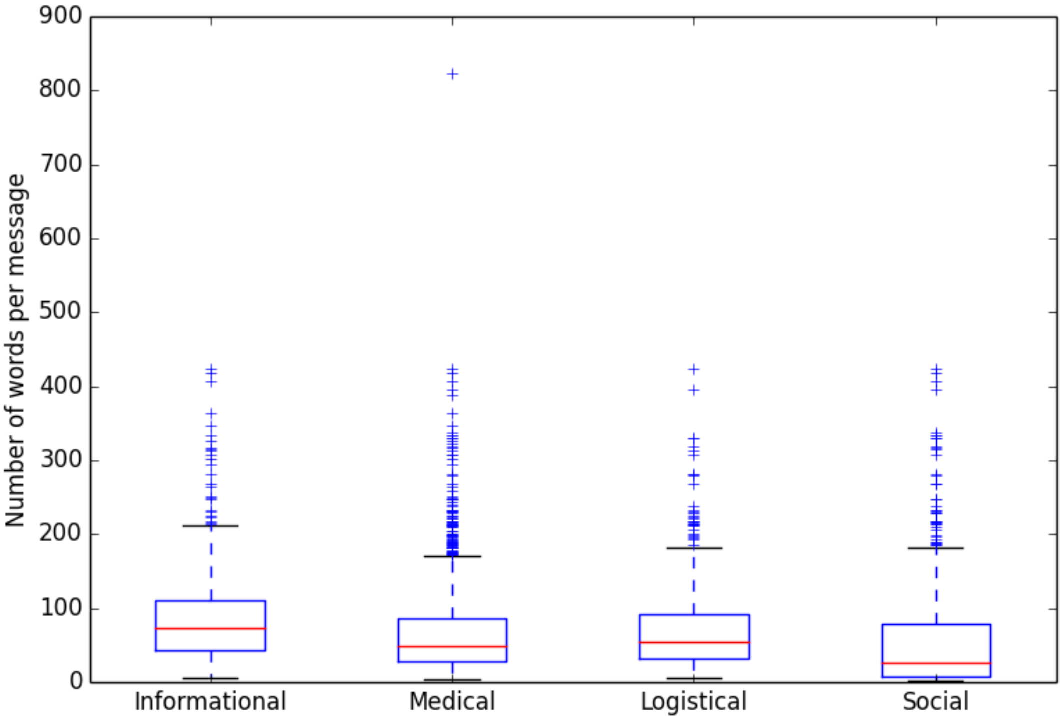


Figure 4.6 Statistical summary for the words in portal messages in informational, medical, logistical, and social categories. Each message is represented by a blue line. Boxes indicates the messages lengths range between first quantile (25th percentile), and third quantile (75th percentile). Lower black line represents extreme values within 1.5 time first percentile, and the upper black line is the 1.5 of third quantile and any value above the black line is considered as an outlier. Red lines indicate the median (second quantile or 50th percentile) of the words counts

Features without semantics: BoW and bag of phrases features

Using BoW that lack syntax and semantics, the AUCs for classifying informational and logistical communications using random forest were 0.8027 and 0.9280, which were higher than the AUC of logistic regression by 0.034 and 0.025%0, respectively. The AUCs for classifying medical and social communications using logistic regression and random forest were 0.885 and 0.827. The accuracy of logistic regression and random forest were around 0.88, 0.83, and 0.77 for informational, medical, and social communications. The accuracy of classifying logistical communications using logistic regression was 0.893, higher than the accuracy of random forest by 0.013.

Using features that lack semantics preserve very basic syntax such as bag of phrases, the highest AUC for classifying informational, social, and logistical communications were 0.786, 0.812, and 0.857 using random forest, while logistic regression classified the medical communications with AUC 0.851. The AUCs for classification of medical, social, and logistical communication using logistic regression were 0.804, 0.79, and 0.858, and were higher than accuracy of random forest for same categories. Table 4.3 summarizes the performance metrics of classifiers trained on bag of words and bag of phrases.

Table 4.3 Performance metrics of classifiers trained on features without semantics: bag of words and bag of phrases. BoW: Bag of words, RF: Random Forest, LR: Logistic regression. AUC = area under the curve, ACC = Accuracy. The highest AUC of classifying a specific category (e.g. social) is bolded

	BoW RF		**BoW LR**		**Bag of Phrases RF**		**Bag of Phrases LR**	
	AUC	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**
Informational	**0.803**	0.877	0.769	**0.880**	**0.788**	0.876	0.761	0.872
Medical	0.884	**0.833**	**0.890**	0.830	0.837	0.725	**0.852**	**0.804**
Social	**0.828**	**0.773**	0.827	0.770	**0.812**	0.725	0.797	**0.790**
Logistical	**0.928**	0.880	0.903	**0.893**	**0.857**	0.775	0.835	**0.858**

Syntax in graph

Random forests classified informational, medical, social, and logistical communications represented by graph vectors with AUCs of 0.832, 0.914, 0.845, and 0.889, respectively, which were higher than AUC of logistic regression (Table 4.4a). For graph representation, logistic regression had higher accuracy than random forest for all categories except medical messages.

Semantics without context in paragraph vector and Word2Vec vectors aggregation

We trained random forest and logistic regression using two features representations built on the semantics of the words more than the context: paragraph vectors and aggregated Word2Vec embedding. Using logistic regression to classify informational and medical communications represented by paragraph vectors yielded 0.763 and 0.845 AUC, which were slightly higher than using random forest by 0.008 and 0.038. On the other hand, using random forests to classify the social and logistical communications represented by paragraph vectors yielded AUCs of 75%, and

78%, and outperformed logistic regression by roughly 0.02 (Table 4.4b). The AUC of DNN using Word2Vec were higher than using paragraph vectors and lower than using graph vectors (Table 4.4c). For DNN, using the words vectors' average instead of the sum yielded similar AUC. Table 4.4 summarizes the metrics of classifiers using graph-based representation, paragraph vectors, and DNN.

Table 4.4 Performance metrics of prediction models that implemented (a) graph representation, (b) paragraph representation, or (c) DNN using sum of word vectors. Bold text represents the highest AUC and accuracy

(a) Graph representation

	Random forest		**Logistic regression**	
	AUC	**ACC**	**AUC**	**ACC**
Informational	**0.832**	0.877	0.826	**0.883**
Medical	**0.914**	**0.849**	0.851	0.746
Social	**0.845**	0.745	0.814	**0.816**
Logistical	**0.889**	0.789	0.866	**0.836**

(b) Paragraph representation

	Random forest		**Logistic regression**	
	AUC	**ACC**	**AUC**	**ACC**
Informational	0.725	0.874	**0.763**	0.876
Medical	0.837	0.831	**0.846**	0.822
Social	**0.752**	0.787	0.736	0.795
Logistical	**0.784**	0.788	0.762	0.784

(c) DNN using sum of word vectors

	Informational	**Medical**	**Social**	**Logistical**
AUC	0.768	0.892	0.833	0.815
ACC	0.861	0.827	0.812	0.787

Semantics and context in CNN classification

Constructing the features with context and semantics, MHAVPara2Vec with CNN had the highest AUCs: 0.909, 0.916, 0.936, and 0.944 for informational, medical, social, and logistical communications, respectively. The CNN using MHAVPara2Vec vectors yielded a higher AUC than the CNN using Google vector by 0.06-0.025 as shown in Table 4.3. Word embedding generated by SDWord2Vec yielded 0.899, 0.880, 0.925, and 0.931 AUC for informational, medical, social, and logistical communications, which were the same or slightly lower than using Google vectors. Using an integer coding vectors that lack the semantics of the words had the lowest AUC compared to the other CNN models that implemented semantically generated vectors. The AUC of CNN using MHAVWord2Vec was not significantly different from CNN using MHAVPara2Vec for all categories except for informational where using MHAVPara2Vec yielded a significantly higher AUC (Table 4.5). A CNN model that implemented MHAVPara2Vec had the highest accuracy except for medical message where using SDWord2Vec had an AUC of 0.892, the highest accuracy.

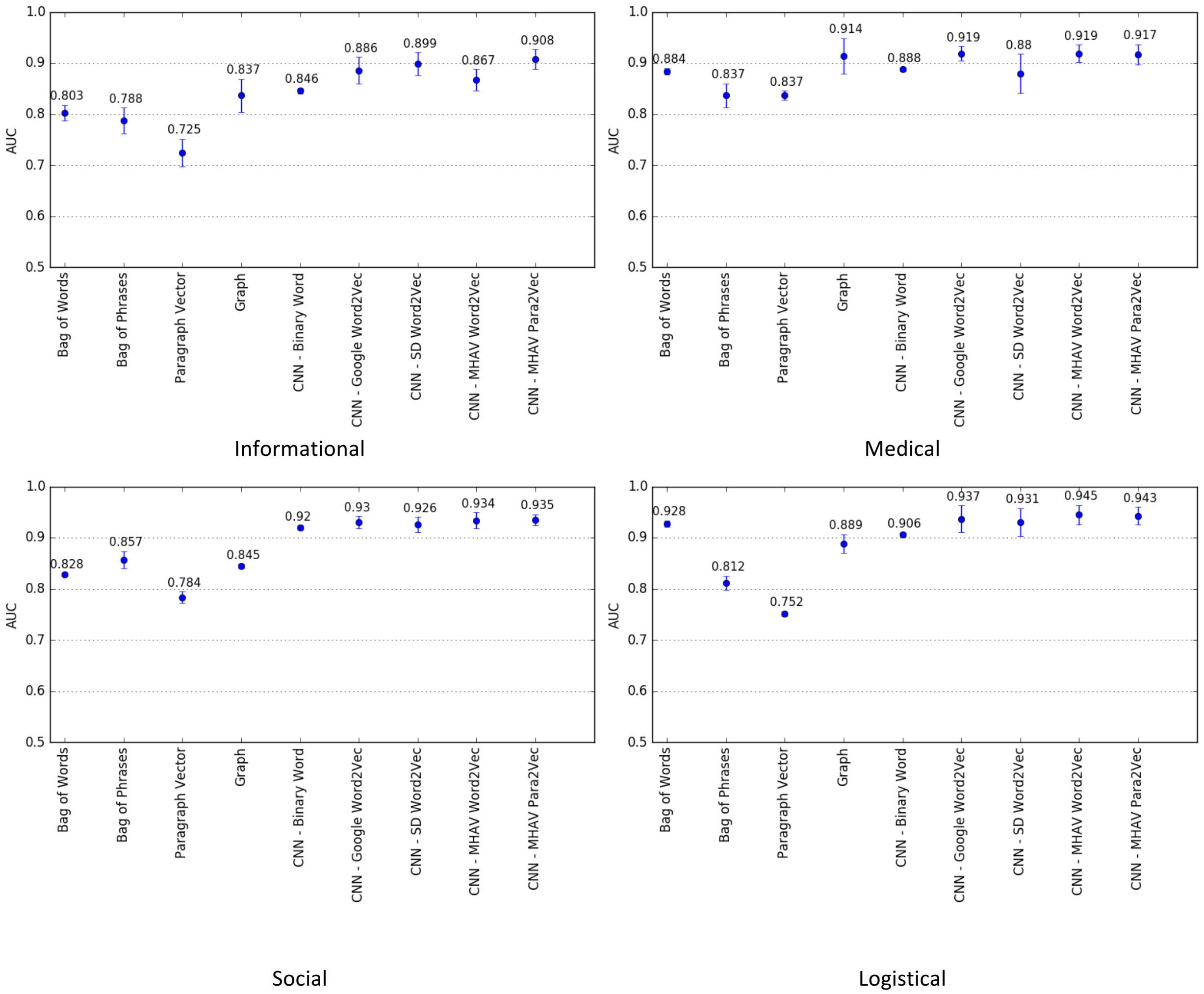
Table 4.5 The performance metrics of CNN models using word vectors as features. AUC = area under the curve, ACC = Accuracy

	Google Vector		**SD Word2Vec**		**MHAVWord2Vec**		**MHAVPara2Vec**		**Integer coding**	
	AUC	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**
Informational	0.887	0.886	0.900	0.851	0.867	0.879	**0.909**	**0.900**	0.846	0.868
Medical	0.919	0.874	0.880	**0.892**	0.919	0.876	**0.916**	0.866	0.888	0.849
Social	0.930	0.889	0.926	0.891	0.935	0.887	**0.937**	**0.896**	0.920	0.884
Logistical	0.938	0.905	0.931	0.906	0.945	0.906	**0.945**	**0.909**	0.907	0.896

Method Comparisons

Using CNN with MHAVPara2Vec embedding had the highest AUC as compared to the other one-versus-all classifiers as depicted in Figure 4.6. Standard bag of words model had AUCs of 0.803, 0.884, 0.828, and 0.928 for informational, medical, social and logistical communications, and outperformed bag of phrases and paragraph vectors. Using the graph representation generated by the Synthesys had higher AUC than bag of words classifier by 0.034, 0.03, 0.017 for informational, medical, and social communications. For logistical communications, the bag of words outperformed the graph representation by 0.03. The AUC for medical and social DNN classifiers were slightly higher than the AUC of the bag of words classifiers. On the other hand, the AUC of informational and logistical classifiers that used bag of words are higher than corresponding DNN classifiers. The variance of all classifiers ranged between 0.006 – 0.035, where graph-based representation had the highest variance and logistical messages using bag of words had lowest variance.

The CNN using random integer embeddings yielded AUCs of 0.846, 0.888, 0.920, and 0.906 for informational, medical, social, and logistical communications, respectively, and they were higher than the AUCs for bag of words classifiers. The highest AUC differences between bag of words and CNN were 0.106, 0.032, 0.109, and 0.017 for informational, medical, social, and logistical communication classification, respectively, when we used the CNN with MHAVPara2Vec.



Informational

Medical

Social

Logistical

Figure 4.7 Comparison between classification models using different features with various complexities. AUC: area under the curve, CNN: Convolutional Neural Network, SD: Synthetic Derivative, MHAV: My Health At Vanderbilt

Compared to other classifiers, the CNN yielded higher precision values that ranged between 0.2659 - 0.3461, 0.9147 - 0.9237, 0.7368 - 0.9182, 0.6839 - 0.7816 for informational, medical, social, and logistical communications, respectively. The precision values of CNN informational and social classifiers were the highest. On the other hand, classifying medical communication using the graph representation had the highest precision with 0.9451, and classifying logistical messages using bag of words yielded the highest precision with 0.8333.

Similarly, the CNN models yielded high recall values with the following ranges 0.8448 - 0.9006, 0.7531 - 0.8556, 0.7912 - 0.8257, 0.7522 - 0.8539 for informational, medical, social, and logistical communications, respectively. Classifying logistical messages using CNN had the highest recall amongst other logistical classifiers. The recall value- for classifying informational communications using paragraph vector representations was higher by only 0.009 compared to CNN informational classifier. Using paragraph vector representation to classify the medical communications generated highest recall value of 0.9082, while using phrases representation generated the highest recall of 0.8810. The following tables summarize the precision and recall values for all classifiers.

Table 4.6 A comparison of precision values for all classifiers. BoW: Bag of words, RF: Random Forest, LR: Logistic regression, CNN: Convolutional Neural Network, SD: Synthetic Derivative, MHAV: My Health At Vanderbilt

Classifier	Informational	Medical	Social	Logistical
BoW - RF	0.245	0.903	0.649	0.750
BoW - LR	0.271	0.912	0.820	0.833
Phrases - RF	0.309	0.891	0.462	0.509
Phrases - LR	0.221	0.920	0.465	0.628
Graph - RF	0.318	0.937	0.529	0.582
Graph - LR	0.297	0.945	0.576	0.682
DNN	0.298	0.870	0.605	0.509
Paragraph vector - RF	0.172	0.873	0.494	0.374
Paragraph vector - LR	0.210	0.874	0.579	0.382
CNN - Integer coding	0.266	0.917	0.747	0.713
CNN - Google Vector	0.293	0.924	0.771	0.782
CNN - SDWord2Vec	0.319	0.915	0.918	0.684
CNN - MHAVWord2Vec	0.325	0.921	0.737	0.726
CNN - MHAVPara2Vec	0.346	0.916	0.786	0.767

Table 4.7 A comparison of recall values of all classifiers. BoW: Bag of words, RF: Random Forest, LR: Logistic regression, CNN: Convolutional Neural Network, SD: Synthetic Derivative, MHAV: My Health At Vanderbilt

Classifier	Informational	Medical	Social	Logistical
BoW - RF	0.727	0.709	0.658	0.761
BoW - LR	0.788	0.847	0.540	0.672
Phrases - RF	0.676	0.825	0.857	0.787
Phrases - LR	0.730	0.793	0.881	0.720
Graph - RF	0.757	0.618	0.750	0.853
Graph - LR	0.730	0.793	0.726	0.800
DNN	0.515	0.888	0.684	0.433

Paragraph vector - RF	0.909	0.908	0.579	0.687
Paragraph vector - LR	0.879	0.847	0.474	0.821
CNN - Integer coding	0.875	0.753	0.791	0.752
CNN - Google Vector	0.901	0.850	0.802	0.782
CNN - SDWord2Vec	0.845	0.851	0.821	0.833
CNN - MHAVWord2Vec	0.845	0.856	0.826	0.854
CNN - MHAVPara2Vec	0.874	0.851	0.820	0.830

Discussion

In this chapter, we sought to improve the classification of patient portal messages by incorporating semantics and the context of the words in creating features. We compared different text representations and various classifiers to identify four semantic types of consumer health communications in patient portal messages. Implementing features that considered the semantics and context of words provided the most accurate classifications and increased the number of correctly categorized messages with important clinical content such as the presentation of a new problem or adverse effects of a medication. Specifically, MHAVPara2Vec along with a CNN improved classification by 0.105 for identifying medical and social content in patient portal messages compared to all other methods. This work improved upon prior classification efforts by Cronin and colleagues [231]. Nevertheless, the precision of the informational classifier is still lower than 0.50, which suggests a room for improvement in classification and text representation.

Classifiers trained using bag of words or phrases (i.e., the crudest, standard feature representations that lack semantics) had acceptable performance but could be readily improved. Although representing messages using phrases preserves the partial order of the words or basic syntactic representation, it did not improve the accuracy of the classifiers, likely due to the large feature space. The results demonstrate that using a graph representation of the text improves classifier performance slightly since it implements the words and all syntactic relations between them. Nevertheless, the similarities between words and their meaning are not included in the graph representation. For example, "appointments" and "appointment" are similar words, but each one of them is represented with a different node and a different edge with the "noun" node.

Moreover, the three aforementioned standard representations do not include the context surrounding words or the semantics of the words.

Creating paragraph vector generates a dense representation that consolidates the sub-contexts of messages into a single vector. This approach is suitable for paragraphs that discuss one topic such as movie reviews. For text that deliver different messages or contain various topics, paragraph vector overlook the local context scattered at different sections of the text, generating a combined text representation for all the different topics in the text. This mixed presentation does not generate distinguishable features to train a classifier. For instance, MHAV has been broadly deployed and adopted across all clinical specialties, and thus its patient portal messages involve a wide variety of health concerns [286–289]. The paragraph vector approach might be more effective if applied to messages from a particular subspecialty, messages that contain one of the patients need only, or messages with mutually exclusive labels.

The analysis indicates that CNN models outperform other models for categorization of portal messages. A CNN's ability to highlight local context regardless of the word location in the message might contribute to the improvement in classification, especially when classifying social and informational communications. For example, a CNN classifier correctly predicted labels for acknowledgment messages: "ok", "Thanks. [Name]", and "That works for me. [Name]". In addition, CNN predicted the category of messages with common words (e.g., "how long until these test results return?"), which are difficult to classify using simple representations such as BoW. Nevertheless, the CNN classifier still misses some positive labels, especially messages that contain both medical and informational content. For informational communication classification, around 27% of false negative messages had medical and informational content.

One of the common classifier errors, especially in binary CNN models, occurs in labeling patients' needs that share very similar sub-contexts. Some informational messages, where patients inquire about a procedure, include medical details. For example, the informational message "Do we need a clinic visit before Dr. [Name] would administer the SI injection?" includes a sub-phrase about SI injection, a procedure (i.e., intervention) performed to diagnose or treat sacroiliac joint dysfunction. The sub-phrase by itself contains a medical need; however, an informational need is detected when considering the entire message. The medical classifier labels the needs in the message as medical since it localizes the sub-phrase, and it predicts only two outputs: medical and non-medical. Increasing the size of CNN window might improve the performance for sentences with longer context since the window of creating context representation will be wider; however, it might reduce the accuracy of classification for short messages.

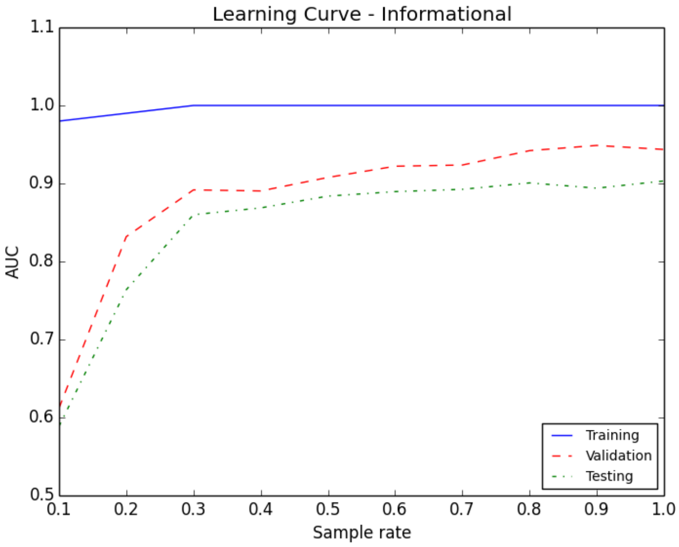
Another example is misclassifying non-logistical messages that include facilities and contact information such as locations names and phone numbers. For instance, the following medical message “this evening my medication still were not filled. My pharmacy is Rite Aid in **PLACE. **PHONE”, contains requests for medication refill and information about a pharmacy. The binary logistical classifier detects the location and place and labels the message as logistical. The main limitation of CNN is forcing all messages to have a fixed-length set by the longest one. Portal messages vary in length especially in medical and logistical categories (Figure 4.1), which might affect the performance of this classifier.

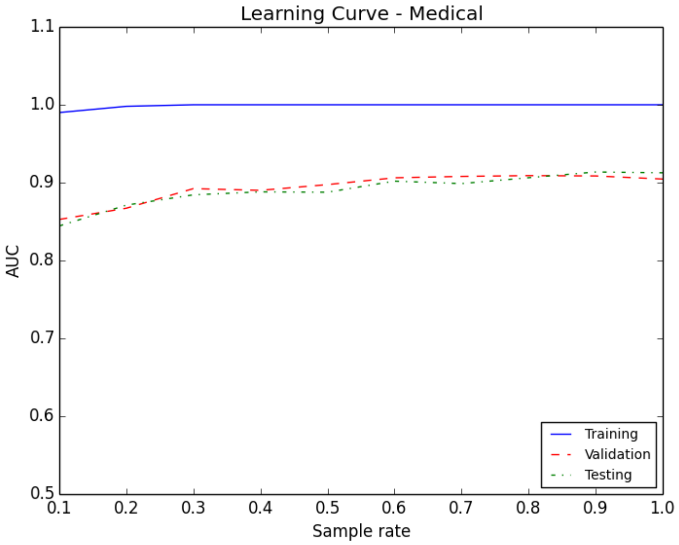
Using a representation that considers the words’ semantic outperforms an embedding that assigns random integer vectors to words. In addition, training the Word2Vec model on a local dataset, especially on documents with the same writing style, generates words vectors that improve classification results as illustrated in Table 4.3. Word embeddings learned from paragraphs, and word context (i.e., Paragraph2Vec) also provides vectors that improve classification performance for some categories.

The methods in this chapter have certain limitations, which can be broadly divided into methodological and dataset limitations. First, the dataset was relatively small, and more labeled data could improve classification accuracy and precision. The size of the data might raise a concern about its generalizability when training deep learning models. However, in some cases, obtaining a larger, manually labeled dataset can be an increasingly labor-intense and costly task. Nevertheless, choosing the appropriate architecture of the CNN model can lessen this problem. We plotted a learning curve to further understand the impact of the training dataset size.

A learning curve summarizes the behavior of the learning by plotting the training, validation, and test errors (or any other performance metrics) as the size of the training dataset increases [290]. The x-axis represents the percentage of the full training dataset sampled to train the model. The y-axis represents that AUC of training the model of the corresponding portion of the training dataset. For a well-trained model, the error or accuracy of validation and test set will be the same as to the training set when the size of the sampled training dataset is closer to 100%. The validation and test curves approximate to each other in a generalizable model. The learning curves of CNN models in Figure 4.7 show that adding more data can improve the informational and social classifiers. The medical and logistical classifiers might improve but the differences between the test and the validation sets demonstrate that those two classifiers perform well on new datasets as shown in the second and last sub-figures in Figure 4.7.

During training, we tuned the weights of the word embedding to predict whether a specific need is expressed in the communication message. Tuning the parameters and performing the grid search are the primary keys to repeating the analysis on another dataset. We computed learning curves for the optimal CNN models trained on MHAVPara2Vec to understand the impact of the dataset size. Learning curves demonstrate the predictive performance of a model by showing the relationship between the number of training cases and the performance metrics (e.g., accuracy, AUC) [291]. We observed that the learning curves of AUC plateaus as depicted in the learning curve below. These plots indicate that the potential improvements from adding more training data might improve the model slightly, especially for medical and logistical classifiers. Several studies examined whether a CNN could be trained on a small image dataset to generate an acceptable model [292,293]. The studies demonstrated that identifying the correct CNN architecture and the appropriate parameters such as initialization of weights could improve the generalizability of the CNN model trained on small datasets [292,293].


Learning Curve - Informational
AUC
Sample rate
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Training
Validation
Testing


Learning Curve - Medical
AUC
Sample rate
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Training
Validation
Testing

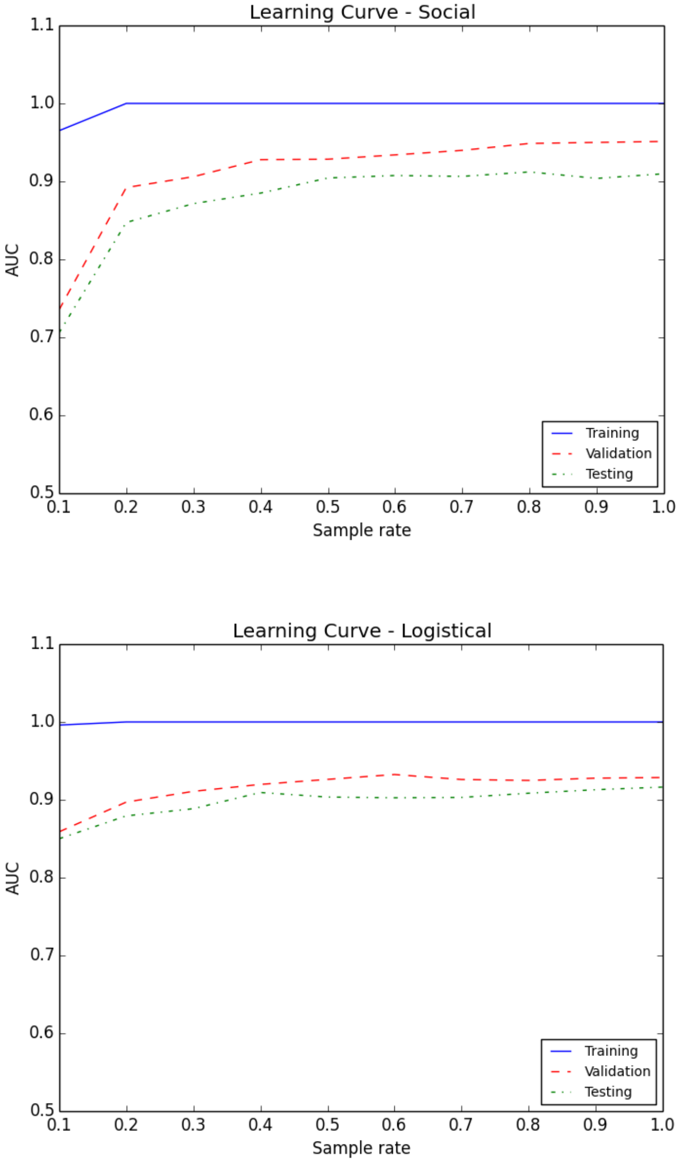


Figure 4.8 The learning curves for the "Para2Vec MHAV" CNN classifiers: informational, medical, logistical, and social classifiers

Second, we trained a binary classifier to classify multi-label messages. Messages that contain multiple labels can be misclassified. Multi-label classification is a growing field, and there is active work to extend single label classifiers such as CNN to multi-label classifiers. Training a multi-label classifier that detects the different types of needs expressed in patient messages is another possible extension to our study.

Third, the study was conducted using data from a single tertiary care institution in the Southeastern United States, and regional dialects or unique problems addressed at this regional referral center may influence the content of the portal messages and thus, classifier performance. Finally, we used de-identified dataset where main identifiers such as names, phone have been replaced by tokens (e.g. [NAME], [PHONE]). The performance of classifier for some categories might be biased due to the existence of a specific de-identification words such as [PHONE], which appeared in more than half of logistical messages.

Summary

Learning the text representations including the semantics and the context of the words may improve the text classifiers performance. This improvement could help triage and prioritize important message sent via patient portals, detect satisfaction messages with complaints or negative feedback, or provide possible topics for messages between providers. This aim demonstrated that classification models that use bag of words have lower accuracy than graph-based representations that consider syntactic aspects in the portal messages. The best classification results are attained when incorporating semantics and context by applying CNN to feature vectors generated by Word2Vec and Paragraph2Vec. This research shows that considering both the semantics and context of the message improves the identification of the types of communication in patient portal messages.

It is worth noting that the clinical communication messages we used in this study have been manually cleaned of the noise and irrelevant information. Hence, we attempted to represent noisy clinical communication messages using word embedding and predict the risk of readmission using CNN. The AUC we obtained from this experiment did not exceed 60%. Further investigations revealed that patient messages were too noisy as unprocessed messages contain the entire thread of replies. Hence, clinical documents to filter out irrelevant information is necessary.

CHAPTER 5 SEMANTICS AND CONTEXT REPRESENTATION IN RETRIEVING RELEVANT CLINICAL INFORMATION

Introduction

EMR contain an abundance of narrative data. EMRs often organize clinical documents by date or title. Providers must sift through documents to compile information required to plan a treatment, or assess a disease prognosis [294]. Reading documents to distinguish between relevant, irrelevant, or duplicate information is an overwrought and time-consuming task, especially for complex patients [85,295]. Information overload and locating data of-interest are difficulties central to electronic clinical documentation [85].

Information retrieval algorithms were developed to search and extract information from clinical documents [80,296–300]. NLP algorithms automatically extract clinical entities and variables that can be useful to detect patient phenotypes [168,301–303]. These techniques still have challenges in extracting clinical data due to the large volume of data, complexity, variability, and domain-specific language in clinical documents [36,80,304]. Moreover, training models that use traditional NLP algorithms requires annotated datasets, which need extensive resources to create [36]. The generalizability of the models depends on the quality of the annotated dataset due to the variability in the annotation [36,303]. Addressing the gap in implementing semantics to retrieve clinical information is still in the early stages.

This chapter focuses on learning representations for sentences in clinical documents that can extract sentences related to a phenotype. Researchers have been training deep learning models to learn words and documents to classify text, answer questions, and summarize text [184,191–193,195,200,201,272,305,306].

Deep learning models can handle complex relations and require fewer hand-crafted features [306]. Deep learning models also outperform traditional NLP algorithms in a clinical and non-clinical text [88,202,307]. In the previous chapter, the second aim proposed a CNN model to learn a higher representation for the portal messages. CNN model requires all the inputs to have the same dimensions. Sentences and documents vary in length which can raise some concerns over implementing CNN. Another CNN limitation is the inability to learn long contextual information and creating a representation only for the part of the text inside the filter window [203]. Recursive models, such as RNN, overcome those two problems by learning a representation that accounts for the context regardless of the input lengths. This work proposes deep learning model that learns a representation for the sentence in the clinical documents by leveraging the phenotypes assigned to the documents. The model integrates the semantics embedded in the word vectors, and the context persevered long-short-term memory model to learn a phenotype-dependent representation for sentences in clinical documents.

Objective

The primary aim was to develop an approach that automatically extracts information relevant to a phenotype from clinical documents, without using an annotated dataset. The extraction model learns the complex association between phenotypes and sentences in the clinical documents, as depicted in Figure 5.1. A sentence in a clinical document is relevant to a phenotype if it contained information about the phenotype, including medications, procedures, symptoms, or diagnostic tests. To achieve this objective, the analysis is focused on the following questions:

1. Can the extraction model use the correlations between the documents' phenotypes and the sentences in the clinical documents to identify sentences relevant to a phenotype?
2. Does the granularity in defining a phenotype influence the accuracy of extracting the relevant sentences?
3. Does the learned representation outperform the traditional features such as phenotype keywords in identifying and extracting relevant sentences?

Clinical significance

Training machine learning models that extract sentences and information relevant to a disease have been studies over the past decades. Training the machine learning model require annotated datasets which are few espcially public or shared ones, Medical infomratian need a machine learning approach that does not require annotation. The signifcance of our proposed model is:

1. Highlighting and extracting the clinical data about a phenotype from a document.
2. Presenting the extracted sentences to healthcare providers before a visit.
3. Upgrading the current scanning process by presenting the documents and the sentences that contain the relevant clinical information regardless of their type or insertion date.
4. Establishing a search tool or information retrieval based on the phenotype.

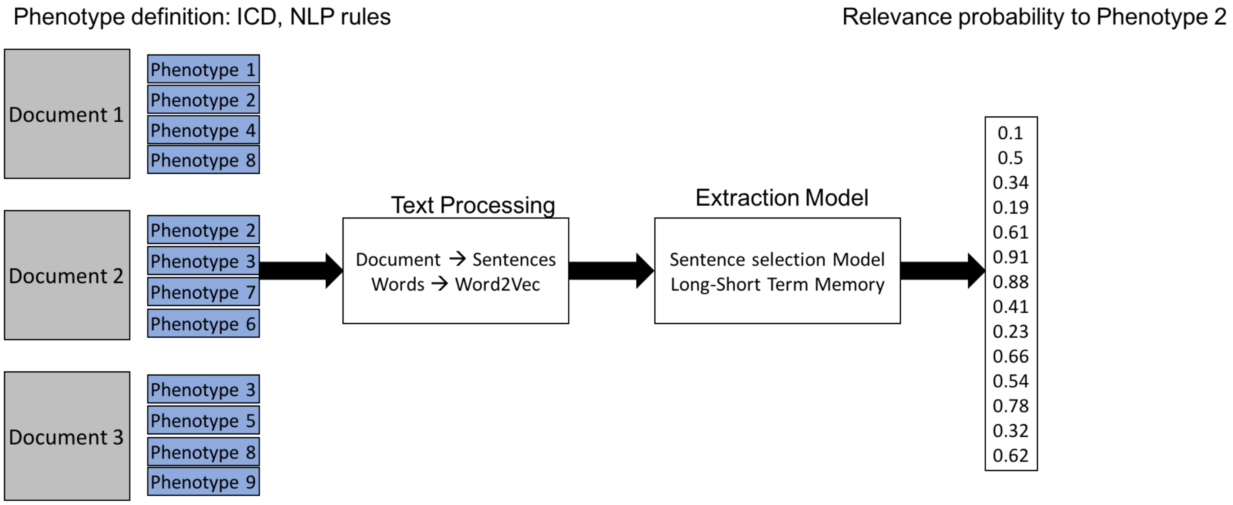


Figure 5.1 Pipeline for training the phenotype sentences selection model using phenotypes assigned to documents and sentences inside the documents

Background

Clinical documents are heterogeneous data sources. Healthcare providers document patients' symptoms, medical history, and medications in a narrative format. This flexibility and the freedom in documentation can encumber extracting the data needed for subsequent clinic visits or research [308]. Manual data extraction from documents is a laborious task that requires time and effort, which both are valuable assets in the clinical environment [36,309].

Traditional information extraction methods rely on manually created features which could limit their efficiency [310]. Rule-based tools, such as KnowledgeMap, cTakes, and MedLEE, combine syntactic analysis, regular expressions, and medical terminologies such as unified medical language system (UMLS), LOINC, and RXNorm to extract relevant medical concepts and structure clinical documents [144,146,147,164,257,311]. Rule-based models might overlook some concepts since clinical documents usually contain local abbreviations, acronyms, and misspellings [312]. Moreover, these models depend on medical terminologies that require ongoing maintenance and all possible variations of terms which can be enormous [313,314].

Structured data and simple NLP rules are used to extract notes for cohorts, retrieve relevant entities, and assign phenotypes to clinical documents [100,304,315]. Most structured data such as ICD and current procedural terminology (CPT) codes are reported in clinical documents. The codes are inserted in a structured format within a time window, that varies from a few hours or couple of days, from the time of creating clinical documents. The order of structured data inside the document does not match their chronological order. Hence, the structured data cannot be directly mapped or linked to its location in the document.

Machine learning models automate extraction of clinical data from documents. However, training the extraction model still requires annotated datasets. Creating an annotated dataset requires human resources, domain expertise, and time, which can be hard to obtain [301]. As a result, training datasets are small and cannot be shared due to various concerns, such as privacy and annotation standards issues [37]. Small datasets, lack of inter-institutional sharing, and lack of standardization can limit the generalizability of information extraction models. A solution that lessens some of the annotation burdens is now a necessity in the era of big data.

Semantics in informational retrieval

Several studies demonstrated the effectiveness of the words' semantics in developing the retrieval system. One approach exploited the UMLS concepts and their co-occurrence in the same document to calculate a semantic relatedness score of concepts to a queried concept [316]. Another study compared retrieving episode of care in clinical notes and discharge summaries using random index vectors (i.e., one-hot vectors), and Word2Vec vectors [317]. The authors learned three different context vectors: adding the index vectors for surrounding words, adding note index, and adding ICD index to each random index vector. They compared the scores learned from context vectors to Word2Vec vectors. The authors found that Word2Vec or semantics had the highest precision in retrieving information related to clinical concepts [317].

Deep learning and natural language Processing

The last chapter introduced some applications to deep learning in NLP and the study in the previous chapter implemented CNN to create higher representations for the text and classify the patient messages sent via patient portals. RNN models apply the same computation at each sequence element and include all previous sequence elements [318]. LSTM, an RNN variation, specifies the amount of data from previous and current sequence elements that are included in the calculation [319,320]. Researchers have applied RNN and LSTM in NLP to analyze sentiment, perform entity recognition, and classify text [321–323]. Gao, *et* al. trained a hierarchical LSTM with attention model that identified sentences related to cancer in pathology reports in the SEER dataset [324]. Lui, *et*. al trained RNN model on an annotated dataset to extract clinical entity and personal health information [325].

Materials and methods

We trained the model to select relevant sentences from clinical documents for two phenotypes that might benefit from locating relevant information: AMI and cancer. AMI is an acute, specific condition that can have very long outpatient notes. Ischemic heart is a less specific condition that may have broader narratives in a clinical narrative for AMI. Cancer is a chronic condition where patient information such as complications and toxicities is extracted to develop evidence-based therapy regimen [326]. Healthcare providers would normally scan clinical documents to extract relevant sentences for AMI and cancer.

For this approach, we split documents into sentences to avoid processing very long input. We trained a sentence extraction model using a LSTM with attention model to distinguish relevant phenotype sentences (e.g. "A 50 years old male patient with stage 3 lung cancer") from irrelevant sentences (e.g. "patient has one daughter", "patient had renal failure") as depicted in Figure 5.2. LSTM accounts for word context and processes the sentence based on the classification output.

The model processed and labeled each sentence individually. Sentences were considered relevant if they included clinical entities associated with the phenotype of interest such as medications, diagnostic tests, procedures, symptoms; while irrelevant sentences did not have relevant entities.

Dataset

We extracted discharge summaries and clinical notes from the SD, a de-identified version of VUMC's EMR, from 2008 to 2013 [274]. The documents are associated with a set of phenotypes, as shown in Figure 5.1. We chose three cohorts to train three sentence extraction models:

Cohort 1: Acute Myocardial Infarction (AMI)

Since discharge summaries recap the treatment that AMI patients underwent, WE trained an AMI sentence extraction model on case documents that are the discharge summaries inserted in the patient chart up to 10 days after the ICD9 code 410.X, allowing to capture 90% of AMI patients [327]. For instance, if the 410.9 code was inserted on 03-04-2009, the patient's discharge summary written between 03-04-2009 and 03-14-2009 is a case document. Control discharge summaries belonged to patients diagnosed with other phenotypes. In total, WE included 5,038 case summaries and 177,813 control summaries.

Cohort 2: Ischemic Heart

A phenotype can be defined with high specificity using ICD codes or with lower granularity using the ICD parent codes, PheWAS codes, or ICD chapters [328]. In the AMI model, we applied a specific AMI definition to label the sentences. To evaluate the effect of granularity in defining a phenotype, we used PheWAS or Phecode to generalize AMI to the ischemic heart phenotype. Any ICD9 code within the range 410.0-414.9 has an ischemic heart Phecode. We extracted case summaries inserted 10 days after the ischemic heart code. The control documents were summaries inserted for other codes. We included 30,385 case documents and 152,466 control documents.

Cohort 3: Cancer

We represented all cancers using one general phenotype to obtain a large dataset for training. Using billing codes to extract cancer notes can decrease the positive predictive values (PPV) and sensitivity of the extraction [329]. Assigning cancer phenotypes to notes using NLP rules can increase the sensitivity and PPV [330]. Hence, we extracted cancer notes that: 1) were written by a provider in hematology and oncology, and 2) had all the following keywords: cancer, tumor, stage, staging, hematology, and oncology. To create the control dataset, we sampled clinical notes inserted that: 1) were issued from a clinic, other than hematology or oncology, and 2) had none of the previous keywords. Our cancer cohort dataset had 17,954 cancer notes and 429,358 control notes.

Sentence selection model

We extracted documents for case and control cohorts to train the sentences extraction model. The case documents belonged to patients diagnosed with the phenotype of interest at the time of the documentation during admission, or clinical visit. The control documents belong to patients who were not diagnosed with the phenotype at the documentation time. A patient might have documents in both cohorts if he or she had other diagnoses at different times. We tokenized the documents into sentences and processed each sentence independently as illustrated in Figure 5.2. All sentences extracted from case documents were marked as relevant or positive (i.e. 1) while the sentences extracted from control documents were negative or irrelevant (i.e. 0). Obviously not all sentences in a document are relevant, but this approximation was needed as phenotype information is not specified more granularly.

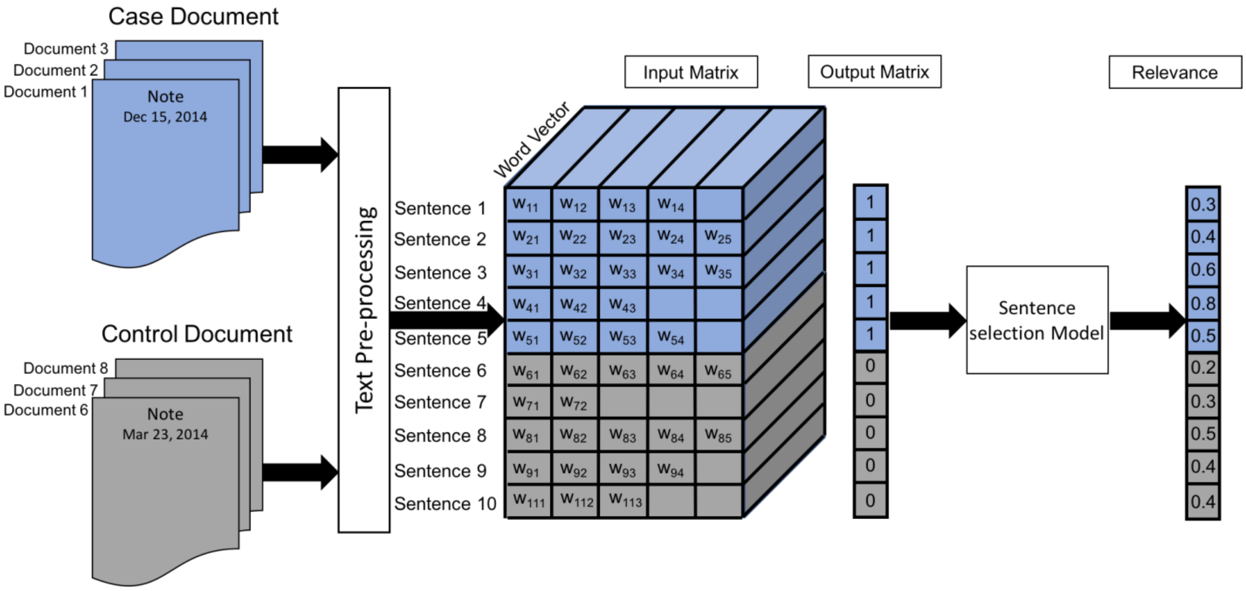


Figure 5.2. Model for extracting sentences with information about the phenotype of interest. Case documents include information about the target phenotype. Control documents include information about phenotypes other than the target. Text pre-processing: tokenizing the documents into sentences, and representing words by word embedding. Input matrix: sentences represented by words' vectors. Output matrix: a label whether the sentence is originated from a case or control document

LSTM Sentence selection model

After splitting the documents into sentences, we removed the non-alphabetical and numerical characters and converted all the words into lower case. We trained a LSTM with attention model, depicted in Figure 5.3, that has four layers:

1. An input layer that represents each word in the sentence using the Word2Vec embedding. To get word embedding for the words, we trained a Word2Vec model on 2.4 million clinical documents in the SD using the parameters: 50 for the minimum word count, 15 for the window size, and 100 for the number of hidden units.
2. LSTM layer composed of sequentially connected LSTM units, where each unit corresponds to a word in the sentence. The units are sequentially connected, where each unit feed its output to its successor [320]. Each LSTM unit at element t is a collection of four vectors in the representation space: input gate i_t, forget gate f_t, output gate o_t, and a hidden representation h_t, as depicted in Figure 5.4. A memory cell c_t allows the LSTM to preserve the cell state for a long time and it is a function of the input gate. The input gate i_t determines the amount of current input x_t to be stored in the memory cell, and forget gate controls the

amount of current input x_t to be forgotten, as inferred from the equations set 1. There are different variations for LSTM; Appendix A provides a description for LSTM and its equations.

3. Attention layer that creates a context vector to identify the words that are highly associated with the output [331]. LSTM layer creates a fixed vector at the last LSTM cell that represents the entire sentence. The relationship or the dependency between the words in the input sentence and the output (i.e., classes, sentence) will be lost [332]. The attention layer is a dense layer that pays attention to each word in the sentence. It applies a Softmax function to relate each word in the sentence to the output, and produces a predictive distribution for the words in the sentence, as Figure 5.5 shows [333]. The context vector of attention provides a dependency score for each word regardless of its location. Appendix A describe the attention layer and its equation in detail.

4. An output layer that applies a Softmax regression to generate a probability value of the sentence relevance or relevance score.

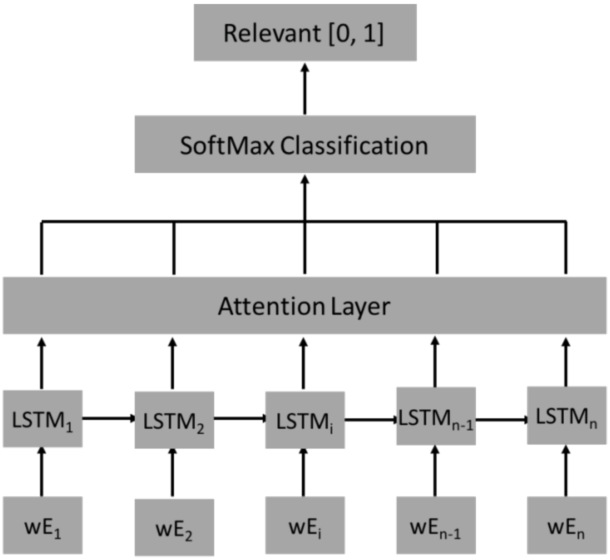


Figure 5.3 Sentence selection model layers. wE$_i$ is word embedding of word i in the sentence, LSTM$_i$ is LSTM unit at step i, n is the number of the words in the sentence

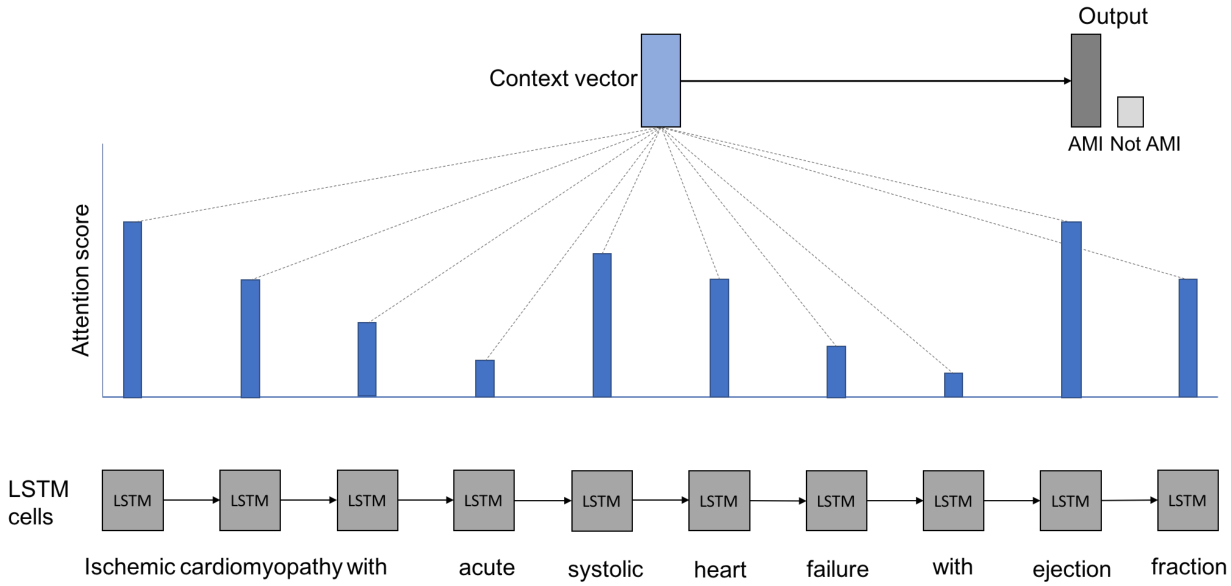


Figure 5.4 The distribution of attention scores for the words in the sentence in relation to the relevance label: AMI or not AMI related. The figure is an illustration for the distribution of attention scores

In the rest of the chapter, a LSTM with attention model and sentence extraction model will be mentioned interchangebly.

Sentence selection model parameters

I employed a human-guided search to select the optimal model amongst the models trained on the following parameters combinations: 100, 200, 300 for LSTM hidden units; 0.001, 0.005, 0.01, 0.2 for learning rate; Tanh and ReLU for activation function; and Adadelta, Adam, and Adagrad for optimization. We trained the model on 90% of the dataset and validated on 10% of the dataset once (i.e. there was no cross validation), for 5 to 10 iterations or epochs. The optimal model parameters were 100, Adam, 0.001, and Tanh for hidden units, optimization, learning rate, and activation function respectively, which is the combination that yielded the highest AUC and accuracy on the validation. We trained the final model for 7 epochs because the AUC and the accuracy of the validation set declined after 7 epochs.

Gold Standard

I assessed the sentence extraction models on an annotated dataset. From the SD, we randomly selected discharge summaries for 100 patients diagnosed with AMI and clinical notes for 100 cancer patients created between 2014 and 2016. To annotate the documents, we recruited three medical student annotators via VUMC PyBossa, a local VUMC crowdsourcing environment [334]. We asked the annotators to select the sentences that had diagnostic tests, medications, procedures, or symptoms related to AMI, or cancer. We standardized the annotation by tokenizing the documents into sentences before the annotation. Sentences were relevant if at least two annotators labeled them as relevant; otherwise, the sentences label were irrelevant.

Model evaluation using gold standard

I compared the LSTM with attention sentence extraction model to two other, established models:

1. Word2Vec extraction model: we created a keyword-based extraction model that extracts relevant sentences using the phenotype's concepts and their similar terms. Using clinician feedback, we defined the possible words that healthcare providers might use when they search for AMI or cancer information. We picked four AMI keywords: heart attack, AMI, acute myocardial infarction, and angina. For cancer, we selected tumor and cancer keywords.

   To expand the keyword search list, we identified similar words for each keyword. Word2Vec model provides a list of the top similar words for a keyword by calculating the cosine similarities between the vectors of a keyword and other words in the vocabulary. A Word2Vec model specifically trained on an EMR and note subsets was used to create comprehensive list of word similarity scores [335,336]. For the above keywords, we attained the list of similar words using the approach previously described and validated by Ye, *et* al [335]. We retrieved and combined the top 3,000 words similar to the AMI keywords and generated a list of 6560 AMI keywords. We repeated the same process for cancer keywords and created a list of 4474 cancer keywords.

   The Word2Vec model attempts to identify sentences with similar words in it to the phenotype. The model assigned a score to each sentence by summing the cosine similarity of the similar words in the sentence. If a word was not in the similarity keyword list, it was excluded. Specifically, for a sentence *j* S_j with sequence of words $w_1, w_2, w_3, w_i, \dots, w_t$, where

w_1 w_2 w_i are the words in the phenotype's keywords list, the following equation generated the relevance score:

$$\mathrm{Word2Vec}_{\mathrm{score}} = S_j \textit{ similarity score using Word2V}$$
$$= \frac{\sum_{i=1}^{n} \mathrm{SimScore}_i(\textit{phenotype keyword}, w_i)}{n}, \forall w \textit{ in } S_j$$

Where:

n: the number of words in both S_j and top similar words list. We excluded words that are not in the list of top similar words

SimScore$_i$(w_i): the similarity score between word w_i and a phenotype keyword (e.g., SimScore$_i$(resection, cancer)

The calculated Word2Vec$_{\text{score}}$ heuristic of S_j is a probability value that ranges between 0 and 1.

2. Combination of LSTM with attention and Word2Vec$_{\text{score}}$ similarity score: we combined the scores of LSTM with attention and Word2Vec$_{\text{score}}$ to evaluate whether the two relevance scores identify the same relevant sentences. We fitted a logistic regression to learn the coefficients α, β in the following equation:

$$\textit{Senetence}_j \textit{ combined score} = \textit{Logit}(\alpha LSTM_{Score} + \beta Word2Vec_{socre})$$

We used the cancer model to extract cancer sentences, and the AMI and the ischemic heart models to identify AMI sentences. The sentence extraction models generated a relevance probability for each annotated sentence in the gold standard. We used the relevance probabilities or relevance scores to calculate the AUC for the three models on the gold standard. In our experiment, we had one gold standard dataset. One AUC value cannot show if the difference between two models is significant. To test the significance, we sampled 80% of the sentences and calculated the AUC for the three scores' models. We repeated the sampling process 100 times. We applied paired t-test to evaluate if the difference between the AUC of the three scores is significant.

The gold standard combine all annotators responses to create the true positive values. The combined standard does not demonstrate the impact of including a specific annotator and or the correctness of the annotations of the corresponding annotators while fixating the labels of the other annotators. The correctness of the labels generated by annotator can be evaluated by introducing noise to the his/her labels, holding the other annotators' labels fixed, and reevaluate the noisy gold standard. Data perturbation, a technique that distorts a dataset, can create a noisy gold standard [337,338]. Perturbation is applied on images or labels to reevaluate and regularize machine learning models to reduce the misclassification or prediction [339,340]. Data perturbation can be applied to show the sensitivity of the model towards possible variations and changes in samples. Assessing the quality of the annotation for the annotators can benefit from a similar analysis. The perturbations are introduced to the labels that the annotators created by switching each sentence from labeled to unlabeled or from positively annotated to negatively annotated. Perturbation can be gradually introduced as the percentage of the perturbed sentences increases. The model performance is reassessed for the perturbed versions of the gold standard. Applying perturbation analysis on the three annotators labeling will evaluate each annotator labels in the gold standard.

Results

I trained AMI and ischemic heart sentence extraction models on 11 million sentences and trained cancer sentence extraction model on 12 million sentences. The percentages of positive sentences in training were 2.7% for AMI, 16% for ischemic heart, and 20% for cancer, as reported in Table 5.1. The models were validated on 1.2 million AMI and ischemic heart sentences, and on three million cancer sentences.

Table 5.1 Documents description for AMI and cancer datasets.

Cohort	**Maximum sentence length**	**Training sentences**		**Validation sentences**	
		All (n)	**Positive (n, %)**	**All (n)**	**Positive (n, %)**
AMI	416	11,090,556	298,842 (2.7%)	1,232,284	51,536 (4.1%)
Ischemic heart	416	11,090,556	1,807,832 (16%)	1,232,284	231,163 (19%)

Cancer	218	12,090,000	2,411,025 (20%)	3,022,501	651,984 (22%)

Model evaluation using annotated dataset

The annotated dataset included 100 discharge summaries that had 16,513 sentences total and 100 cancer notes with 31,926 sentences total. Kappa and F1 or agreement scores between annotators in identifying AMI relevant sentences were higher than annotating relevant cancer sentences. Annotator 1 had the highest agreement scores with other annotators with kappa scores 0.606 in AMI documents and 0.558 in cancer notes, as reported in Table 5.2. This annotator also had the highest agreement with the cancer sentence extraction model and one of the highest agreement scores with AMI sentence extraction model (the difference AMI kappa scores between Annotator 1 and Annotator 2 was less than 0.007 as Table 5.2 shows).

Table 5.2 Agreement between annotators, and between the annotator and our sentence extraction model using Kappa and F1 scores.

	AMI		**Cancer**	
	Kappa	**F1**	**Kappa**	**F1**
Annotators 1 and 2	0.606	0.686	0.154	0.184
Annotators 1 and 3	0.576	0.676	0.559	0.633
Annotators 2 and 3	0.562	0.650	0.120	0.153
Annotator 1 and model	0.183	0.233	0.133	0.344
Annotator 2 and model	0.190	0.237	0.014	0.053
Annotator 3 and model	0.139	0.191	0.095	0.348

Comparing the three sentence extraction models, the AUC of applying AMI sentence extraction model on AMI gold standards was 0.862 which was higher than using Word2Vec scores alone by 0.024, as shown in Table 5.3. Combining AMI Word2Vec and sentence extraction scores using the coefficients 6.93 and 8.01 yielded 0.888 for AUC.

Selecting relevant AMI sentences with ischemic heart sentence extraction yielded 0.8523 which was higher than using Word2Vec scores by 0.025. We used the coefficients 4.3289 and 6.0554 to combine the scores generated by the ischemic heart sentence extraction model and Word2Vec. Combining those scores outperformed both individual models with an AUC of 0.8867. Selecting AMI relevant sentences using AMI model had higher AUC than selecting AMI sentences using ischemic heart model.

Using the cancer sentence extraction model to identify cancer-related sentences yielded AUC of 0.683 (Table 5.3), which was lower than using similarity scores of Word2Vec. Combining the score using 1.1620 and 5.533 coefficients for LSTM and Word2Vec scores outperformed the two models and yielded 0.822 AUC.

Using the median of words' scores to calculate the $\text{Word2Vec}_{\text{scores}}$, instead of the mean, yielded lower AUC. However, using the maximum score of the words in the sentences as $\text{Word2Vec}_{\text{score}}$ yielded slightly higher AUC by 0.019 and 0.009 for AMI and cancer respectively, compared to using the mean in $\text{Word2Vec}_{\text{score}}$ equation. The p-value of the paired t-test on 100 AUC values for the sub-sampled gold standard dataset shows that the difference between the AUC of the three models is less than 0.05.

Table 5.3 The Area Under the Curve (AUC) values for predicting the relevant sentences in gold standard datasets using: 1) LSTM with attention, 2) Word2Vec similarity score, and 3) combining of LSTM with attention and Word2Vec.

Cohort	LSTM	Word2Vec	LSTM+Word2Vec	
			Coefficients α, β	AUC
AMI	0.862	0.839	6.931, 8.005	0.888
Ischemic	0.852	0.830	4.329, 6.055	0.887
Cancer	0.683	0.811	1.162, 5.533	0.822

Since the kappa and F1 scores were low, we reevaluated the models using a more rigid gold standard (or tails gold standard). The strict dataset included only sentences with full agreement that were selected or were not selected by all annotators. The AUC of predicting the relevant sentences in this rigid gold standard was 0.922 for LSTM in the AMI cohort (Table 5.4), a value 0.06 higher than predicting all the sentences (Table 5.3). Using the strict dataset also raised the AUC for LSTM in the cancer cohort to 0.721 (Table 5.4), 0.035 higher than the previous approach (Table 5.3).

Table 5.4 Area Under the Curve (AUC) for predicting relevant sentences in the strict gold standard datasets using LSTM with attention, Word2Vec similarity scores, and combining scores of LSTM with Word2Vec

Cohort	LSTM	W2V	LSTM+ Word2Vec	
			Coefficients α, β	AUC
AMI	0.922	0.888	10.109, 9.408	0.942
Ischemic	0.912	0.882	6.058, 6.800	0.941
Cancer	0.721	0.864	1.345, 5.846	0.871

We plotted the cumulative distribution function (CDF) for the models' scores to analyze the scores learned by the AMI and cancer sentence extraction models (Figures 5.6-5.8). The x-axes in the plots represent sentences relevance scores. The y-axes represent the percentage of sentences that have scores equal to the corresponding score on x-axis, or less. We created three CDF each for AMI, ischemic heart and cancer phenotypes, the CDF shows the scores of positively-annotated sentences (i.e., selected by annotator, solid lines), and the second CDF depicts the scores of negatively-annotated sentences (i.e., not selected by annotator, dashed lines). For positively-annotated sentences, the closer the corner of CDF plot to the bottom right or high scores corner, the higher the relevance score assigned by the model to related sentences (i.e., higher true positive value). For negatively-annotated sentences, the closer the CDF corner to the top left corner, the lower the relevance scores assigned for irrelevant sentences (i.e., higher true negative value).

As shown in Figure 5.6, for the AMI dataset, around half of the positively-annotated sentences had relevance scores higher than 0.2, while 80% of the negatively-annotated dataset had 0.1 scores or lower. For the ischemic heart model, Figure 5.7 demonstrates that the scores of 65% of positively annotated sentences were higher than 0.5, while 80% of negatively annotated sentences’ scores were lower than 0.4. Figure 5.8 shows the cancer model assigned 0.5 or higher for 77% of positively-annotated sentences; while the scores of 42% negatively-annotated cancer sentences were lower than 0.5.

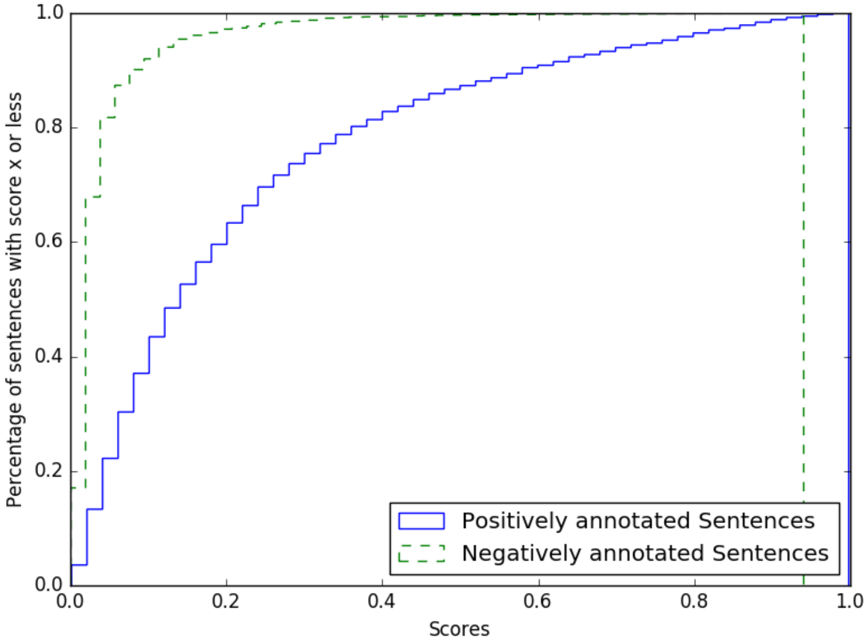


Figure 5.5 CDF distribution of AMI sentences scores generated by LSTM with attention model for positively and negatively annotated sentences in the AMI gold standard dataset

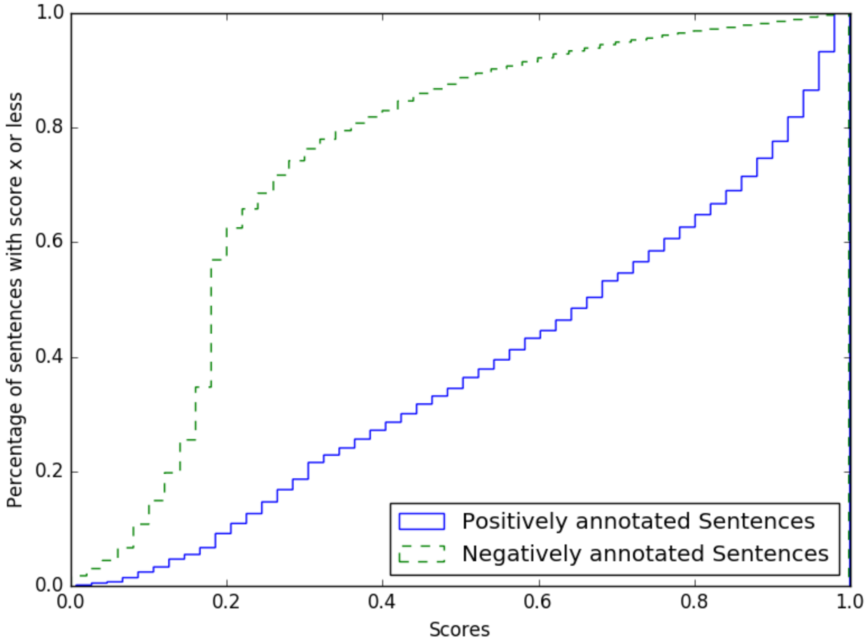


Figure 5.6 CDF distribution of ischemic heart sentences scores generated by LSTM with attention model for positively and negatively annotated sentences in the AMI gold standard dataset

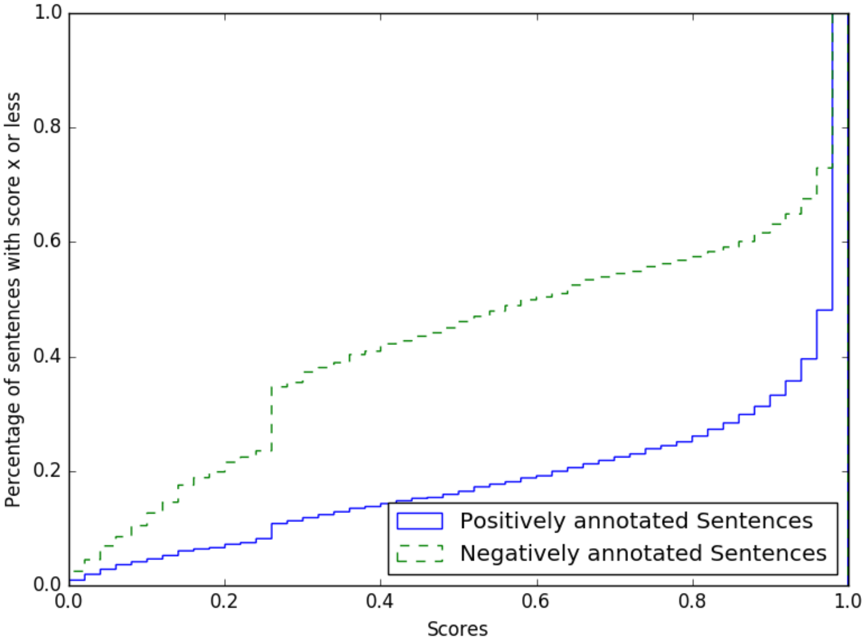
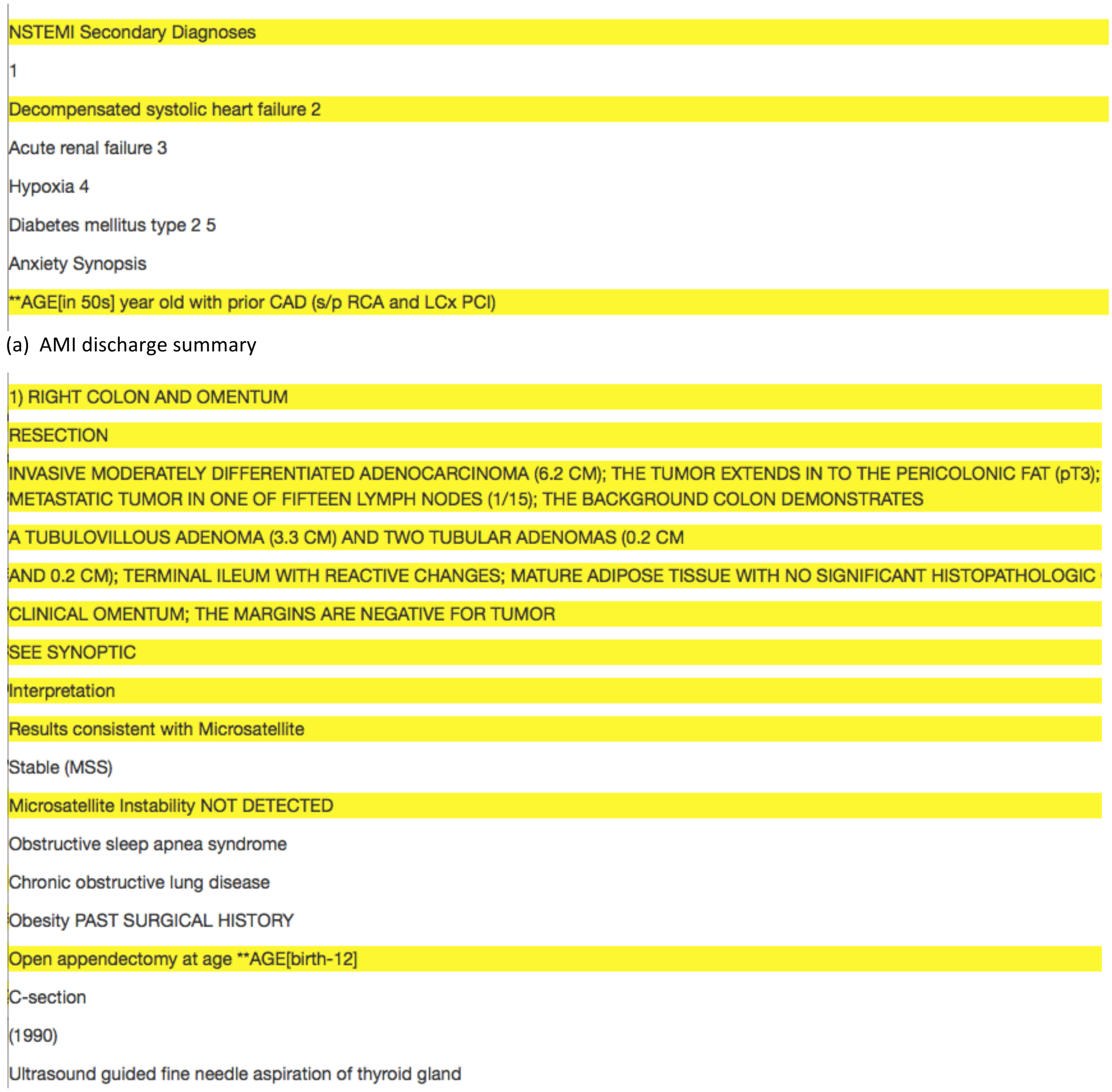


Figure 5.7 CDF distribution of cancer sentences scores generated by LSTM with attention model for positively and negatively annotated sentences in the cancer gold standard

The Pearson correlation between sentence selection and Word2Vec scores were all positive and the correlations for ischemic heart models scores was the highest. The Pearson correlation between the two models were 0.556, 0.447, and 0.282 for ischemic heart, AMI, and cancer models respectively.

One of the model implementation is highlighting the relevant sentences. Using the CDF plots, we selected the score threshold to highlight relevant sentences. For each model, we identified the highlight threshold where 60% of positively annotated sentences has that corresponding score or higher. We highlighted the AMI sentences that have LSTM with attention scores higher than 0.2 for AMI sentences and cancer sentences with LSTM scores higher than 0.7. The sentences that include AMI findings and diagnosis were highlighted in AMI discharge summary (Figure 5.9a). Sentences that describe the cancer location and stage were also highlighted (Figure 5.9b).

The model implementation can be expanded to focus the attention of the reader on the individual words inside the sentences. The attention layer creates an attention distribution for the words in the sentence or an attention score for each word. The scores of the words can be visualized to show the importance of the words and their relative importance to the phenotype, as shown in Figure 5.10a and Figure 5.10b.



NSTEMI Secondary Diagnoses

1

Decompensated systolic heart failure 2

Acute renal failure 3

Hypoxia 4

Diabetes mellitus type 2 5

Anxiety Synopsis

**AGE[in 50s] year old with prior CAD (s/p RCA and LCx PCI)

(a) AMI discharge summary

1) RIGHT COLON AND OMENTUM

RESECTION

INVASIVE MODERATELY DIFFERENTIATED ADENOCARCINOMA (6.2 CM); THE TUMOR EXTENDS IN TO THE PERICOLONIC FAT (pT3); METASTATIC TUMOR IN ONE OF FIFTEEN LYMPH NODES (1/15); THE BACKGROUND COLON DEMONSTRATES

A TUBULOVILLOUS ADENOMA (3.3 CM) AND TWO TUBULAR ADENOMAS (0.2 CM

AND 0.2 CM); TERMINAL ILEUM WITH REACTIVE CHANGES; MATURE ADIPOSE TISSUE WITH NO SIGNIFICANT HISTOPATHOLOGIC

CLINICAL OMENTUM; THE MARGINS ARE NEGATIVE FOR TUMOR

SEE SYNOPTIC

Interpretation

Results consistent with Microsatellite

Stable (MSS)

Microsatellite Instability NOT DETECTED

Obstructive sleep apnea syndrome

Chronic obstructive lung disease

Obesity PAST SURGICAL HISTORY

Open appendectomy at age **AGE[birth-12]

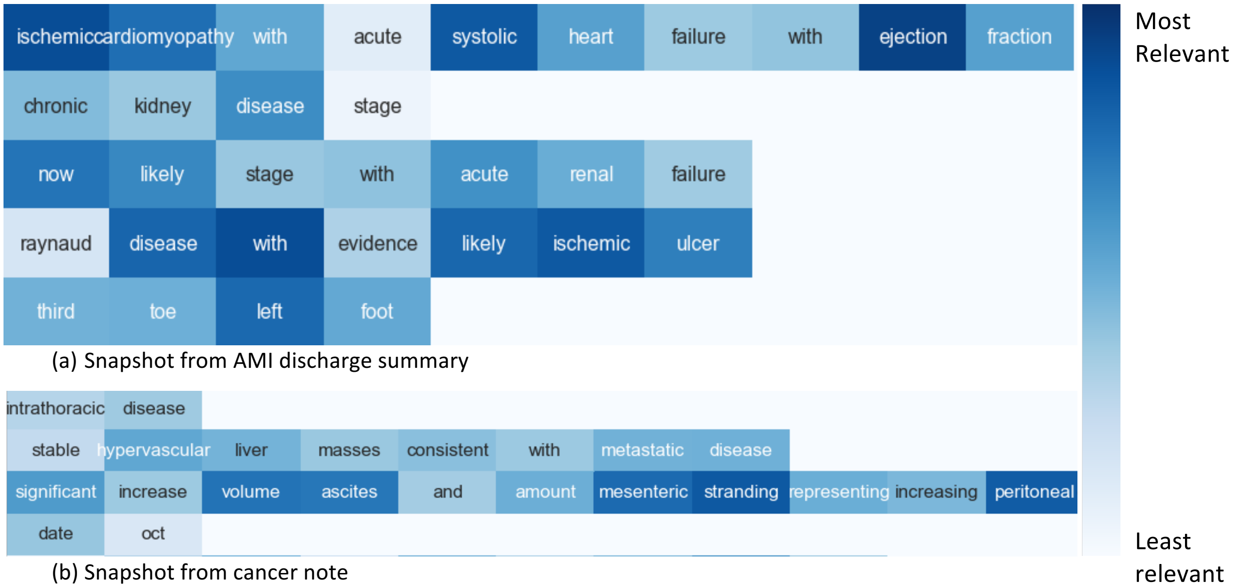
C-section

(1990)

Ultrasound guided fine needle aspiration of thyroid gland

(b) Cancer note

Figure 5.8 A visual example of the relevant sentences selected by LSTM model and highlighted with a yellow background: (a) AMI discharge summary (b) cancer note

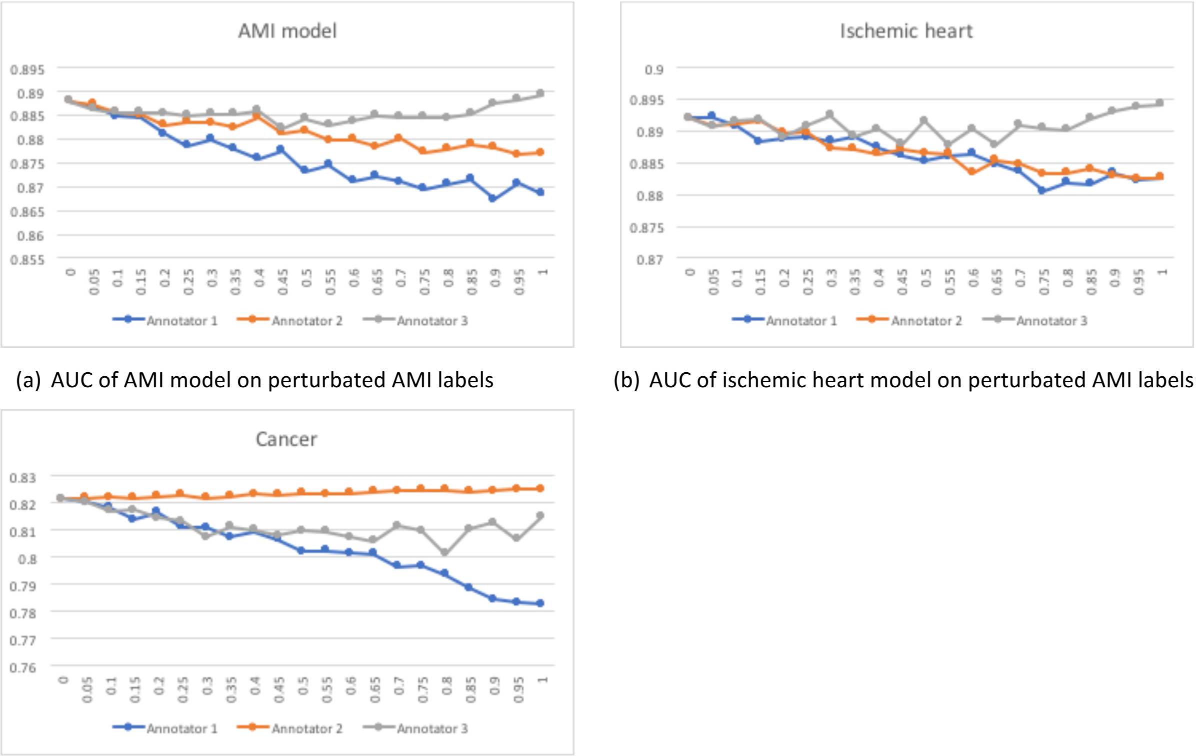


(a) Snapshot from AMI discharge summary

(b) Snapshot from cancer note

Figure 5.9 A heat map for the attention scores for the words in sentences from: (a) AMI discharge summary (b) cancer notes. Darker color indicates higher attention scores which indicate higher relevance to the phenotype.

The model AUC should decrease as the perturbation of good annotators labels increase. Annotator 1 (blue line in Figure 5.11) had the highest agreement with the model and perturbing his/her annotation lead to 2% decrease in AUC for AMI and ischemic heart models, and 4% for cancer model with full perturbation, as seen in Figures below. The AUC value of the AMI and ischemic heart had a similar declining trend for Annotator 2 (orange line) but with a lower slope; however, the perturbation of Annotator 2 labels did not affect on the model performance given the labels of the other annotators, as seen in Figure 5.11. Excluding Annotator 3 labels from the gold standard (plotted in grey) influenced the model AUC values differently. There was an apparent increase in AUC when most of Annotator 3 positively-annotated sentences were switched to negatively-annotated sentences.



(a) AUC of AMI model on perturbated AMI labels

(b) AUC of ischemic heart model on perturbated AMI labels

(c) AUC of Cancer model on perturbated cancer labels

Figure 5.10 Perturbation of positive labels in the gold standard provided by annotators and their effect on the reported AUC of the model: (a) Reevaluating AMI model (b) Reevaluating ischemic heart model (c) Reevaluating cancer model

Discussion

The results show that LSTM sentence extraction models, depicted in Figures 5.1 and 5.2), can locate and extract relevant sentences to a phenotype in clinical documents. Training an LSTM model can improve the extraction of relevant sentences compared to using keywords search only. Both LSTM and Word2Vec models complement each other and combining their output increased the extraction accuracy.

The models proposed in this chapter outperform the conventional methods such as concept extraction tools. For instance, KnowledgeMap could only identify a subset of relevant AMI and cancer sentences. The AUC of identifying relevant sentences using KnowledgeMap were 0.5273 and 0.6060 for AMI and cancer, respectively, which were lower than the AUC of Word2Vec and LSTM with attention models.

Word2Vec yielded higher relevance scores for some sentences with words similar to the phenotype's keywords. LSTM generated higher scores for sentences that mentioned the phenotype contextually and had relevant concepts such as diagnosis, symptoms, and medications. For instance, in AMI gold standard, the Word2Vec scores of the sentences “ischemic cardiomyopathy” and “2d echocardiogram read by dr” were higher than the LSTM score. This shows that Word2Vec can extract relevant sentences that contain similar words better than LSTM.

The LSTM model assigned phenotype-relevance scores higher than Word2Vec for the AMI sentences “major epicardial coronary arteries are widely patented,” “placed on beta-blocker therapy,” and “there is disease in the very small distal branches of the circumflex.” In the annotated cancer notes, LSTM assigned scores higher than Word2Vec for the sentences: “adenocarcinoma positive for kras g12c treatment,” and “left lower lobe wedge resection.” Word2Vec provided higher scores for the sentences “but proved to be chondrosarcomatous at exploration,” and “retinoblastoma status post radiation and bilateral enucleation treatment.”. Combining LSTM with Word2Vec scores increased the identifications of relevant sentences.

Sentence extraction models can highlight the relevant sentences for providers, especially the ones that do not include direct information or the phenotype search keywords. The model could improve the quality of retrieved information presented to healthcare providers.

The low kappa scores and the slightly lower than expected AUC for the sentence extraction models can be explained by the mislabeled sentences in the gold standard. The annotators selected sentences that had low scores and did not include phenotype data in some cases. None of the clinical concepts in the sentences “obstructive sleep apnea”, “date social history and family history” and “no subtype nasal” are related to AMI or cancer. On the other hand, some sentences that the extraction model missed are relevant to other phenotypes. For example, the annotated sentences “Lisinopril” and “she was given Lovenox therapeutic dosing and transferred for further management” and “I personally reviewed her face and neck ct,” and “ct sinuses were” are relevant to AMI and cancer and other phenotypes such as deep vein thrombosis, high blood pressure, heart failure.

Sentence extraction models identified related sentences that some annotators overlooked. For instance, all models assigned scores higher than 0.7 for the sentences “estimated gfr 54” and “she underwent chest radiography” in AMI documents, and “carcinoid sxs are stable” and “enhancing left adrenal lesion likely representing metastatic disease” in cancer notes. Some annotators did not label those sentences as relevant. One application for the model is pre-annotating possible relevant sentences for annotators, as shown in Figure 5.9. Pre-annotating can increase the agreement scores amongst annotators who have different levels of expertise [303,341]. It may also reduce the annotation time and effort [303,341].

The discrimination between the scores of positively-annotated and negatively-annotated AMI sentences was higher than the cancer sentences discrimination as inferred from CDF plots (Figures 5.6 and 5.8). Using general phenotype definition could affect the performance of sentence extraction model. In the training, we applied the most general cancer definition while we chose a more granular, specific definition to train the AMI model. We also observed lower performance when we used ischemic heart definition to identify AMI relevant sentences, which is a less specific than AMI definition (Table 2 and 3).

The reliability of cancer annotation was not high which might affect the reported AUC of cancer sentence extraction model. Annotating cancer notes is challenging, especially for non-experienced annotators who received minimal or no clinical training in oncology. The level of clinical expertise required to annotate cancer notes may explain the low performance of the cancer model. The kappa and F1 between “Annotator 1 and Annotator 2” and between “Annotator 2 and Annotator 3” scores were lower than 0.2 in the cancer dataset (Table 2). Including only sentences with a full agreement in the strict gold standard increased the AUC of identifying relevant sentences. The level of disagreement might have affected the reported performance of the model. For instance, a first-year medical student does not have enough training in oncology as a fourth-year medical student, which underlied kappa and F1 scores where a fourth-year medical student (Annotator 1) had the highest values.

We believe that the sentence extraction model can have multiple implementations in the clinical environment. The model can serve as a tool to gather the facts from unstructured data about a phenotype from the past visits and present them to health care providers, as visualized in Figure 5.9. On the words level, the scores of the words can focus the attention on the words relative to the phenotype. In AMI sentences, heart, ejection, cardiomyopathy, and ischemic are more relative to the AMI condition than with, stage, or foot. Highlighting or locating words, as shown in Figure 5.10. The words score can denoise the sentences from irrelevant word to perform a data-driven preprocessing step.

This study has some limitations. First, we evaluated the models on a small annotated dataset. It will be helpful to compare models using larger datasets that include broader disease definitions and longer term patient data. Second, the quality of the annotation was low to medium. The recruited annotators have different levels of clinical expertise. Recruiting more annotators with the required experience would provide a more accurate evaluation. Third, our cancer note extraction methods might be biased since we used NLP rules. In our future work, we are planning to create another note extraction method or less biased extraction method for cancer note. Fourth, we trained a binary model for each phenotype. Training a multi-phenotype model that accounts for the co-occurrence of phenotypes which could be a possible extension to our model and would have a valuable application in the clinical environment.

Summary

This chapter introduced a model to learn a representation for sentences to extract the ones relevant to a phenotype such as diagnoses and medications from clinical documents. The study demonstrates a sentence extraction model can learn representations and identify relevant sentences in clinical documents for acute myocardial infarction and cancer. We were able to combine methods to identify relevant sentences with greater accuracy than Word2Vec or LSTM extraction models alone. Learning informative sentences representations depends upon specificity in phenotype definitions, while successful sentence extraction is dependent on annotator experience. This representation learning and extraction approach could help providers extract relevant patient information in narrative clinical documents, thus increasing efficiency and improving patient care.

CHAPTER 6 CONCLUSION

EMRs store heterogeneous data that describe patients' health over their lifespan. Researchers leverage the power of machine learning to analyze clinical data and build models that help decision-making. Researchers convert clinical data into machine-readable features. Preparing informative features that can highlight the important patterns in input samples is a primary success factor in training a model.

Creating a useful feature representation depends on the predictors and output. Longitudinal dynamic features and time-dependent outputs should be represented differently than static features and fixed outputs. Text-based? features that highlight the words' meaning and context allow the machine learning model to mimic human processing of the text, which in turn could improve the quality of models.

This dissertation focused on learning and creating feature representations of structured and unstructured clinical data. The central insight is learning and forming features that embed the distinct input patterns that are descriptive to different outputs. In longitudinal dynamic data where patients' health status changes over time, constructing and feeding the updated information can improve the prediction model and mimic the way healthcare providers reevaluate the upcoming patient's outcome or disease. Unstructured data (i.e. clinical documents) is another valuable source in EMR. Learning and creating features and representations for the clinical documents is a critical step in training machine learning models that have various applications such as phenotyping, predicting outcome, identifying the type of questions and need in patients messages, or retrieving information. Our research and investigations contribute to the biomedical informatics field by addressing common challenges encountered by researchers who are training machine learning models using EMR data.

One challenge in using EMR to predict patient outcomes is capturing the change in the data that can influence the outcome. For instance, having complications after discharge can increase the probability of an unavoidable emergency visit. Chapter 3 described a dynamic model that predicts the risk of readmission or death after being hospitalized for a phenotype. The proposed model creates temporal representations for structured data and dynamic features for information inserted after discharge to predict the outcome of patients. The post-discharge model mimics the healthcare provider assessment by incorporating the changes in patient health status after being discharged. The post-discharge model outperformed models trained on traditional static features such as LACE and at-discharge models in predicting high risk patients who had a hip fracture or cognitive heart failure by at least 20% in the AUC value. This work fills the gap in the current readmission and outcome prediction models that create a static representation and leave out the new information inserted before the event of interest.

Another medical informatics challenge is identifying effective clinical text mining algorithms that integrate semantics and word context into representations. Creating those features might need manual efforts which can be hard to obtain and might limit the models' generalizability. Chapter 4 describes a method to learn text representations that integrates both the semantics and context of words in patients' portal messages. Learning an informative representation of patients' messages can improve the detection of patients' needs, and thus prioritizing the response to the messages according to their urgency. The work in chapter 4 examined and compared different messages representations. Some of representations lack context and semantics, while other incorporate semantics or context or both. The analysis shows that learning a representation that includes semantics and context lead to a better identification of patients' needs in messages. The AUC values were higher by 3%-10% using semantic and context representations compared to traditional features such as bag of words. This work built on and improved the studies that Cronin and colleagues performed in the same area.

Extracting relevant clinical text and identifying disease information have significant applications in the clinical environment. Information extraction can facilitate chart reviews by presenting the relevant parts of clinical documents to healthcare providers. However, training extraction models often require manually curated datasets. Annotating datasets need human resources. The necessity of annotated datasets and the cost of preparing them can hinder the development of extraction models. Chapter 5 introduced a machine learning model that extracts sentences relevant to a phenotype without using an annotated dataset. The extraction model learns semantic representations for sentences using the correlations between the content and the phenotypes of the clinical documents. Sentence extraction models performed better than KnowledgeMap, a medical concepts extraction tool, and keywords search. The proposed deep learning extraction method might reduce the need for annotated datasets during the training. Moreover, it offers biomedical informatics researchers a chance to utilize the reachable clinical datasets and train extraction models that fits the writing and linguistic styles in their organizations. Other researchers might prefer the usage of medical terminology but combining both source can improve the NLP models and reduce the need for de-identifying, annotating, and sharing datasets that have small size in most cases.

This dissertation developed methods to mine data from EMR and converted them into numerical representations. The EMR data is heterogeneous, dynamic, and changeable. Every encounter in healthcare organization leaves a digital footprint. However, the tremendous amount of clinical data can be hard to analyze and comprehend manually in a busy environment such as healthcare. Moreover, some analyses require ongoing complex computations and finding associations that humans are incapable of performing promptly. Machine learning models can perform complex computations, learn associations, and apply the learned patterns on unseen data. The raw EMR data needs to be transformed and converted into digital features and variables that the machine learning models understand to find the patterns and solve the clinical problems.

The Dynamic nature in data should be captured to predict a time-varying outcome such as readmission. Our analysis proved that using dynamic features to train prediction models improves the prediction. Creating the dynamic features might need some understanding of the sources of time-varying features. The clinical experts can help during the creation of the dynamic features by identfying the sources of dynamic data that can improve the models. Besides, they can identify the CDS tools that can be built or utilized to collect the dynamic variables and new clinical events.

Integrating the semantics and the context in the text representation can increase the accuracy of the output of the machine learning models such as classifications and information extraction. However, the current NLP methods require manually annotated datasets and curated language and medical dictionaries to train the models. Applying the deep learning methods in the NLP field shows promising results and has many applications in healthcare such as information extraction, summarizations, and answering questions. The proposed NLP methods can reduce the necessity of manually embedding the scientific and medical knowledge to create and represent the clinical text. The presented NLP methods are not a replacement for the human, and it builds on the existing clinical knowledge stored EMR data to reduce the time and effort wasted on tedious tasks such as annotating datasets for training models or triaging patients' messages according to the requests and needs of the patients. The dissertation proposed methods to address those challenges: creating dynamic representations for changeable EMR data, and learning text representation while embedding semantics and context.

CHAPTER 7 FUTURE WORK

This dissertation proposed representations of structured and unstructured that improved machine learning model accuracy. However, there is always room for improvement. This chapter discusses extensions or future directions to the work proposed in this dissertation, and, more broadly, other possible future research in medical informatics. First, the chapter describes the possible improvements on creating dynamic representations for patients after a baseline or major event (e.g. discharge). The Chapter explore the future of applying deep learning in clinical NLP including classifications and information extraction.

Dynamic features

In the dynamic feature model, the post-discharge model predicted the outcome of patients using the structured post-discharge data. Learning a dynamic representation of the unstructured post-discharge data, such as clinical communication, and combining it with dynamic structured features might improve the identification of adverse events that might lead to unavoidable readmissions. In our future work, we are planning to apply the methods and techniques used here to create a text representation for the clinical documents and improve the dynamic model proposed in the first aim (please refer to page 10 for a general description, and page 30 for the specific objectives of the first aim). The future model will create a text representation for the unstructured data and clinical documents that are inserted into the patients EMR record whether clinical notes during outpatient appointments or patients' portal messages exchanged between healthcare providers and patients. Our final goal is developing a method that creates a representation for the dynamic structured and unstructured data and trains our machine learning models on the combined or learned representation for the two types of data. The future model can utilize the sources of new information after a significant event (e.g. discharge, transient ischemic attack) such as patient portals and outpatient data, creating informative representations for EMR data including structured (laboratories, medications, etc.), clinical documents, images etc., and combining those representations to build decision aid tools such as high risk patients identification and new phenotypes.

Deep learning in clinical NLP

Integrating semantics and context of words enhanced the text representations for classification. However, the learned representations can be improved. Adding more messages can increase the model's generalizability and accuracy. A semi-supervised model can be trained to learn representation on larger unannotated datasets since annotating all messages requires human and time resources [31]. Learning representation for other clinical tasks can evaluate the model scalability. A similar representation learning model can be applied to identify the patients' satisfaction and complaints in messages, detect adverse events in patients' portal messages and extract the billing codes from the messages if they exist. For instance, patients communicate their concerns about their health via messages in patients' portals. They seek confirmations, suggestions, advices, and opinions about their current health status. Patients who had a major surgery and cannot leave their houses may send their surgeon or primary care physician discomforts, complications, or unusual physical symptoms. Learning and creating representations for those messages can offer a patient-reported source of data that the EMR has but still not fully utilized.

Learning sentences' representation while considering semantics and context improved the extraction of sentences related to a phenotype. However, the extraction model learned representation to extract relevant sentences for one phenotype. Hundreds or thousands of models are needed to cover all phenotypes. Learning a representation for sentences or words for multiple ICD codes can be more efficient, and it will account for the co-occurrence of multiple ICD codes. A multi-ICD attention model can identify relevant words for ICD codes using attention layer only, as proposed in [332]. The proposed model has a short training time and learns scores for words in a long text regardless of their location [332]. This strategy would allow our model to link the ICD codes to the relevant words inside the documents. Identifying the relevant words-to-ICD can: 1) help biomedical informatics researchers find new associations between words and diseases, 2) learn the words that can be associated with multiple diseases, 3) identify words for disease that were not included in UMLS, 4) remove the noisy words for clinical documents as a pre-processing step, 5) reduce the time that clinical experts spend on creating and annotating relevant words and benefit from their knowledge on refining the output.

A more specific multi-ICD model can be trained to predict a subset of ICD for a general phenotype. For example, a model that predicts cancer billing codes for oncology notes can list the possible set of billing codes for the clinical provider at the end of the documentation. Training a model for cancer codes might increase the discrimination of words' attention scores distribution that is relevant to specific cancer types such as lung cancer. Clinical researchers hesitate to use the billing codes in their research since only a few ICD assigned to each document based on the billing process, and the codes do not reflect all the ICD codes in the documents. Training specific multi-ICD model can help researchers expand the list of ICD codes they implement in their research.

Moreover, it can explain some the variability in the coding for similar patients. Physicians and billing coders assign ICD codes depending on the bill that will be sent to a payer. Hence, similar patients might have different ICD codes in their record. Showing all ICD codes can explain this coding variability by listing all ICD. Showing all the ICD will list the missing ones and the different diseases and comorbidities that can lead to the difference in assigning the billing codes.

Relevant sentence extraction models can have multiple implementations. The model can preprocess and remove irrelevant noisy sentences or words before training a machine learning model. For instance, the cancer sentence extraction model can identify and clean the general cancer notes before training a classifier to detect specific cancer billing codes. Another implementation is creating a decision support tool that summarizes patients documents from past visits. The tool extracts and displays the sentences about the phenotype that the healthcare provider wants to treat. This would enable researchers to mine the clinical documents efficiently, identify the clinical contents inside the documents, and increase the usage of clinical documents by learning efficient and more concise representation for the documents.

In conclusion, our future work would focus on addressing limitations to our current study: expanding the creation of dynamic features and training dynamic models and improving the clinical document representation leading to better clinical text mining. It would also build upon the foundation of work created here. Future work including building dynamic prediction models for any clinical events (e.g., the probability of being diagnosed with a phenotype, disease prognosis) and applying advanced deep learning models in clinical text mining. The future could elevate this dissertation by enabling researchers to train machine learning models that have new information feed loop that feed the new patients EMR data and mine the clinical documents that we are only able to obtain the tip of the iceberg of the clinical information stored in them.

APPENDIX A

This appendix describes the LSTM model and equations of LSTM and the attention layer that we used in our models.

LSTM description

Equations set 1 shows LSTM equation as described in Zaremba and Sutskever [324]. The output gate controls the amount of current state that will be passed to next LTSM cell. Finally, as the last equation in the equations set shows, the h_t is a function of the memory cell and the output cell, and it represents the learned representation from the current and all the previous LSTM sequence elements.

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)} \quad (1)$$

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}$$

$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}$$

$$u_t = tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1}$$

$$h_t = o_t \odot \tanh{(c_t)}$$

Where W is weight matrix for the current cell, U is weight matrix for the hidden representation of previous LSTM cell, b superscripts i, f and o correspond to the gates: input, forget and output. Figure A.1 depicts the LSTM cell and its gates.

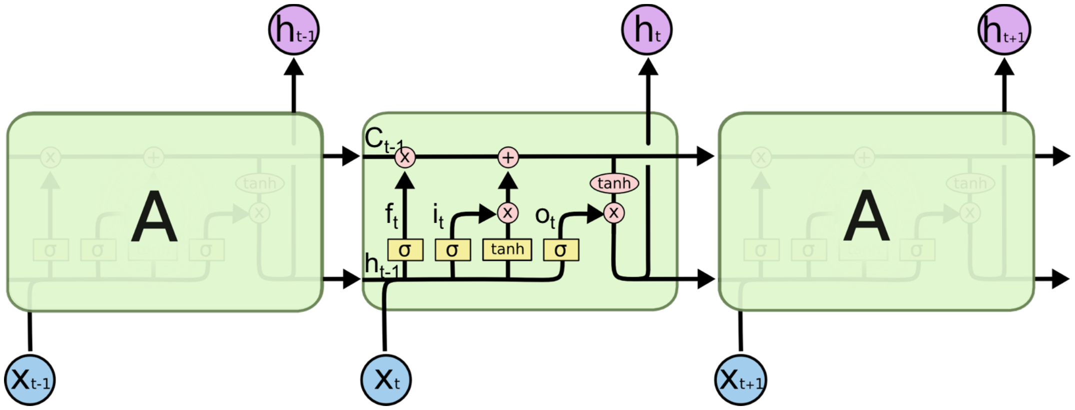


Figure 0.1 A detailed visualization of gates in LSTM unit. f is the forget gate, i is the input gate, o is the output gate, h is the hidden representation of current unit, and c is the memory cell

Attention layer equations

Given the hidden states in LSTM layer $h_1, h_2, \ldots, h_{t-1}$ and context vector v_t, a simple concatenation layer that combines previous hidden states and context vector to calculate the attention at hidden state h_t. An alignment vector $a_t(s)$ is created by combining each source hidden state (hidden states that preceded h_t) $\bar{h}_s$ with the current target hidden unit at step s, h_t using the following equation:

$$a_t(s) = \frac{\exp\left(score\left(h_t, \bar{h}_s\right)\right)}{\sum_{i=1}^{\acute{s}} \exp\left(score\left(h_t, \bar{h}_s\right)\right)}$$

Where score values are calculated using the following equation:

$$score = \tanh(\bar{h}_s W_a h_t + b_a)$$

W_a is the trainable weight matrix of attention. In our current analysis, we used the following equation to simplify the model:

$$score = \tanh(W_a h_t + b_a)$$

A context vector is created by calculating the weighted vector of input sequence using the alignment vector a_t, as shown in chapter 5, Figure 5.5.

REFERENCES

1 Botsis T, Hartvigsen G, Chen F, *et al.* Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci* 2010;**2010**:1–5.http://www.ncbi.nlm.nih.gov/pubmed/21347133 (accessed 28 Nov 2017).

2 Safran C, Bloomrosen M, Hammond WE, *et al.* Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc* 2007;**14**:1–9. doi:10.1197/jamia.M2273

3 Friedman DJ, Parrish RG, Ross DA. Electronic health records and US public health: current realities and future promise. *Am J Public Health* 2013;**103**:1560–7. doi:10.2105/AJPH.2013.301220

4 Demner-Fushman D, Mork JG, Shooshan SE, *et al.* UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. *J Biomed Inform* 2010;**43**:587–94. doi:10.1016/j.jbi.2010.02.005

5 Shortliffe EH. Computer programs to support clinical decision making. *JAMA* 1987;**258**:61–6.http://www.ncbi.nlm.nih.gov/pubmed/3586293 (accessed 14 Nov 2017).

6 Hunt DL, Haynes RB, Hanna SE, *et al.* Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *JAMA* 1998;**280**:1339–46.http://www.ncbi.nlm.nih.gov/pubmed/9794315 (accessed 14 Nov 2017).

7 Miller RA, Masarie jr FE. Use of the quick medical reference (QMR) program as a tool for medical education. *Methods Inf Med* 1989;**28**:340–5.

8 Pulley JM, Denny JC, Peterson JF, *et al.* Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project. *Clin Pharmacol Ther* 2012;**92**:87–95. doi:10.1038/clpt.2011.371

9 Olesen JB, Lip GYH, Hansen ML, *et al.* Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. *BMJ* 2011;**342**:d124. doi:10.1136/BMJ.D124

10 Schwartz PJ, La Rovere MT, Vanoli E. Autonomic nervous system and sudden cardiac death. Experimental basis and clinical observations for post-myocardial infarction risk stratification. *Circulation* 1992;**85**:I77-91.http://www.ncbi.nlm.nih.gov/pubmed/1728509 (accessed 28 Nov 2017).

11 Fonarow GC, Adams KF, Abraham WT, *et al.* Risk Stratification for In-Hospital Mortality in Acutely Decompensated Heart Failure&lt;SUBTITLE&gt;Classification and Regression Tree Analysis&lt;/SUBTITLE&gt; *JAMA* 2005;**293**:572. doi:10.1001/jama.293.5.572

12 Boriani G, Botto GL, Padeletti L, *et al.* Improving stroke risk stratification using the CHADS2 and CHA2DS2-VASc risk scores in patients with paroxysmal atrial fibrillation by continuous arrhythmia burden monitoring. *Stroke* 2011;**42**:1768–70. doi:10.1161/STROKEAHA.110.609297

13 Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. *J Biomed Inform* 2015;**56**:229–38. doi:10.1016/j.jbi.2015.05.016

14 Churpek MM, Yuen TC, Winslow C, *et al.* Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards.

Crit Care Med 2016;**44**:368–74. doi:10.1097/CCM.0000000000001571

15 Barnes S, Hamrock E, Toerper M, *et al.* Real-time prediction of inpatient length of stay for discharge prioritization. *J Am Med Informatics Assoc* 2016;**23**:e2–10. doi:10.1093/jamia/ocv106

16 Thoeny HC, Ross BD. Predicting and monitoring cancer treatment response with DW-MRI. *J Magn Reson imaging* 2010;**32**:2–16. doi:10.1002/jmri.22167.Predicting

17 Schroth W, Antoniadou L, Fritz P, *et al.* Breast cancer treatment outcome with adjuvant tamoxifen relative to patient CYP2D6 and CYP2C19 genotypes. *J Clin Oncol* 2007;**25**:5187–93. doi:10.1200/JCO.2007.12.2705

18 Paul D, Su R, Romain M, *et al.* Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Comput Med Imaging Graph* 2017;**60**:42–9. doi:10.1016/j.compmedimag.2016.12.002

19 Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. *J Biomed Inform* 2015;**57**:28–37. doi:10.1016/j.jbi.2015.07.010

20 McCoy TH, Castro VM, Cagan A, *et al.* Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study. *PLoS One* 2015;**10**:e0136341. doi:10.1371/journal.pone.0136341

21 Jagannatha AN, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. *Proc Conf Empir Methods Nat Lang Process Conf Empir Methods Nat Lang Process* 2016;**2016**:856–65.http://www.ncbi.nlm.nih.gov/pubmed/28004040 (accessed 25 Nov 2017).

22 Collins SA, Cato K, Albers D, *et al.* Relationship between nursing documentation and patients' mortality. *Am J Crit Care* 2013;**22**:306–13. doi:10.4037/ajcc2013426

23 Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514–8. doi:10.1136/jamia.2010.003947

24 Uzuner Ö, South BR, Shen S, *et al.* 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552–6. doi:10.1136/amiajnl-2011-000203

25 Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Informatics Assoc* 2013;**20**:806–13. doi:10.1136/amiajnl-2013-001628

26 Chen SY, Stem M, Schweitzer MA, *et al.* Assessment of postdischarge complications after bariatric surgery: A National Surgical Quality Improvement Program analysis. *Surgery* 2015;**158**:777–86. doi:10.1016/j.surg.2015.04.028

27 Miller TA, Bethard S, Dligach D, *et al.* Extracting Time Expressions from Clinical Text. 2015;:81–91.https://aclanthology.info/pdf/W/W15/W15-3809.pdf (accessed 10 Nov 2017).

28 Sahu SK, Anand A, Oruganty K, *et al.* Relation extraction from clinical texts using domain invariant convolutional neural network. Published Online First: 30 June 2016.http://arxiv.org/abs/1606.09370 (accessed 10 Nov 2017).

29 Cronin RM. *Automatic Classification of Patient-Generated Messages From a Patient Portal*. 2015.

30 Garla VN, Brandt C. Ontology-guided feature engineering for clinical text classification. *J Biomed Inform* 2012;**45**:992–8. doi:10.1016/j.jbi.2012.04.010

31 Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management. *J Biomed Inform* 2013;**46**:869–75. doi:10.1016/j.jbi.2013.06.014

32 Denecke K, Deng Y. Sentiment analysis in medical settings: New opportunities and challenges. *Artif Intell Med* 2015;**64**:17–27. doi:10.1016/j.artmed.2015.03.006

33 South BR, Shen S, Jones M, *et al.* Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *Summit on Translat Bioinforma* 2009;**2009**:1–32.http://www.ncbi.nlm.nih.gov/pubmed/21347157 (accessed 12 Nov 2017).

34 Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA . Annu Symp proceedings AMIA Symp* 2011;**2011**:189–96.http://www.ncbi.nlm.nih.gov/pubmed/22195070 (accessed 12 Nov 2017).

35 Denny JC. Chapter 13: Mining Electronic Health Records in the Genomics Era. *PLoS Comput Biol* 2012;**8**:e1002823. doi:10.1371/journal.pcbi.1002823

36 Velupillai S, Mowery D, South BR, *et al.* Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearb. Med. Inform. 2015. doi:10.15265/IY-2015-009

37 Chapman WW, Nadkarni PM, Hirschman L, *et al.* Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011;**18**:540–3. doi:10.1136/amiajnl-2011-000465

38 Ancker JS, Shih S, Singh MP, *et al.* Root Causes Underlying Challenges to Secondary Use of Data. *AMIA Annu Symp Proc* 2011;**2011**:57–62. doi:10.1126/science.277.5322.9o

39 O'Malley KJ, Cook KF, Price MD, *et al.* Measuring diagnoses: ICD code accuracy. Health Serv. Res. 2005. doi:10.1111/j.1475-6773.2005.00444.x

40 Nguyen H, Patrick J. Text Mining in Clinical Domain. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. 2016. doi:10.1145/2939672.2939720

41 Feldman K, Faust L, Wu X, *et al.* Beyond Volume: The Impact of Complex Healthcare Data on the Machine Learning Pipeline. Published Online First: 1 June 2017.http://arxiv.org/abs/1706.01513 (accessed 1 Dec 2017).

42 Lasko TA, Denny JC, Levy MA. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS One* 2013;**8**:e66341. doi:10.1371/journal.pone.0066341

43 L'Heureux A, Grolinger K, Elyamany HF, *et al.* Machine Learning With Big Data: Challenges and Approaches. *IEEE Access* 2017;**5**:7776–97. doi:10.1109/ACCESS.2017.2696365

44 García S, Luengo J, Herrera F. Dealing with Noisy Data. Springer, Cham 2015. 107–45. doi:10.1007/978-3-319-10247-4_5

45 Nguyen H, Patrick J. Text Mining in Clinical Domain: Dealing with Noise. doi:10.1145/2939672.2939720

46 Lee C, Luo Z, Ngiam KY, *et al.* Big Healthcare Data Analytics: Challenges and Applications. Springer, Cham 2017. 11–41. doi:10.1007/978-3-319-58280-1_2

47 Zhu X, Wu X. Class Noise vs. Attribute Noise: A Quantitative Study. *Artif Intell Rev* 2004;**22**:177–210. doi:10.1007/s10462-004-0751-8

48 Brodley CE, Friedl MA. Identifying Mislabeled Training Data. *J Artif Intell Res* Published

Online First: 1 June 2011. doi:10.1613/jair.606

49 Miotto R, Li L, Kidd BA, *et al.* Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Published Online First: 2016. doi:10.1038/srep26094

50 Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. *Nat Biotechnol* 2017;**35**:342–6. doi:10.1038/nbt.3780

51 Wells BJ, Chagin KM, Nowacki AS, *et al.* Strategies for handling missing data in electronic health record derived data. *EGEMS (Washington, DC)* 2013;**1**:1035. doi:10.13063/2327-9214.1035

52 Raghunathan TE. What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. *Annu Rev Public Health* 2004;**25**:99–117. doi:10.1146/annurev.publhealth.25.102802.124410

53 Little RJ, D'Agostino R, Cohen ML, *et al.* The Prevention and Treatment of Missing Data in Clinical Trials. *N Engl J Med* 2012;**367**:1355–60. doi:10.1056/NEJMsr1203730

54 Bounthavong M, Watanabe JH, Sullivan KM. Approach to Addressing Missing Data for Electronic Medical Records and Pharmacy Claims Data Research. *Pharmacother J Hum Pharmacol Drug Ther* 2015;**35**:380–7. doi:10.1002/phar.1569

55 Walsh CG, Ribeiro JD, Franklin JC. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clin Psychol Sci* 2017;**5**:457–69. doi:10.1177/2167702617691560

56 Polnaszek B, Gilmore-Bykovskyi A, Hovanes M, *et al.* Overcoming the challenges of unstructured data in multisite, electronic medical record-based abstraction. *Med Care* Published Online First: 2016. doi:10.1097/MLR.0000000000000108

57 Stavrianou A, Andritsos P, Nicoloyannis N. Overview and semantic issues of text mining. *ACM SIGMOD Rec* 2007;**36**:23. doi:10.1145/1324185.1324190

58 Berry MW, Castellanos M. *Survey of text mining II: Clustering, classification, and retrieval*. 2008. doi:10.1007/978-1-84800-046-9

59 Hu J, Fang L, Cao Y, *et al.* Enhancing text clustering by leveraging Wikipedia semantics. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*. New York, New York, USA: : ACM Press 2008. 179. doi:10.1145/1390334.1390367

60 Singh M, Murthy A, Singh S. Prioritization of Free-Text Clinical Documents: A Novel Use of a Bayesian Classifier. *JMIR Med Informatics* 2015;**3**:e17. doi:10.2196/medinform.3793

61 Choi E, Bahadori MT, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *arXiv* 2015;**56**:1–12.http://arxiv.org/abs/1511.05942

62 Goldstein BA, Navar AM, Pencina MJ, *et al.* Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Informatics Assoc* 2017;**24**:198–208. doi:10.1093/jamia/ocw042

63 Zhao J, Papapetrou P, Asker L, *et al.* Learning from heterogeneous temporal data in electronic health records. *J Biomed Inform* 2017;**65**:105–19. doi:10.1016/j.jbi.2016.11.006

64 Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. Published Online First: 24 June 2012.http://arxiv.org/abs/1206.5538 (accessed 30 Nov 2017).

65 Anderson M, Antenucci D, Bittorf V. Brainwash: A Data System for Feature Engineering.

EecsUmichEdu Published Online First: 2013. doi:10.1.1.244.9089

66 Sun J, McNaughton CD, Zhang P, *et al.* Predicting changes in hypertension control using electronic health records from a chronic disease management program. *J Am Med Inform Assoc* Published Online First: 2014. doi:10.1136/amiajnl-2013-002033

67 Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. *Yearb Med Inform* 2014;**9**:215–23. doi:10.15265/IY-2014-0009

68 Ramani G V, Uber PA, Mehra MR. Chronic heart failure: contemporary diagnosis and management. *Mayo Clin Proc* 2010;**85**:180–95. doi:10.4065/mcp.2009.0494

69 Shahar Y, Combi C. Timing is everything. Time-oriented clinical information systems. *West J Med* 1998.

70 Singh A, Nadkarni G, Gottesman O, *et al.* Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *J Biomed Inform* Published Online First: 2015. doi:10.1016/j.jbi.2014.11.005

71 Madkour M, Benhaddou D, Tao C. Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain. Comput. Methods Programs Biomed. 2016. doi:10.1016/j.cmpb.2016.02.007

72 Batal I, Fradkin D, Harrison J, *et al.* Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data. *KDD proceedings Int Conf Knowl Discov Data Min* 2012;**2012**:280–8. doi:10.1145/2339530.2339578

73 Batal I, Valizadegan H, Cooper GF, *et al.* A temporal pattern mining approach for classifying electronic health record data. *ACM Trans Intell Syst Technol* 2013;**4**:1–22. doi:10.1145/2508037.2508044

74 Teramukai S, Okuda Y, Miyazaki S, *et al.* Dynamic prediction model and risk assessment chart for cardiovascular disease based on on-treatment blood pressure and baseline risk factors. *Hypertens Res* Published Online First: 2016. doi:10.1038/hr.2015.120

75 Tangri N, Inker LA, Hiebert B, *et al.* A Dynamic Predictive Model for Progression of CKD. *Am J Kidney Dis* Published Online First: 2017. doi:10.1053/j.ajkd.2016.07.030

76 Nelson B, McGorry PD, Wichers M, *et al.* Moving from static to dynamic models of the onset of mental disorder a review. JAMA Psychiatry. 2017. doi:10.1001/jamapsychiatry.2017.0001

77 Nahid P, Dorman SE, Alipanah N, *et al.* Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. doi:10.1093/cid/ciw376

78 ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification. https://www.cdc.gov/nchs/icd/icd9cm.htm (accessed 26 Nov 2017).

79 Kasthurirathne SN, Dixon BE, Gichoya J, *et al.* Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. *J Biomed Inform* 2017;**69**:160–76. doi:10.1016/j.jbi.2017.04.008

80 Jiang M, Chen Y, Liu M, *et al.* A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Informatics Assoc* 2011;**18**:601–6. doi:10.1136/amiajnl-2011-000163

81 Zhong G, Wang L-N, Dong J. An Overview on Data Representation Learning: From Traditional Feature Learning to Recent Deep Learning. *J Financ Data Sci* Published Online

First: 2016. doi:10.1016/j.jfds.2017.05.001

82 Patel VL, Kaufman DR, Arocha JF. Emerging paradigms of cognition in medical decision-making. *J Biomed Inform* 2002;**35**:52–75.http://www.ncbi.nlm.nih.gov/pubmed/12415726 (accessed 18 Nov 2017).

83 Van Vleck TT, Stein DM, Stetson PD, *et al.* Assessing data relevance for automated generation of a clinical summary. *AMIA . Annu Symp proceedings AMIA Symp* 2007;**2007**:761–5.http://www.ncbi.nlm.nih.gov/pubmed/18693939 (accessed 18 Nov 2017).

84 Zhang R, Pakhomov S, Lee JT, *et al.* Navigating longitudinal clinical notes with an automated method for detecting new information. *Stud Health Technol Inform* 2013;**192**:754–8.http://www.ncbi.nlm.nih.gov/pubmed/23920658 (accessed 18 Nov 2017).

85 Farri O, Pieckiewicz DS, Rahman AS, *et al.* A qualitative analysis of EHR clinical document synthesis by clinicians. *AMIA . Annu Symp proceedings AMIA Symp* 2012;**2012**:1211–20.http://www.ncbi.nlm.nih.gov/pubmed/23304398 (accessed 17 Nov 2017).

86 South BR, Shen S, Jones M, *et al.* Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC Bioinformatics* 2009;**10 Suppl 9**:S12. doi:10.1186/1471-2105-10-S9-S12

87 Sulieman L, Fabbri D, Wang F, *et al.* Predicting Negative Events: Using Post-discharge Data to Detect High-Risk Patients. *AMIA Annu Symp Proc* 2016.

88 Sulieman L, Gilmore D, French C, *et al.* Classifying patient portal messages using Convolutional Neural Networks. *J Biomed Inform* Published Online First: 2017. doi:10.1016/j.jbi.2017.08.014

89 Panahiazar M, Taslimitehrani V, Pereira NL, *et al.* Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. In: *Studies in Health Technology and Informatics*. 2015. doi:10.3233/978-1-61499-512-8-369

90 Hao T, Rusanov A, Boland MR, *et al.* Clustering clinical trials with similar eligibility criteria features. *J Biomed Inform* Published Online First: 2014. doi:10.1016/j.jbi.2014.01.009

91 Geyer HL, Scherber RM, Dueck AC, *et al.* Distinct clustering of symptomatic burden among myeloproliferative neoplasm patients: Retrospective assessment in 1470 patients. *Blood* Published Online First: 2014. doi:10.1182/blood-2013-09-527903

92 Wilson JR, Grossman RG, Frankowski RF, *et al.* A Clinical Prediction Model for Long-Term Functional Outcome after Traumatic Spinal Cord Injury Based on Acute Clinical and Imaging Factors. *J Neurotrauma* Published Online First: 2012. doi:10.1089/neu.2012.2417

93 Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit Med* 2018;**1**:18. doi:10.1038/s41746-018-0029-1

94 Jackson RG, Patel R, Jayatilleke N, *et al.* Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* 2017;**7**:e012012. doi:10.1136/bmjopen-2016-012012

95 Xu H, Stenner SP, Doan S, *et al.* MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;**17**:19–24. doi:10.1197/jamia.M3378

96 Walsh C, Hripcsak G. The effects of data sources, cohort selection, and outcome

definition on a predictive model of risk of thirty-day hospital readmissions. *J Biomed Inform* Published Online First: 2014. doi:10.1016/j.jbi.2014.08.006

97 Valko M, Hauskrecht M. Feature importance analysis for patient management decisions. *Stud Health Technol Inform* 2010;**160**:861–5.http://www.ncbi.nlm.nih.gov/pubmed/20841808 (accessed 2 Dec 2017).

98 Capurro D, PhD MY, van Eaton E, *et al.* Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. *EGEMS (Washington, DC)* Published Online First: 2014. doi:10.13063/2327-9214.1079

99 Scheurwegs E, Luyckx K, Luyten L, *et al.* Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *J Am Med Informatics Assoc* Published Online First: 2016. doi:10.1093/jamia/ocv115

100 Shivade C, Raghavan P, Fosler-Lussier E, *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Informatics Assoc* 2014;**21**:221–30. doi:10.1136/amiajnl-2013-001935

101 Kusiak A. Feature transformation methods in data mining. *IEEE Trans Electron Packag Manuf* Published Online First: 2001. doi:10.1109/6104.956807

102 Zhang X, Yu FX, Karaman S, *et al.* Learning discriminative and transformation covariant local feature detectors. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. 2017. doi:10.1109/CVPR.2017.523

103 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science (80-)* Published Online First: 2006. doi:10.1126/science.1127647

104 Fung G, Dy JG, Masaeli M. From Transformation-Based Dimensionality Reduction to Feature Selection. *Proc 27th Int Conf Mach Learn* 2010.

105 Nargesian F, Samulowitz H, Khurana U, *et al.* Learning feature engineering for classification. In: *IJCAI International Joint Conference on Artificial Intelligence*. 2017. doi:10.24963/ijcai.2017/352

106 MTW, Liu H, Motoda H. Feature Extraction Construction and Selection: A Data Mining Perspective. *J Am Stat Assoc* Published Online First: 1999. doi:10.2307/2669967

107 Suarez-Alvarez MM, Pham DT, Prostov MY, *et al.* Statistical approach to normalization of feature vectors and clustering of mixed datasets. *Proc R Soc A Math Phys Eng Sci* Published Online First: 2012. doi:10.1098/rspa.2011.0704

108 Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. *J Mach Learn Res* Published Online First: 2003. doi:10.1016/j.aca.2011.07.027

109 Lustgarten JL, Gopalakrishnan V, Grover H, *et al.* Improving classification performance with discretization on biomedical datasets. *AMIA Annu Symp Proc* 2008.

110 Maslove DM, Podchiyska T, Lowe HJ. Discretization of continuous features in clinical datasets. *J Am Med Informatics Assoc* Published Online First: 2013. doi:10.1136/amiajnl-2012-000929

111 Irani K, Fayyad U. Multi-lnterval Discretization of Continuous-Valued Attributes for Classification learning. *Proc Natl Acad Sci U S A* Published Online First: 1993. doi:10.1109/TKDE.2011.181

112 Kerber R. Chimerge: Discretization of numeric attributes. *Proc tenth Natl Conf Artif Intell* 1992.

113 Dougherty J, Kohavi R, Sahami M. Supervised and Unsupervised Discretization of Continuous Features. In: *Machine Learning Proceedings 1995*. 1995. doi:10.1016/B978-1-55860-377-6.50032-3

114 Monti S, Cooper G. A latent variable model for multivariate discretization. *Seventh Int Work …* Published Online First: 1999. doi:10.1.1.50.3822

115 Hripcsak G, Albers DJ, Perotte A. Parameterizing time in electronic health record studies. *J Am Med Informatics Assoc* Published Online First: 2015. doi:10.1093/jamia/ocu051

116 Lasko TA. Nonstationary Gaussian Process Regression for Evaluating Clinical Laboratory Test Sampling Strategies. *Proc Conf AAAI Artif Intell* Published Online First: 2015. doi:10.1161/CIRCRESAHA.116.303790.The

117 Zhao J, Henriksson A, Kvist M, *et al.* Handling Temporality of Clinical Events for Drug Safety Surveillance. *AMIA Annu Symp Proc* 2015.

118 Post AR, Harrison JH. Temporal Data Mining. *Clin Lab Med* 2008;**28**:83–100. doi:10.1016/j.cll.2007.10.005

119 Zafar Nezhad M, Zhu D, Sadati N, *et al.* A Predictive Approach Using Deep Feature Learning for Electronic Medical Records: A Comparative Study. *arXiv* 2018.

120 Le Q V, Ranzato M, Monga R, *et al.* Building high-level features using large scale unsupervised learning. *Int Conf Mach Learn* Published Online First: 2011. doi:10.1109/MSP.2011.940881

121 Coates A, Arbor A, Ng AY. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. *Aistats 2011* Published Online First: 2011. doi:10.1109/ICDAR.2011.95

122 Mayfield E, Penstein-Rosé C. Using feature construction to avoid large feature spaces in text classification. In: *Proceedings of the 12th annual conference on Genetic and evolutionary computation - GECCO '10*. 2010. doi:10.1145/1830483.1830714

123 Mwangi B, Tian TS, Soares JC. A review of feature reduction techniques in Neuroimaging. Neuroinformatics. 2014. doi:10.1007/s12021-013-9204-3

124 Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2010. doi:10.1002/wics.101

125 Lever J, Krzywinski M, Altman N. Points of Significance: Principal component analysis. Nat. Methods. 2017. doi:10.1038/nmeth.4346

126 Rasmussen MA, Colding-Jørgensen M, Hansen LT, *et al.* Multivariate evaluation of pharmacological responses in early clinical trials - A study of rIL-21 in the treatment of patients with metastatic melanoma. *Br J Clin Pharmacol* Published Online First: 2010. doi:10.1111/j.1365-2125.2009.03600.x

127 Halai AD, Woollams AM, Lambon Ralph MA. Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex* Published Online First: 2017. doi:10.1016/j.cortex.2016.04.016

128 Federolf PA, Boyer KA, Andriacchi TP. Application of principal component analysis in clinical gait research: Identification of systematic differences between healthy and medial knee-osteoarthritic gait. *J Biomech* Published Online First: 2013. doi:10.1016/j.jbiomech.2013.06.032

129 Geraci J, Wilansky P, de Luca V, *et al.* Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evid Based Ment*

Health 2017;:ebmental-2017-102688. doi:10.1136/eb-2017-102688

130 Denny JC, Peterson JF, Choma NN, *et al.* Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc* 2010;**17**:383–8. doi:10.1136/jamia.2010.004804

131 Soguero-Ruiz C, Hindberg K, Mora-Jiménez I, *et al.* Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *J Biomed Inform* 2016;**61**:87–96. doi:10.1016/j.jbi.2016.03.008

132 Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, *et al.* Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep* 2017;**7**:46226. doi:10.1038/srep46226

133 Raju GS, Lum PJ, Slack RS, *et al.* Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. *Gastrointest Endosc* 2015;**82**:512–9. doi:10.1016/j.gie.2015.01.049

134 Kao A, Poteet S. Text mining and natural language processing. *ACM SIGKDD Explor Newsl* 2005;**7**:1–2. doi:10.1145/1089815.1089816

135 Hotho A, Nürnberger A, Paaß G. A brief survey of text mining. *LDV FORUM - Gld J Comput Linguist Lang Technol* Published Online First: 2005.http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.153.6679 (accessed 2 Dec 2017).

136 Allahyari M, Pouriyeh S, Assefi M, *et al.* A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *arXiv Prepr arXiv170702919* Published Online First: 10 July 2017.http://arxiv.org/abs/1707.02919 (accessed 2 Dec 2017).

137 Liddy E. Natural Language Processing. *Encycl Libr Inf Sci* Published Online First: 1 January 2001.http://surface.syr.edu/cnlp/11 (accessed 2 Dec 2017).

138 Névéol A, Zweigenbaum P, Processing SE for the IYS on CNL. Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. *Yearb Med Inform* 2015;**10**:194–8. doi:10.15265/IY-2015-035

139 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;**18**:544–51. doi:10.1136/amiajnl-2011-000464

140 Wu S, Miller T, Masanz J, *et al.* Negation's Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. *PLoS One* 2014;**9**:e112774. doi:10.1371/journal.pone.0112774

141 Dligach D, Bethard S, Becker L, *et al.* Discovering body site and severity modifiers in clinical texts. *J Am Med Informatics Assoc* 2014;**21**:448–54. doi:10.1136/amiajnl-2013-001766

142 2015AB UMLS National Drug File - Reference Terminology Source Information. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/ (accessed 7 Dec 2016).

143 SNOMED International. https://www.snomed.org/snomed-ct/ (accessed 8 Dec 2017).

144 RxNorm. https://www.nlm.nih.gov/research/umls/rxnorm/ (accessed 8 Dec 2017).

145 Davis MF, Sriram S, Bush WS, *et al.* Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *J Am Med Inform Assoc* 2013;**20**:e334-40. doi:10.1136/amiajnl-2013-001999

146 Denny JC, Irani PR, Wehbe FH, *et al.* The KnowledgeMap project: development of a

concept-based medical school curriculum database. *AMIA . Annu Symp proceedings AMIA Symp* 2003;**2003**:195–9.http://www.ncbi.nlm.nih.gov/pubmed/14728161 (accessed 3 Dec 2017).

147 Savova GK, Masanz JJ, Ogren P V, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507–13. doi:10.1136/jamia.2009.001560

148 Friedman C, Liu H, Shagina L, *et al.* Evaluating the UMLS as a source of lexical knowledge for medical language processing. *Proceedings AMIA Symp* 2001;:189–93.http://www.ncbi.nlm.nih.gov/pubmed/11825178 (accessed 14 Nov 2017).

149 Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *J Biomed Inform* 2013;**46**:1088–98. doi:10.1016/j.jbi.2013.08.004

150 ConText/NegEx | BLUlab. http://blulab.chpc.utah.edu/content/contextnegex (accessed 8 Dec 2017).

151 Szlosek DA, Ferrett J. Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems. *EGEMS (Washington, DC)* 2016;**4**:1222. doi:10.13063/2327-9214.1222

152 Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514–8. doi:10.1136/jamia.2010.003947

153 Turney PD, Pantel P. From frequency to meaning: Vector space models of semantics. *J Artif Intell Res* Published Online First: 2010. doi:10.1613/jair.2934

154 Rumshisky A, Ghassemi M, Naumann T, *et al.* Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry* 2016;**6**:e921. doi:10.1038/tp.2015.182

155 i2b2: Informatics for Integrating Biology &amp; the Bedside. https://www.i2b2.org/index.html (accessed 3 Dec 2017).

156 Marafino BJ, John Boscardin W, Adams Dudley R. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform* 2015;**54**:114–20. doi:10.1016/J.JBI.2015.02.003

157 Liao KP, Cai T, Savova GK, *et al.* Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* 2015;**350**:h1885. doi:10.1136/BMJ.H1885

158 Al-Haddad MA, Friedlin J, Kesterson J, *et al.* Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. *HPB* 2010;**12**:688–95. doi:10.1111/j.1477-2574.2010.00235.x

159 Temple MW, Lehmann CU, Fabbri D. Natural Language Processing for Cohort Discovery in a Discharge Prediction Model for the Neonatal ICU. *Appl Clin Inform* 2016;**7**:101–15. doi:10.4338/ACI-2015-09-RA-0114

160 Wu ST, Sohn S, Ravikumar KE, *et al.* Automated chart review for asthma cohort identification using natural language processing: an exploratory study. *Ann Allergy Asthma Immunol* 2013;**111**:364–9. doi:10.1016/j.anai.2013.07.022

161 Rosenbloom ST, Denny JC, Xu H, *et al.* Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Informatics Assoc* 2011;**18**:181–6. doi:10.1136/jamia.2010.007237

162 Walji MF, Kalenderian E, Tran D, *et al.* Detection and characterization of usability problems in structured data entry interfaces in dentistry. *Int J Med Inform* 2013;**82**:128–38. doi:10.1016/j.ijmedinf.2012.05.018

163 Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Inform Assoc* 2004;**11**:104–12. doi:10.1197/jamia.M1471

164 LOINC — The freely available standard for identifying health measurements, observations, and documents. https://loinc.org/ (accessed 8 Dec 2017).

165 Karystianis G, Sheppard T, Dixon WG, *et al.* Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. *BMC Med Inform Decis Mak* 2016;**16**:18. doi:10.1186/s12911-016-0255-x

166 Ford E, Carroll JA, Smith HE, *et al.* Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016;:ocv180. doi:10.1093/jamia/ocv180

167 Resnik P, Niv M, Nossal M, *et al.* Communication of Clinically Relevant Information in Electronic Health Records: A Comparison between Structured Data and Unrestricted Physician Language | Perspectives. *Perspect Heal Inf Manag* Published Online First: 2008.http://perspectives.ahima.org/communication-of-clinically-relevant-information-in-electronic-health-records-a-comparison-between-structured-data-and-unrestricted-physician-language/

168 Ford E, Carroll JA, Smith HE, *et al.* Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Informatics Assoc* 2016;**23**:1007–15. doi:10.1093/jamia/ocv180

169 Meystre SM, Friedlin FJ, South BR, *et al.* Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol* 2010;**10**:70. doi:10.1186/1471-2288-10-70

170 Fort K, Nazarenko A, Rosset S. Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis. *Proc COLING 2012 Tech Pap pages 895–910, COLING 2012, Mumbai, December 2012* 2012;:895–910.

171 Dumitrache A, Aroyo L, Welty C. Achieving Expert-Level Annotation Quality with CrowdTruth The Case of Medical Relation Extraction. http://ceur-ws.org/Vol-1428/BDM2I_2015_paper_3.pdf (accessed 12 Nov 2017).

172 Marcheggiani D, Sebastiani F. On the Effects of Low-Quality Training Data on Information Extraction from Clinical Reports. *J Data Inf Qual* 2017;**9**:1–25. doi:10.1145/3106235

173 Murphy KP. *Machine Learning: A Probablistic Perspective*. 2012. doi:10.1007/SpringerReference_35834

174 Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. doi:10.1023/A:1010933404324

175 Shmueli G. To Explain or To Predict? 2009. doi:10.2139/ssrn.1351252

176 Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics* Published Online First: 2018. doi:10.1186/s12859-018-2264-5

177 Goodfellow Ian, Bengio Yoshua CA. *Deep Learning*. MIT Press 2016. http://www.deeplearningbook.org

178 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44. doi:10.1038/nature14539

179 Bastien F, Bengio Y, Bergeron A, *et al.* Deep Self-Taught Learning for Handwritten Character Recognition. *Found Trends® Mach Learn* 2010;**2**:1–127. doi:10.1561/2200000006

180 Urban G, Geras KJ, Kahou SE, *et al.* Do Deep Convolutional Nets Really Need to be Deep and Convolutional? *Nature* 2016;**521**:436–44. doi:10.1038/nature14539

181 Sutskever I, Hinton GE. Deep, Narrow Sigmoid Belief Networks Are Universal Approximators. *Neural Comput* 2008;**20**:2629–36. doi:10.1162/neco.2008.12-07-661

182 Schmidhuber J. Deep learning in neural networks: An overview. *Neural Networks* 2015;**61**:85–117. doi:10.1016/J.NEUNET.2014.09.003

183 Sze V, Chen Y-H, Yang T-J, *et al.* Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Published Online First: 27 March 2017.http://arxiv.org/abs/1703.09039 (accessed 24 May 2017).

184 Kim Y. Convolutional Neural Networks for Sentence Classification. *arXiv14085882 [cs]* Published Online First: August 2014. doi:10.3115/v1/D14-1181

185 Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst 25* 2012;:1097–105.http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

186 Lawrence S, Giles CL, Ah Chung Tsoi, *et al.* Face recognition: a convolutional neural-network approach. *IEEE Trans Neural Networks* 1997;**8**:98–113. doi:10.1109/72.554195

187 Cire??an DC, Meier U, Gambardella LM, *et al.* Convolutional neural network committees for handwritten character classification. In: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. 2011. 1135–9. doi:10.1109/ICDAR.2011.229

188 Nguyen P, Tran T, Wickramasinghe N, *et al.* Deepr: A Convolutional Net for Medical Records. Published Online First: 25 July 2016.http://arxiv.org/abs/1607.07519 (accessed 1 Dec 2017).

189 Bajor JM, Lasko TA. PREDICTING MEDICATIONS FROM DIAGNOSTIC CODES WITH RECURRENT NEURAL NETWORKS. *ICLR* Published Online First: 2016.https://pdfs.semanticscholar.org/7ebf/ef7098ae93b7596a275ab17cc136f958262b.pdf (accessed 1 Dec 2017).

190 Lipton ZC, Kale DC, Elkan C, *et al.* Learning to Diagnose with LSTM Recurrent Neural Networks. Published Online First: 11 November 2015.http://arxiv.org/abs/1511.03677 (accessed 1 Dec 2017).

191 Bradbury J, Socher R. MetaMind Neural Machine Translation System for WMT 2016. ;**2**:264–7.http://www.statmt.org/wmt16/pdf/W16-2308.pdf (accessed 19 Nov 2017).

192 Xiong C, Merity S, Socher R. Dynamic Memory Networks for Visual and Textual Question Answering. *ICML* 2016;:8.http://arxiv.org/abs/1603.01417

193 Socher R, Pennington J, Huang EH, *et al.* Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In: *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics 2011. 151–61. doi:10.1.1.224.9432

194 Joulin A, Grave E, Bojanowski P, *et al.* Bag of Tricks for Efficient Text Classification. *arXiv160701759 [cs]* Published Online First: July 2016. doi:1511.09249v1

195 Mikolov T, Sutskever I, Chen K, *et al.* Distributed Representations of Words and Phrases and their Compositionality. *arXiv13104546 [cs, stat]* Published Online First: October 2013.http://arxiv.org/abs/1310.4546 (accessed 7 Sep 2016).

196 Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: : Association for Computational Linguistics 2014. 1532–43.http://www.aclweb.org/anthology/D14-1162 (accessed 7 Sep 2016).

197 Mikolov T, Corrado G, Chen K, *et al.* Efficient Estimation of Word Representations in Vector Space. *Proc Int Conf Learn Represent (ICLR 2013)* Published Online First: 2013. doi:10.1162/153244303322533223

198 Trask A, Gilmore D, Russell M. Modeling Order in Neural Word Embeddings at Scale. *arXiv150602338 [cs]* Published Online First: June 2015.http://arxiv.org/abs/1506.02338 (accessed 22 Dec 2016).

199 Nikfarjam A, Sarker A, O'Connor K, *et al.* Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Informatics Assoc* 2015;**22**:671–81. doi:10.1093/jamia/ocu041

200 Le Q V., Mikolov T. Distributed Representations of Sentences and Documents. *arXiv14054053 [cs]* Published Online First: May 2014. doi:10.1145/2740908.2742760

201 Kageback M, Mogren O, Tahmasebi N, *et al.* Extractive Summarization using Continuous Vector Space Models. *Proc 2nd Work Contin Vector Sp Model their Compos* 2014;:31–9. doi:10.1007/978-3-642-14834-7_15

202 Lai S, Xu L, Liu K, *et al.* Recurrent Convolutional Neural Networks for Text Classification. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015. 2267–73.http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9745 (accessed 7 Sep 2016).

203 Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014. doi:10.3115/v1/P14-1062

204 Conneau A, Schwenk H, Barrault L, *et al.* Very Deep Convolutional Networks for Natural Language Processing. *KI - Künstliche Intelligenz* 2016;**26**:357–63. doi:10.1007/s13218-012-0198-z

205 Xue X, Yin X. Topic modeling for named entity queries. *Proc 20th ACM Int Conf Inf Knowl Manag - CIKM '11* 2011;:2009. doi:10.1145/2063576.2063877

206 Grnarova P, Schmidt F, Hyland SL, *et al.* Neural Document Embeddings for Intensive Care Patient Mortality Prediction. *arXiv161200467 [cs]* Published Online First: December 2016.http://arxiv.org/abs/1612.00467 (accessed 22 Jan 2017).

207 Fontein DBY, Klinten Grand M, Nortier JWR, *et al.* Dynamic prediction in breast cancer: Proving feasibility in clinical practice using the TEAM trial. *Ann Oncol* Published Online First: 2015. doi:10.1093/annonc/mdv146

208 Hubbard A, Munoz ID, Decker A, *et al.* Time-dependent prediction and evaluation of variable importance using superlearning in high-dimensional clinical data. In: *Journal of Trauma and Acute Care Surgery*. 2013. doi:10.1097/TA.0b013e3182914553

209 Hospital Guide to Reducing Medicaid Readmissions. https://www.ahrq.gov/sites/default/files/publications/files/medreadmissions.pdf (accessed 26 Nov 2017).

210 Farrar S, Ryan M, Ross D, *et al.* Using discrete choice modelling in priority setting: an application to clinical service developments. *Soc Sci Med* 2000;**50**:63–75.http://www.ncbi.nlm.nih.gov/pubmed/10622695 (accessed 26 Nov 2017).

211 Mitton C, Donaldson C. Health care priority setting: principles, practice and challenges. *Cost Eff Resour Alloc* 2004;**2**:3. doi:10.1186/1478-7547-2-3

212 Kansagara D, Englander H, Salanitro A, *et al.* Risk Prediction Models for Hospital Readmission. *JAMA* 2011;**306**:1688. doi:10.1001/jama.2011.1515

213 Ben-Chetrit E, Chen-Shuali C, Zimran E, *et al.* A simplified scoring tool for prediction of readmission in elderly patients hospitalized in internal medicine departments. *Isr Med Assoc J* 2012;**14**:752–6.http://www.ncbi.nlm.nih.gov/pubmed/23393714 (accessed 26 Nov 2017).

214 Van Walraven C, Dhalla IA, Bell C, *et al.* Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj* 2010;**182**:551–7. doi:10.1503/cmaj.091117

215 Kansagara D, Englander H, Salanitro A, *et al.* Readmission risk modeling: A systematic review. *J Gen Intern Med* 2011;**26**:S125. doi:http://dx.doi.org/10.1007/s11606-011-1730-9

216 Merkow RP, Ju MH, Chung JW, *et al.* Underlying Reasons Associated With Hospital Readmission Following Surgery in the United States. *JAMA* 2015;**313**:483. doi:10.1001/jama.2014.18614

217 Morris MS, Deierhoi RJ, Richman JS, *et al.* The Relationship Between Timing of Surgical Complications and Hospital Readmission. *JAMA Surg* 2014;**149**:348. doi:10.1001/jamasurg.2013.4064

218 Opheim A, Danielsson A, Alt Murphy M, *et al.* Early prediction of long-term upper limb spasticity after stroke. *Neurology* 2015;**85**:873–80. doi:10.1212/WNL.0000000000001908

219 Hersh AM, Masoudi FA, Allen LA. Postdischarge Environment Following Heart Failure Hospitalization: Expanding the View of Hospital Readmission. *J Am Heart Assoc* 2013;**2**:e000116–e000116. doi:10.1161/JAHA.113.000116

220 DeLia D, Tong J, Gaboda D, *et al.* Post-Discharge Follow-Up Visits and Hospital Utilization by Medicare Patient 2007–2010. *Medicare Medicaid Res Rev* 2014;**4**:E1–19. doi:10.5600/mmrr.004.02.a01

221 Hernandez AF, Greiner MA, Fonarow GC, *et al.* Relationship Between Early Physician Follow-up and 30-Day Readmission Among Medicare Beneficiaries Hospitalized for Heart Failure. *JAMA* 2010;**303**:1716. doi:10.1001/jama.2010.533

222 Jackson C, Shahsahebi M, Wedlake T, *et al.* Timeliness of outpatient follow-up: an evidence-based approach for planning after hospital discharge. *Ann Fam Med* 2015;**13**:115–22. doi:10.1370/afm.1753

223 Sharma G, Kuo Y-F, Freeman JL, *et al.* Outpatient Follow-up Visit and 30-Day Emergency Department Visit and Readmission in Patients Hospitalized for Chronic Obstructive Pulmonary Disease. *Arch Intern Med* 2010;**170**:1664–70. doi:10.1001/archinternmed.2010.345

224 Wang F, Lee N, Hu J, *et al.* Towards heterogeneous temporal clinical event pattern discovery. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. 2012. doi:10.1145/2339530.2339605

225 Caballero K, Akella R. Dynamic Estimation of the Probability of Patient Readmission to the ICU using Electronic Medical Records. *AMIA Annu Symp Proc* 2015.

226 Pedregosa F, Varoquaux G. *Scikit-learn: Machine learning in Python*. 2011. doi:10.1007/s13398-014-0173-7.2

227 Rosen PR, Groshen S, Saigo PE, *et al.* A long-term follow-up study of survival in stage I (T1N0M0) and stage II (T1N1M0) breast carcinoma. *J Clin Oncol* 1989.

228 Saphner T, Tormey DC, Gray R. Annual hazard rates of recurrence for breast cancer after primary therapy. *J Clin Oncol* Published Online First: 1996. doi:10.1200/JCO.1996.14.10.2738

229 Yu K Da, Wu J, Shen ZZ, *et al.* Hazard of breast cancer-specific mortality among women with estrogen receptor-positive breast cancer after five years from diagnosis: Implication for extended endocrine therapy. *J Clin Endocrinol Metab* Published Online First: 2012. doi:10.1210/jc.2012-2423

230 Gundlapalli A V, Carter ME, Palmer M, *et al.* Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. *AMIA Annu Symp Proc* 2013.

231 Cronin RM, Fabbri D, Denny JC, *et al.* Automated Classification of Consumer Health Information Needs in Patient Portal Messages. *AMIA . Annu Symp proceedings AMIA Symp* 2015;**2015**:1861–70.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4765690&tool=pmcentrez&rendertype=abstract (accessed 7 Sep 2016).

232 Miñarro-Giménez J. Applying deep learning techniques on medical corpora from the World Wide Web: a prototypical system and evaluation. *arXiv Prepr arXiv …* 2015;:1–14.http://arxiv.org/abs/1502.03682 (accessed 7 Sep 2016).

233 Amunategui M, Markwell T, Rozenfeld Y. Prediction Using Note Text: Synthetic Feature Creation with word2vec. *arXiv150305123 [cs]* 2015;:13.http://arxiv.org/abs/1503.05123 (accessed 7 Sep 2016).

234 Wakefield DS, Mehr D, Keplinger L, *et al.* Issues and questions to consider in implementing secure electronic patient-provider web portal communications systems. *Int J Med Inform* 2010;**79**:469–77. doi:10.1016/j.ijmedinf.2010.04.005

235 Emont S. Measuring the Impact of Patient Portals: What the Literature Tells Us - CHCF.org. http://www.chcf.org/publications/2011/05/measuring-impact-patient-portals (accessed 21 Nov 2016).

236 Goel MS, Brown TL, Williams A, *et al.* Patient reported barriers to enrolling in a patient portal. *J Am Med Informatics Assoc* 2011;**18**:i8–12. doi:10.1136/amiajnl-2011-000473

237 Detmer D, Bloomrosen M, Raymond B, *et al.* Integrated Personal Health Records: Transformative Tools for Consumer-Centric Care. *BMC Med Inform Decis Mak* 2008;**8**:45. doi:10.1186/1472-6947-8-45

238 Kittler AF, Carlson GL, Harris C, *et al.* Primary care physician attitudes towards using a secure web-based portal designed to facilitate electronic communication with patients. *Inform Prim Care* 2004;**12**:129–38. doi:10.2196/jmir.8.1.e2

239 Zickmund SL, Hess R, Bryce CL, *et al.* Interest in the use of computerized patient portals: Role of the provider-patient relationship. *J Gen Intern Med* 2008;**23**:20–6. doi:10.1007/s11606-007-0273-6

240 Wade-Vuturo AE, Mayberry LS, Osborn CY. Secure messaging and diabetes management: experiences and perspectives of patient portal users. *J Am Med Informatics Assoc* 2013;**20**:519–25. doi:10.1136/amiajnl-2012-001253

241 Haun JN, Lind JD, Shimada SL, *et al.* Evaluating user experiences of the secure messaging tool on the veterans affairs' patient portal system. *J Med Internet Res* 2014;**16**:e75. doi:10.2196/jmir.2976

242 Goel MS, Brown TL, Williams A, *et al.* Disparities in enrollment and use of an electronic patient portal. *J Gen Intern Med* 2011;**26**:1112–6. doi:10.1007/s11606-011-1728-3

243 Osborn CY, Mayberry LS, Wallston KA, *et al.* Understanding patient portal use: Implications for medication management. *J Med Internet Res* 2013;**15**:e133. doi:10.2196/jmir.2589

244 North F, Crane SJ, Stroebel RJ, *et al.* Patient-generated secure messages and eVisits on a patient portal: are patients at risk? *J Am Med Informatics Assoc* 2013;**20**:1143–9. doi:10.1136/amiajnl-2012-001208

245 Boffa M, Weathers A, Ouyang B. Analysis of Patient Portal Message Content in an Academic Multi-specialty Neurology Practice (S11.005). *Neurology* 2015;**84**:S11.005.http://www.neurology.org/content/84/14_Supplement/S11.005 (accessed 27 Feb 2017).

246 Cronin RM, Fabbri D, Denny JC, *et al.* A comparison of rule-based and machine learning approaches for classifying patient portal messages. *Int J Med Inform* 2017;**105**:110–20. doi:10.1016/j.ijmedinf.2017.06.004

247 Robinson JR, Valentine A, Carney C, *et al.* Complexity of medical decision-making in care provided by surgeons through patient portals. *J Surg Res* 2017;**214**:93–101. doi:10.1016/j.jss.2017.02.077

248 Shenson JA, Ingram E, Colon N, *et al.* Application of a Consumer Health Information Needs Taxonomy to Questions in Maternal-Fetal Care. *AMIA Annu Symp Proc* 2015;**2015**:1148–56.http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765568/ (accessed 22 Jan 2017).

249 Purcell GP. Surgical textbooks: past, present, and future. *Ann Surg* 2003;**238**:S34-41.http://www.ncbi.nlm.nih.gov/pubmed/14703743

250 Grevet C, Choi D, Kumar D, *et al.* Overload is overloaded. In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. New York, NY, USA: : ACM 2014. 793–802. doi:10.1145/2556288.2557013

251 Yoo S, Yang Y, Lin F, *et al.* Mining social networks for personalized email prioritization. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*. New York, NY, USA: : ACM 2009. 967. doi:10.1145/1557019.1557124

252 Roberts A, Gaizauskas R, Hepple M, *et al.* Building a semantically annotated corpus of clinical texts. *J Biomed Inform* 2009;**42**:950–66. doi:10.1016/j.jbi.2008.12.013

253 Weingart SN, Hamrick HE, Tutkus S, *et al.* Medication safety messages for patients via the web portal: The MedCheck intervention. *Int J Med Inform* 2008;**77**:161–8.

doi:10.1016/j.ijmedinf.2007.04.007

254 Whittaker S, Matthews T, Cerruti J, *et al.* Am I wasting my time organizing email? In: *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. New York, NY, USA: : ACM 2011. 3449. doi:10.1145/1978942.1979457

255 Dredze M, Lau T, Kushmerick N. Automatically classifying emails into activities. In: *Proceedings of the 11th international conference on Intelligent user interfaces - IUI '06*. New York, NY, USA: : ACM 2006. 70. doi:10.1145/1111449.1111471

256 Coussement K, Van den Poel D. Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decis Support Syst* 2008;**44**:870–82. doi:10.1016/j.dss.2007.10.010

257 Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. *Methods Inf Med* 1993;**32**:281–91. doi:10.1136/jamia.1998.0050001

258 Dumais S, Platt J, Heckerman D, *et al.* Inductive learning algorithms and representations for text categorization. In: *Proceedings of the seventh international conference on Information and knowledge management - CIKM '98*. New York, NY, USA: : ACM 1998. 148–55. doi:10.1145/288627.288651

259 Jackson P, Moulinier I. *Natural Language Processing for Online Applications: Text retrieval, extraction and categorization. Second revised edition*. John Benjamins Publishing 2007.

260 Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 2009;**45**:427–37. doi:10.1016/j.ipm.2009.03.002

261 Scott S, Matwin S. Feature Engineering for Text Classification. In: *Proceedings of the Sixteenth International Conference on Machine Learning*. San Francisco, CA, USA: : Morgan Kaufmann Publishers Inc. 1999. 379–88.http://dl.acm.org/citation.cfm?id=645528.657484 (accessed 23 Jan 2017).

262 Wang P, Domeniconi C. Building semantic kernels for text classification using wikipedia. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. New York, NY, USA: : ACM 2008. 713. doi:10.1145/1401890.1401976

263 Jiang C, Coenen F, Sanderson R, *et al.* Text classification using graph mining-based feature extraction. *Knowledge-Based Syst* 2010;**23**:302–8. doi:10.1016/j.knosys.2009.11.010

264 Sharma R, Raman S. Phrase-based Text Representation for Managing the Web Documents. In: *Proceedings of the International Conference on Information Technology: Computers and Communications*. Washington, DC, USA: : IEEE Computer Society 2003. 165--.http://dl.acm.org/citation.cfm?id=844385.845970 (accessed 23 Jan 2017).

265 Lewis DD. An evaluation of phrasal and clustered representations on a text categorization task. In: *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '92*. New York, NY, USA: : ACM 1992. 37–50. doi:10.1145/133160.133172

266 Coenen F, Leng P, Sanderson R, *et al.* Statistical Identification of Key Phrases for Text Classification. In: *Machine Learning and Data Mining in Pattern Recognition*. Springer, Berlin, Heidelberg 2007. 838–53. doi:10.1007/978-3-540-73499-4_63

267 Mihalcea R, Radev D. *Graph-based Natural Language Processing and Information*

Retrieval. Cambridge University Press 2011.

268 Debusmann R, Kuhlmann M. Dependency Grammar: Classification and Exploration. In: Crocker MW, Siekmann J, eds. *Resource-Adaptive Cognitive Processes*. Springer Berlin Heidelberg 2010. 365–88. doi:10.1007/978-3-540-89408-7_16

269 Schenker A, Last M, Bunke H, *et al.* Classification of Web documents using a graph model. In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* 2003. 240–4. doi:10.1109/ICDAR.2003.1227666

270 Luo Y, Sohani AR, Hochberg EP, *et al.* Automatic lymphoma classification with sentence subgraph mining from pathology reports. *J Am Med Informatics Assoc* 2014;**21**:824–32. doi:10.1136/amiajnl-2013-002443

271 Luo Y, Xin Y, Hochberg E, *et al.* Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. *J Am Med Informatics Assoc* 2015;**22**:1009–19. doi:10.1093/jamia/ocv016

272 Kumar A, Irsoy O, Su J, *et al.* Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. *arXiv* 2015;:1–10.http://arxiv.org/abs/1506.07285 (accessed 3 Oct 2016).

273 Osborn CY, Rosenbloom ST, Stenner SP, *et al.* MyHealthAtVanderbilt: policies and procedures governing patient portal functionality. *J Am Med Informatics Assoc* 2011;**18**:i18–23. doi:10.1136/amiajnl-2011-000184

274 Roden D, Pulley J, Basford M, *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther* 2008;**84**:362–9. doi:10.1038/clpt.2008.89

275 Jackson GP, Shenson JA, Ingram E, *et al.* A Taxonomy of Consumer Health-Related Communications for Characterizing the Content of Patient Portal Messages. *Pers Commun*

276 Natural Language Toolkit (NLTK). NLTK Packag. 2016.http://www.nltk.org/api/nltk.html

277 Amoore L. Cloud geographies: Computing, data, sovereignty. *Prog Hum Geogr* 2016;:0309132516662147. doi:10.1177/0309132516662147

278 Reasoning D. Synthesys Technology Overview. http://www.digitalreasoning.com/resources/Synthesys_v3.9_Technology_Overview_FINAL_Jan_2015.pdf (accessed 4 Dec 2016).

279 Řehůřek R, Sojka P. Gensim: topic modeling for humans. 2010.https://radimrehurek.com/gensim/models/word2vec.html

280 Li M, Carrell D, Aberdeen J, *et al.* De-identification of clinical narratives through writing complexity measures. *Int J Med Inform* 2014;**83**:750–67. doi:10.1016/j.ijmedinf.2014.07.002

281 Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* 2014;**15**:1929–58. doi:10.1214/12-AOS1000

282 Hinton GE, Srivastava N, Krizhevsky A, *et al.* Improving neural networks by preventing co-adaptation of feature detectors. *arXiv12070580 [cs]* Published Online First: July 2012. doi:arXiv:1207.0580

283 Bergstra J, Yamins D, Cox DD. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *12th PYTHON Sci CONF (SCIPY 2013)* 2013;:13–

20.http://hyperopt.github.io/hyperopt/%5Cnhttps://github.com/jaberg/hyperopt%5Cnhttp://www.youtube.com/watch?v=Mp1xnPfE4PY

284 Lex A, Gehlenborg N, Strobelt H, *et al.* UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph (IEEE InfoVis';14)* 2014.

285 UpSet: Visualizing Intersecting Sets. http://www.caleydo.org/tools/upset/ (accessed 7 Dec 2016).

286 Robinson J, Davis SA, Cronin RM, *et al.* Use of a patient portal during hospital admissions to surgical services. *AMIA Annu Symp Proc* 2016;**2016**:1967–76.

287 Masterman M, Cronin RM, Davis SE, *et al.* Adoption of secure messaging in a patient portal across pediatric specialties. *AMIA Annu Symp Proc* 2016;**2016**:1967–76.

288 Shenson JA, Cronin RM, Davis SE, *et al.* Rapid growth in surgeons??? use of secure messaging in a patient portal. *Surg Endosc Other Interv Tech* 2016;**30**:1432–40. doi:10.1007/s00464-015-4347-y

289 Cronin RM, Davis SE, Shenson JA, *et al.* Growth of Secure Messaging Through a Patient Portal as a Form of Outpatient Interaction across Clinical Specialties. *Appl Clin Inform* 2015;**6**:288–304. doi:10.4338/ACI-2014-12-RA-0117

290 Abu-Mostafa YS, Magdon-Ismail M, Lin H-T. *Learning from Data*. 2012. doi:10.1007/s13398-014-0173-7.2

291 Langley P. Machine Learning as an Experimental Science. *Mach Learn* 1988;**3**:5–8. doi:10.1023/A:1022623814640

292 Wagner R, Thom M, Schweiger R, *et al.* Learning convolutional neural networks from few samples. In: *Proceedings of the International Joint Conference on Neural Networks*. 2013. doi:10.1109/IJCNN.2013.6706969

293 Liu S, Deng W. Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size. In: *2015 3rd IAPR Asian Conference on Pattern Recognition*. IEEE 2015. 730–4. doi:10.1109/ACPR.2015.7486599

294 Gibson B, Butler J, Zirkle M, *et al.* Foraging for Information in the EHR: The Search for Adherence Related Information by Mental Health Clinicians. *AMIA . Annu Symp proceedings AMIA Symp* 2016;**2016**:600–8.

295 Reichert D, Kaufman D, Bloxham B, *et al.* Cognitive analysis of the summarization of longitudinal patient records. *AMIA Annu Symp Proc* 2010;**2010**:667–71.

296 Crammer K, Dredze M, Ganchev K, *et al.* ++Automatic code assignment to medical text. *Proc Work BioNLP 2007 Biol Transl Clin Lang Process - BioNLP '07* 2007;:129. doi:10.3115/1572392.1572416

297 Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. In: *BMC Bioinformatics*. 2008. doi:10.1186/1471-2105-9-S3-S10

298 Pakhomov SVS, Buntrock JD, Chute CG. Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques. *J Am Med Informatics Assoc* 2006;**13**:516–25. doi:10.1197/jamia.M2077

299 Tang B, Cao H, Wu Y, *et al.* Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC Med Inform Decis Mak* 2013;**13**. doi:10.1186/1472-6947-13-S1-S1

300 Hanauer DA, Mei Q, Law J, *et al.* Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and

using the Electronic Medical Record Search Engine (EMERSE). *J Biomed Inform* 2015;**55**:290–300. doi:10.1016/j.jbi.2015.05.003

301 Gobbel GT, Reeves R, Jayaramaraja S, *et al.* Development and evaluation of RapTAT: A machine learning system for concept mapping of phrases from medical narratives. *J Biomed Inform* 2014;**48**:54–65. doi:10.1016/j.jbi.2013.11.008

302 Nath, C; Albaghdadi, M. S;Jonnalagadda S. A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports. *PLoS One* 2016;**11**.https://www.scopus.com/inward/record.uri?eid=2-s2.0-84965164707&doi=10.1371%2Fjournal.pone.0153749&partnerID=40&md5=634e48d61564e2bd713978c33d1ae0d6

303 Gobbel Dr GT, Garvin J, Reeves R, *et al.* Assisted annotation of medical free text using RapTAT. *J Am Med Informatics Assoc* 2014;**21**:833–41. doi:10.1136/amiajnl-2013-002255

304 Wang Y, Wang L, Rastegar-Mojarad M, *et al.* Clinical information extraction applications: A literature review. J. Biomed. Inform. 2018;**77**:34–49. doi:10.1016/j.jbi.2017.11.011

305 Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks. https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf (accessed 19 Nov 2017).

306 Young T, Hazarika D, Poria S, *et al.* Recent Trends in Deep Learning Based Natural Language Processing. arXiv. 2017;:1–22.https://arxiv.org/pdf/1708.02709.pdf%0Ahttps://arxiv.org/pdf/1708.02709.pdf%0Ahttp://arxiv.org/abs/1708.02709

307 Peng H, Li J, He Y, *et al.* Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN. *Proc 2018 World Wide Web Conf World Wide Web - WWW '18* Published Online First: 2018. doi:10.1145/3178876.3186005

308 Rosenbloom ST, Stead WW, Denny JC, *et al.* Generating Clinical Notes for Electronic Health Record Systems. *Appl Clin Inform* 2010;**1**:232–43. doi:10.4338/ACI-2010-03-RA-0019

309 Chalmers DJ, Deakyne SJ, Payan ML, *et al.* Feasibility of integrating research data collection into routine clinical practice using the electronic health record. *J Urol* 2014;**192**:1215–20. doi:10.1016/j.juro.2014.04.091

310 Chiticariu L, Li Y, Reiss FR. Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems! In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013. 827–32.http://aclanthology.info/papers/rule-based-information-extraction-is-dead-long-live-rule-based-information-extraction-systems

311 Friedman C, Johnson SB, Forman B, *et al.* Architectural requirements for a multipurpose natural language processor in the clinical environment. *Proc Symp Comput Appl Med Care* 1995;:347–51.

312 Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. *J Biomed Inform* 2009;**42**:923–36. doi:10.1016/j.jbi.2009.07.007

313 Kasthurirathne SN, Dixon BE, Gichoya J, *et al.* Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. *J Biomed Inform* 2017;**69**:160–76. doi:10.1016/j.jbi.2017.04.008

314 Kang N, Singh B, Afzal Z, *et al.* Using rule-based natural language processing to improve disease normalization in biomedical text. *J Am Med Informatics Assoc* 2013;**20**:876–81. doi:10.1136/amiajnl-2012-001173

315 Wang L, Ruan X, Yang P, *et al.* Comparison of three information sources for smoking information in electronic health records. Cancer Inform. 2016;**15**:237–42. doi:10.4137/CIN.S40604

316 Babashzadeh A, Huang J, Daoud M. Exploiting Semantics for Improving Clinical Information Retrieval. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2013. doi:10.1145/2484028.2484167

317 Moen H, Ginter F, Marsi E, *et al.* Care episode retrieval: Distributional semantic models for information retrieval in the clinical domain. In: *BMC Medical Informatics and Decision Making*. 2015. doi:10.1186/1472-6947-15-S2-S2

318 Sutskever I. Training Recurrent neural Networks. *PhD thesis* 2013;:101.

319 Hochreiter S, Urgen Schmidhuber J. LONG SHORT-TERM MEMORY. *Neural Comput* 1997;**9**:1735–80. doi:10.1162/neco.1997.9.8.1735

320 Palangi H, Deng L, Shen Y, *et al.* Deep Sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Trans Audio Speech Lang Process* 2016;**24**:694–707. doi:10.1109/TASLP.2016.2520371

321 Socher R, Perelygin A, Wu J. Recursive deep models for semantic compositionality over a sentiment treebank. In: *Emnlp*. Citeseer 2013. 1631–42. doi:10.1371/journal.pone.0073791

322 Liu Z, Yang M, Wang X, *et al.* Entity recognition from clinical texts via recurrent neural network. *BMC Med Inform Decis Mak* 2017;**17**:67. doi:10.1186/s12911-017-0468-7

323 Yang Z, Yang D, Dyer C, *et al.* Hierarchical Attention Networks for Document Classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016. 1480–9. doi:10.18653/v1/N16-1174

324 Gao S, Young MT, Qiu JX, *et al.* Hierarchical attention networks for information extraction from cancer pathology reports. *J Am Med Informatics Assoc* Published Online First: 16 November 2017. doi:10.1093/jamia/ocx131

325 Luo L, Li L, Hu J, *et al.* A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. *BMC Med Inform Decis Mak* 2016;**16**:114. doi:10.1186/s12911-016-0357-5

326 Zheng S, Jabbour SK, O'Reilly SE, *et al.* Automated Information Extraction on Treatment and Prognosis for Non–Small Cell Lung Cancer Radiotherapy Patients: Clinical Study. *JMIR Med Informatics* Published Online First: 2018. doi:10.2196/medinform.8662

327 Saczynski JS, Lessard D, Spencer FA, *et al.* Declining length of stay for patients hospitalized with AMI: impact on mortality and readmissions. *Am J Med* 2010;**123**:1007–15. doi:10.1016/j.amjmed.2010.05.018

328 Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;**31**:1102–10. doi:10.1038/nbt.2749

329 Liede A, Hernandez RK, Roth M, *et al.* Validation of international classification of diseases

coding for bone metastases in electronic health records using technology-enabled abstraction. *Clin Epidemiol* 2015;**7**:441–8. doi:10.2147/CLEP.S92209

330 Friedlin J, Overhage M, Al-Haddad MA, *et al.* Comparing methods for identifying pancreatic cancer patients using electronic data sources. *AMIA Annu Symp Proc* 2010;**2010**:237–41.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041435&tool=pmcentrez&rendertype=abstract

331 Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* 2014;**abs/1409.0**.http://arxiv.org/abs/1409.0473

332 Vaswani A, Shazeer N, Parmar N, *et al.* Attention Is All You Need. *CoRR* 2017;**abs/1706.0**.http://arxiv.org/abs/1706.03762

333 Luong M-T, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. Published Online First: 17 August 2015.http://arxiv.org/abs/1508.04025 (accessed 19 Nov 2017).

334 Support - PYBOSSA. http://pybossa.com/support/ (accessed 15 Nov 2017).

335 Ye MSC, Fabbri D. Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. *J Biomed Inform* 2018. doi:https://doi.org/10.1016/j.jbi.2018.05.014

336 Google. Google Word2Vec. 2013.https://code.google.com/p/word2vec/)

337 Wilson RL, Rosen PA. Protecting Data through ‘ Perturbation ’ Techniques : The Impact on Knowledge Discovery in Databases. *Group* Published Online First: 2003. doi:10.4018/jdm.2003040102

338 Kargupta H, Datta S, Wang Q, *et al.* Random-data perturbation techniques and privacy-preserving data mining. *Knowl Inf Syst* Published Online First: 2005. doi:10.1007/s10115-004-0173-6

339 Zheng S, Song Y, Leung T, *et al.* Improving the Robustness of Deep Neural Networks via Stability Training. *CoRR* 2016;**abs/1604.0**.http://arxiv.org/abs/1604.04326

340 Miyato T, Maeda SI, Ishii S, *et al.* Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. IEEE Trans. Pattern Anal. Mach. Intell. 2018. doi:10.1109/TPAMI.2018.2858821

341 Lingren T, Deleger L, Molnar K, *et al.* Evaluating the impact of pre-annotation on annotation speed and potential bias: Natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J Am Med Informatics Assoc* 2014;**21**:406–13. doi:10.1136/amiajnl-2013-001837