Improving Clinical Decision Making With Natural Language Processing And Machine Learning

by

Alexander William Forsyth

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

Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Author .....
Department of Electrical Engineering and Computer Science
May 26 2017

Certified by .....
Regina Barzilay
Professor
Thesis Supervisor

Accepted by .....
Christopher Terman
Chairman, Masters of Engineering Thesis Committee

Improving Clinical Decision Making With Natural Language Processing And Machine Learning

by

Alexander William Forsyth

Submitted to the Department of Electrical Engineering and Computer Science
on May 26 2017, in Partial Fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

This thesis focused on two tasks of applying natural language processing (NLP) and machine learning to electronic health records (EHRs) to improve clinical decision making. The first task was to predict cardiac resynchronization therapy (CRT) outcomes with better precision than the current physician guidelines for recommending the procedure. We combined NLP features from free-text physician notes with structured data to train a supervised classifier to predict CRT outcomes. While our results gave a slight improvement over the current baseline, we were not able to predict CRT outcome with both high precision and high recall. These results limit the clinical applicability of our model, and reinforce previous work, which also could not find accurate predictors of CRT response. The second task in this thesis was to extract breast cancer patient symptoms during chemotherapy from free-text physician notes. We manually annotated about 10,000 sentences, and trained a conditional random field (CRF) model to predict whether a word indicated a symptom (positive label), specifically indicated the absence of a symptom (negative label), or was neutral. Our final model achieved 0.66, 1.00, and 0.77 F1 scores for predicting positive, neutral, and negative labels respectively. While the F1 scores for positive and negative labels are not extremely high, with the current performance, our model could be applied, for example, to gather better statistics about what symptoms breast cancer patients experience during chemotherapy and at what time points during treatment they experience these symptoms.

Thesis Supervisor: Regina Barzilay
Title: Professor

Acknowledgments

I would like to thank Charlotta Lindvall for her help throughout the entire project, from idea formulation, to problem design, to the implementation and specific details of this thesis. Without her medical knowledge and without her dedication to meeting frequently with me, answering all of my questions, and pushing me to work harder, this thesis would not have been possible. I would also like to thank Regina Barzilay for her support throughout this project. Again, without her expertise and help with the technical aspects of the thesis, none of this would have been possible. Finally, I would like to thank my friends at MIT as well as my family for providing support throughout this thesis as well as a much needed escape from work from time to time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis outline

This first chapter provides a brief motivation for using natural language processing (NLP) and machine learning (ML) algorithms with electronic health records. This chapter then introduces the clinical problems addressed in the thesis and the relevant medical background.

Chapter two focuses on the task of predicting cardiac resynchronization therapy (CRT) outcomes. Chapter two provides background on the NLP and ML algorithms used, and describes the methodology and results we obtained.

Chapter three focuses on the task of extracting breast cancer chemotherapy side effects from free text physician notes. As in chapter two, chapter three first provides background on the NLP and ML algorithms used, and then describes our methodology and results.

Chapter four highlights our primary conclusions about both of the described tasks, and then discusses next steps both in improving our work, and in putting this work to clinical use.

1.2 Motivations for natural language processing and machine learning with electronic health records

The amount of data in electronic health records (EHRs) is already massive, and is expanding rapidly. However, physicians today make most clinical decisions without significant analysis of these health records besides manual review. EHRs can contain both structured data (e.g. laboratory values and demographics) as well as unstructured data (e.g. free-text physician notes). The volume of available data means that manual review prior to even critical clinical decisions is either inefficient or impossible. Additionally, a significant portion of the data is unstructured, often in the form of free-text physician notes. Thus, natural language processing (NLP) algorithms in combination with machine learning could be used to improve clinical decision making by efficiently making conclusions from electronic health records.

1.2.1 Electronic health records background

EHR use has grown rapidly over the past decade. A 2016 brief by the American Hospital Association showed that at least basic EHR use has increased from 9.4% of hospitals participating in 2008 to 83.8% of hospitals participating in 2015 [14]. A significant cause for this increase has been funding from the Health Information Technology for Economic and Clinical Health Act of 2009. EHRs contain a variety of information constituting a patient's full medical history. Despite this wealth of information, physicians primarily view EHRs as a byproduct of healthcare instead of as an opportunity to improve clinical decision making [34].

Currently, there are three primary barriers to fully utilizing EHRs. First is patient privacy and security concerns; second is the fragmented and incompatible nature of the variety of EHR platforms used today; and third is the data itself, much of which is unstructured free-text [29, 34]. To varying degrees, all of these barriers are relevant to this thesis. The issue of patient privacy can significantly slow down research efforts as gaining access to protected health data takes time through an often lengthy approval

process. While this issue is contentiously debated both in academic literature and in public policy, and while both gaining access to protected health data and securely managing it were crucial to this thesis, discussing this issue is not a major focus of the thesis itself. The second issue, that of a variety of often incompatible systems used to store EHRs, is also not the primary focus of this thesis. However, this issue is very relevant to the implementation of our methods because specific data management frameworks had to be implemented for each of the two datasets we used. The third issue, that most of the raw data in EHRs is free-text, is the primary focus of this thesis, and is discussed further in the next section.

1.2.2 Review of natural language processing with electronic health records

Natural language processing, often in combination with machine learning has been used on EHRs for a variety of tasks in recent years. The task most relevant to this thesis is clinical decision support (CDS) as well as the subcategory of CDS involving information extraction. Clinical decision support, which uses EHRs to aid physicians in clinical decision making, often uses NLP to extract features from free-text notes that can be used by some machine learning model. Information extraction involves extracting pre-defined types of information from unstructured free-text.

Computers have been used to aid in clinical decision making since at least 1987 when the goal of clinical decision support was defined as to "help health professionals make clinical decisions, deal with medical data about patients or with the knowledge of medicine necessary to interpret such data" [8]. Some examples of CDS systems include tools for focusing attention (e.g. flagging abnormal values out of a large list), tools for providing patient-specific recommendations based on that patient's medical history, and tools for information management, which can include information extraction [24]. Some specific examples of recent information extraction tasks include: 1) automatic assignment of ICD-9-CM codes to clinical free text [26], 2) de-identification of discharge summaries within the i2b2 initiative [37], and 3) patient smoking status

discovery from discharge summaries also within the i2b2 initiative [36].

1.3 Heart failure and cardiac resynchronization therapy

Chronic heart failure (HF) is one of the leading causes of death in the US; it affects nearly 6 million patients and contributes to nearly 300,000 deaths per year, being listed as the primary cause of death for at least 68,000 [23]. Cardiac resynchronization therapy (CRT), an implantable therapy for heart failure developed more than a decade ago, has shown to be an effective therapy for select HF patients [4]. While CRT is effective in the majority of cases, an estimated one-third of treatments are ineffective without a well-understood cause [5]. Thus, our first goal in this thesis was to apply machine learning and NLP techniques to more precisely predict CRT non-response.

1.3.1 Heart failure and CRT background

The primary pathophysiology in HF is impaired myocardial contractility leading to ventricular hypertrophy and dilation leading to disordered electromechanical coupling [4]. Ventricular dyssynchrony reduces myocardial efficiency and cardiac output, worsening cardiac performance [4]. CRT resynchronizes electromechanical coupling by placing leads to the right and left ventricles to excite the right ventricle and left ventricular (LV) lateral wall simultaneously [4]. It has been shown that CRT can improve chronic LV systolic dysfunction and lead to reduction in chamber size as well as improvement in LV ejection fraction (LVEF) [4, 19, 32, 22, 39].

The current clinical guidelines for prescribing CRT are displayed in the decision tree in figure 1-1. The guidelines make a recommendation based on 7 concrete variables: New York Heart Association (NYHA) class, LVEF, QRS duration, left bundle branch block (LBBB), sinus rhythm, ischemic cardiomyopathy and the presence of comorbidities [35]. In addition to not considering variables besides these seven, the current guidelines leave significant room for interpretation that varies from hospital

to hospital and even from doctor to doctor with recommendations as vague as "CRT might be reasonable" [35].

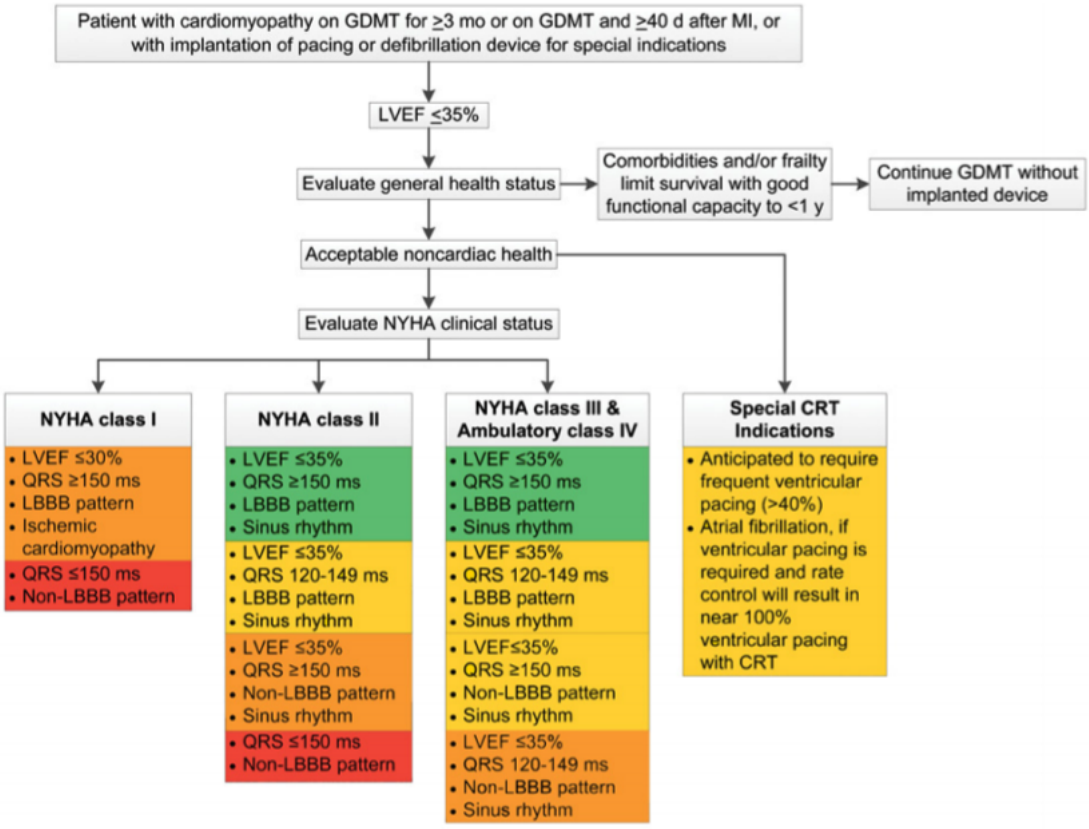

Figure 1-1: The current clinical guidelines for CRT eligibility. The recommendation for treatment is determined by color with green meaning "CRT is recommended", yellow meaning "CRT is reasonable", orange meaning "CRT might be reasonable", and red meaning "CRT is not recommended" [35].

Prior research has determined the change in LVEF from before CRT until a fixed amount of time (typically 6-18 months) post-procedure to be a reasonable metric of success for CRT [10]. In general, patients whose LVEF significantly increases post-procedure are considered "responders", and those whose LVEF decreases are considered "non-responders". However, the exact cutoff LVEF increase for which CRT is considered a success is not well defined, and the choice for a cutoff point might vary according to cardiologist recommendation when attempting to define a binary success versus non-success of CRT, such as in the context of a machine learning problem.

1.3.2 Predicting CRT response

Between 2002 and 2010, there were an average of 41,578 CRT procedures per year in the United States alone, and that number has likely increased in recent years [31]. While specifically defining "non-responders" to CRT is not clear cut, an estimated one-third of CRT procedures are considered ineffective [5]. In addition to having worse clinical outcomes in terms of pump failure and death, CRT nonresponse leads to procedural risks and costs that may be preventable with improved patient selection. The cumulative cost of CRT was calculated in 2005 as between $60,000 and $82,000 [9]. Thus, an estimated $800 million to $1.1 billion is spent in the United States per year on ineffective CRT procedures. Therefore, predicting CRT non-response more precisely than the current decision tree guideline could save significant amounts of money and could lead to better outcomes and better quality of life for patients for whom CRT would be ineffective.

Previous work has sought to identify individual predictors of CRT non-response, but no additional variables have been added to the current decision tree guideline, and no combination of predictors has been identified that predicts CRT non-response with high precision. One study sought to predict CRT response with echocardiographic parameters besides LVEF, but did not find any significant predictors [6]. Another study found no pre-procedure predictors of CRT response from a set of variables including NYHA class, QRS duration, LVEF, age, gender, duration of HF symptoms, and many others [17]. Unlike in the previous results, another study found that both QRS duration and LVEF were significant ($p < 0.05$) predictors of short-term CRT response [18]. However, one of the largest studies of CRT response, the 2008 Predictors of Response to CRT Trial (PROSPECT), a fifty-three center prospective study, found that no single echocardiographic measure of dyssynchrony could be recommended to improve patient selection for CRT [7].

1.4 Patient symptoms during breast cancer chemotherapy

In 2016 in the United States alone, there were about 250,000 cases of and about 40,000 deaths related to female breast cancer [30]. About 36% of early stage (stages I and II) and about 80% of stage III breast cancer patients receive some type of chemotherapy [21]. Chemotherapy side effects are frequently debilitating and can be deadly. Despite this, chemotherapy side effects are not always well documented outside of results from clinical trials. Drug side effects and patient symptoms in general are most often not documented in a structured format in medical records. Instead, symptoms and side effects are documented in physician free-text notes. Extracting symptoms from these free-text notes would have a number of useful clinical applications including: providing more documentation than just clinical trials of side effects to chemotherapy drugs and enabling further clinical decision support systems to be built to assess the risk of side effects and to warn patients and physicians of potentially deadly side effects.

1.4.1 Breast cancer chemotherapy background

The number of chemotherapy regimens for non-metastatic breast cancer (stages I-III) is limited (relative to the number for metastatic cancer), and thus, the side effects of these drugs are easier to study. The most common chemotherapy regimens include some combination of the following eight drugs: doxorubicin (Adriamycin), docetaxel (Taxotere), paclitaxel (Taxol), cyclophosphamide (Cytoxan), trastuzumab (Herceptin), methotrexate (Trexall), and fluorouracil (5-FU/Adrucil). Some common side effects for these drugs include: nausea, vomiting, hair loss, mucositis, and fatigue. More serious side effects for these drugs can include: difficulty breathing, difficulty swallowing, seizures, and even death.

A 2006 study of 12,239 women under age 63 with breast cancer found that "chemotherapy-related serious adverse effects might be more common than reported

by large clinical trials and lead to more patient suffering and health care expenditures than previously estimated" [12]. In the same study, 10.5% of patients had a hospitalization or emergency room visit related to a chemotherapy side effect, and the average cost per patient experiencing a serious side effect was about $15,000 [12].

For most chemotherapy drugs, knowledge of side effects as well as the frequency and severity of these side effects is obtained primarily through clinical trials. During clinical trials, chemotherapy adverse effects are assessed and reported by clinicians. However, a 2004 study found that physician reporting was neither sensitive nor specific in detecting common chemotherapy side effects [11]. A 2001 study of a variety of drug trials found that severity of clinical adverse effects was adequately defined in only 39% of trial reports [13]. Additionally, a review of drug randomized controlled trials found that external validity of the trials was usually inadequate, meaning that results such as drug-related adverse effects found in trials might not accurately reflect results found in the general public [28].

1.4.2 Information extraction

Since chemotherapy side effects and patient symptoms in general are presented primarily in physician free text notes, they must be extracted from the free text in order to create structured data suitable, for example, to determine the frequency of adverse side effects for each chemotherapy drug. Extracting these symptoms manually is a time consuming task, which usually must be performed by physicians or trained professionals, thus making this task inefficient to impossible, depending on the volume of data. Information extraction involves automatically extracting pre-defined structured information from unstructured data using natural language processing. Thus, information extraction could be used to efficiently extract patient symptoms including drug side effects from physician notes.

There are essentially two approaches to information extraction. First is a rule-based approach, in which a pre-designed set of rules is used to extract information. E.g. the phrase "patient reports having X" could be used to extract all reports of symptoms, denoted by X, that match that exact structure in the free text. This

approach works well when a relatively small set of rules can extract most of the desired information (e.g. when the data of interest is in a semi-structured format in the free text). However, this approach does not work well when the desired information can appear in a large variety of contexts within the free text; in this case, the set of rules required to extract the information of interest would be infeasibly large. Physicians describe patient symptoms in a highly variable manner, and thus, a simple rule-based approach would not be sufficient for this task.

The second approach to information extraction is to use machine learning. Instead of using a pre-defined set of rules, supervised ML algorithms are used to learn patterns from a labeled set of free text notes. Machine learning-based information extraction systems have been used, for example, to extract tumor characteristics from breast pathology reports [38], to identify discontinued medications from free text [3], and to extract events and their temporal relation from free text notes [33]. An ML-based approach usually requires a manually annotated dataset, which requires significant initial investment. However, an ML-based approach can be much more generalizable than a rule-based approach, and can work well for tasks where the set of extraction rules is very large, is unknown, or both. Thus, in this thesis, we apply an ML-based approach to extracting patient symptoms from free text physician notes.

Chapter 2

Prediction of cardiac resynchronization therapy outcomes

The first task addressed in this thesis and the focus of this chapter is predicting the outcome of cardiac resynchronization therapy. We used NLP techniques to extract features from physician free-text notes, and used these features in addition to structured data values to train a supervised machine learning classifier. The trained classifier slightly outperformed the currently used decision tree-based clinical guideline in predicting non-responders to CRT. The relatively meager performance gain compared to the current clinical guideline suggests that CRT non-response prediction is an inherently difficult task, and more work would be required to build a hospital-ready clinical decision support system.

2.1 Feature extraction with natural language processing background

2.1.1 Bag-of-words, n-grams, and tf-idf

Bag-of-words is the primary NLP method we used to extract features from free-text physician notes. The most simple unigram bag-of-words model keeps track of the frequency of each word out of the set of possible words, the vocabulary. However,

this model disregards other linguistic structures such as grammar and word order. A slightly more complex n-gram bag-of-words model keeps track of the frequency of each sequence of words of length n, and can represent some of the spatial information in the text that is lost with a unigram model. Larger values of n can represent more spatial context, but require larger training corpora in order to be useful because the probability of each length n sequence decreases significantly. Because of this tradeoff, small values of n, often 2 or 3, are generally used in practice.

There are a variety of methods for keeping track of word sequence frequencies. Two of the most commonly used methods, both of which we tested in this thesis, are term frequency and term frequency-inverse document frequency (tf-idf). Term frequency is the simplest method and uses just the raw count of each word sequence in the text. For example, if we are using a unigram model with vocabulary ["this", "that", "is", "a", "good", "bad", "thesis"], and our text is "this is a thesis that is good", then our sentence would be represented by the term frequency vector [1, 1, 2, 1, 1, 0, 1]. The second method, tf-idf, weights a word's term frequency by a factor inversely proportional to the number of documents in the entire corpus in which that word appears, adjusting for the fact that some words are used more frequently than others. There are a few methods for calculating tf-idf, but the equations we used are:

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t) \tag{2.1}$$

$$idf(t) = \log \frac{1+n_d}{1+df(d,t)} + 1 \tag{2.2}$$

where $tf(t,d)$ is the term frequency of term t in document d, n_d is the total number of documents, and $df(d,t)$ is the number of documents that contain term t.

2.1.2 Word embeddings

A common problem with the bag-of-words n-gram model is the "curse of dimensionality". The English vocabulary contains hundreds of thousands of words, and thus there are tens to hundreds of billions just of possible bigrams (length 2 n-grams). A specific corpus might contain only a subset of these word sequences, but the number

of observed bigrams in real-word datasets is frequently in the millions or billions. As in many tasks dealing with high dimensionality, we apply a form of dimensionality reduction to transform our very high dimensional data into a lower dimensional space. The technique we apply in this thesis is word embeddings.

Word embedding involves mapping n-grams from a space with one dimension per n-gram in the vocabulary to a continuous vector space with much lower dimension (often in the hundreds). Word embeddings began in 2003 with what Bengio et. al. called "a distributed representation for words which allows each training sentence to inform the model about an exponential number of semantically neighboring sentences" [1].

While there are many software packages available today for word embeddings, one of the most popular, word2vec, was created at Google by Mikolov et. al. in 2013 [20]. Many models for word embeddings today, including word2vec, use neural networks. In this thesis, we use an implementation of the word2vec continuous bag-of-words model in Python packaged with Gensim [27].

2.2 Machine learning classifiers

We trained a supervised ML classifier to predict CRT outcomes. In this thesis, we compared the performance of logistic regression (LR), support vector machine (SVM), and random forest (RF) classifiers. We did not attempt to train a neural model because of the relatively small amount of training examples. We used the implementations of these classifiers in Python, scikit-learn [25].

2.3 Dataset and feature extraction methodology

2.3.1 Dataset

The electronic health record dataset came from the Research Patient Data Registry (RPDR) from Partners hospitals in the Boston area. The dataset includes the full medical history of 2641 patients who received CRT implants between 2003 and 2015.

Included in the dataset is structured data (e.g. billing codes, demographics, medications, diagnoses, etc.) and unstructured data (i.e. physician notes including cardiology reports, radiology reports, etc.).

2.3.2 Cohort and primary outcome

Our primary outcome in determining CRT success was the change in LVEF from a baseline level to a post-procedure level. The baseline level was the value measured closest to the procedure date and within 60 days before CRT. The post-procedure level was the value measured closest to one year and between 6 months and 18 months after CRT. A negative ΔLVEF was considered CRT failure, and a positive ΔLVEF was considered CRT success. The patient dying within 18 months after CRT was considered CRT failure regardless of ΔLVEF. We note that because we looked for primary outcomes up to 18 months after procedure, patients were excluded from analysis who received CRT within 18 months of the last date in the available RPDR dataset.

1651 of the 2641 patients receiving CRT were missing either a baseline or followup LVEF in the acceptable time ranges, or received CRT with 18 months of the end of the dataset's time window, and were thus excluded from further study. Of the 990 patients included in analysis, 249 died, 343 had negative ΔLVEF, and 428 had positive ΔLVEF. The characteristics for these 990 patients at the time of CRT implant are shown in table 2.1.

2.3.3 Structured feature and outcome extraction

The complete list of structured features we used is available in table 2.2. Some of these features, such as demographic information and procedure billing codes, were readily available as structured fields in the RPDR dataset. Other features, including six of the features in the clinical guideline decision tree: LVEF, LBBB pattern, sinus rhythm, QRS, NYHA class, and ischemic cardiomyopathy, were not available as structured data fields, and were extracted via regular expression from cardiology reports and

Table 2.1: Patient characteristics at the time of CRT implant (n=990).

Demographics	
Age, mean (SD)	71.6 (11.8)
Female sex, %	21.9
Non-Hispanic white, %	87.2
Medical history, %	
Non-ischemic heart failure	80.2
Coronary artery disease	74.6
Left bundle branch block	37.8
Ventricular arrhythmia	60.3
Atrial fibrillation	50.8
Chronic kidney disease	24.4
Diabetes mellitus	32.7
Diagnostic studies, mean (SD)	
LVEF, %	24.8 (7.69)
Creatinin, mg/dL	1.69 (1.17)
Sodium, mg/L	137.4 (3.9)
Hemoglobin, (mg/dL)	12.4 (2.0)
Medications, %	
Beta-blocker	92.3
ACE/ARB	77.3

cardiology notes.

We note that our primary outcome, LVEF, is one of the regular expression extracted values. To ensure accuracy in our outcome variable, a physician manually validated the accuracy of LVEF extraction, and found that nearly all extracted values were correct. The incorrectly extracted values were corrected. Because manual verification was necessary, extraction via regular expression was used more so to quickly locate the LVEF values in cardiology reports than to automatically extract values without supervision.

2.3.4 NLP feature extraction

Free-text features were used from both cardiology notes and longitudinal medical record notes (most other physician notes). Bigram word embeddings were trained using all notes both in this CRT dataset and the other RPDR dataset used in the

Table 2.2: Longitudinal EHR data utilized in machine learning algorithms.

Data source	Model features
Demographics	Sex, age at implant, race
Billing codes	ICD-9-CM, CPT
Encounter data	Visit type, length of stay
Laboratory reports	Lab values; most recent and trend
Medication list	Medication and drug class
Cardiology reports	LVEF, QRS, LBBB pattern, Sinus rhythm; most recent and trend

next chapter for symptom extraction. We did not use pre-trained word embeddings or train on a public dataset because we wanted medical specific word embeddings. A general public dataset would likely have been missing many of the words used frequently in medical texts.

Common techniques to improve the quality of these NLP features were used. Some of these techniques include: removing stop words, removing low frequency bigrams, lowercasing all words, and removing numbers and special characters. While we tested both term frequency and tf-idf methods, in the final model, bigram tf-idf word embeddings were used.

2.4 Training and tuning classifiers

2.4.1 Evaluation

Because the goal of this task was to predict non-responders to CRT, non-responders were considered to be the "positive" outcome, and responders were considered to be the "negative" outcome in all of our statistics below. In addition to the definitions of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) in table 2.3, the following statistics are relevant to evaluation:

$$precision = \frac{TP}{TP + FP} \tag{2.3}$$

$$recall = \frac{TP}{TP + FN} \tag{2.4}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision + recall} \tag{2.6}$$

where β in the F_β score is a parameter chosen by a user who places β times as much importance on recall as on precision. We chose $\beta = 0.1$, meaning we valued precision about 10 times as much as recall.

Table 2.3: CRT outcome prediction confusion matrix defintions.

	Predicted non-responder	**Predicted responder**
Actual non-responder	True Positive (TP)	False Negative (FN)
Actual responder	False Positive (FP)	True Negative (TN)

2.4.2 Train, validation, and test splits

Train, validation, and test splits of the 990 patients in the final cohort were used in order to train, select, and finally test the performance of our final models. 200 patients were initially held out from the 990 for the test set. For each potential model, random splits of 67% and 33% of the remaining 790 patients were used for training and validation respectively.

For each of the three classifiers tested (LR, SVM, and RF), a hyperparameter search was performed and the model with the best F_β score on 3 cross validations was selected. We trained one model including NLP features and one model only using structured features for comparison. Once the final models had been selected, they were run on the test dataset in order to obtain final performance metrics.

2.5 Results and discussion

Performances of the final models are reported in table 2.4. Because of the relatively low chosen β value, the best performing models had high precision, but relatively low recall. We note that the model including NLP features performed better than the model including only structured fields in all three of precision, recall, and accuracy. While precision was high for both models, the structured features only model accuracy was only marginally better than the physician recommendation accuracy, and the structured and NLP features model improved that accuracy only slightly.

Table 2.4: CRT outcome prediction classifier results on the test set (*note that the physician recommendation is always "should receive CRT" because our dataset includes only patients who received CRT; thus, the only statistic of interest for this model is accuracy).

Model	$\mathbf{F}_\beta$	**Precision**	**Recall**	**Accuracy**
Physician recommendation*	0.0	0.0	0.0	0.58
Structured features only	0.64	0.75	0.04	0.59
Structured + NLP features	0.83	0.89	0.11	0.64

Predicting non-responders with high precision means that physicians could use the model to flag likely non-responders after initially selecting a patient for CRT. The doctor could then either more extensively review a patient's medical history in order to determine whether CRT is the best option, or she could choose not to recommend CRT at that time. However, the relatively low recall achieved by both of our models means that applying these models in a clinical setting would require additional work. The accuracy of our models offered a negligible improvement over the physician recommendation. These results, like many previous studies on CRT response predictors, suggest that predicting CRT non-response is a difficult task, and that predicting CRT non-response with both high precision and high recall has not yet been achieved.

Chapter 3

Breast cancer symptom extraction

The second task addressed in this thesis and the focus of this chapter is extracting physician indicated patient symptoms from free-text notes. Specifically, we sought to extract patient symptoms during and recently after chemotherapy including side effects of chemotherapy drugs. We manually annotated a dataset of approximately 10,000 sentences and trained a machine learning model to extract patient symptoms from the free-text with high precision and recall. These results have a number of interesting applications including enabling further clinical decision support systems to be built such as predicting whether patient symptoms might be life-threatening, and these results could yield better understanding of all side effects of chemotherapy drugs and the relative frequencies of these side effects.

3.1 Problem formulation and machine learning background

3.1.1 Problem formulation

As discussed in the introduction, there are two primary approaches to information extraction: rule-based approaches and machine learning based approaches. In this thesis, we apply an ML-based approach. We formulate our problem as tagging each word in the dataset with a tag: "positive", "neutral", or "negative". A "positive"

word indicates the presence of a symptom; a "neutral" word is unrelated to a symptom; and, a "negative" word specifically indicates the absence of a symptom. E.g. in the sentence: "patient describes having nausea, vomiting, and pain", the tags for that sequence of words would be ["neutral", "neutral", "neutral", "positive", "positive", "neutral", "positive"]. And, in the sentence: "patient denies having nausea", the tags would be ["neutral", "neutral", "neutral", "negative"]. We note that in the actual text, while most word tags would be clear to a human labeler, there are sentences where the correct tags are not clear even to physicians manually annotating the sentence. Examples of such uncertainty include sentences describing historical symptoms (e.g. symptoms a year ago), sentences where a doctor explains potential side effects of a drug to a patient, and sentences including a long "laundry list" of symptoms that the patient may have had at some point, but which might have been copied and pasted from a previous note.

3.1.2 Conditional random field model

To predict individual word tags, we use conditional random fields (CRFs). We chose CRFs instead of individual classifiers for each word because CRFs predict a sequence of labels given a sequence of input, incorporating structural information about the sequence. In our task, the probability a given word indicates a symptom is dependent not only on the word itself, but on previous words. E.g. in the partial sequence "patient denies having", we know that the next word's tag is likely "negative" even before seeing the next word itself.

CRFs were introduced by Lafferty, McCallum, and Pereira in 2001 as an alternative to hidden Markov models (HMMs) and maximum entropy Markov models (MEMMs) with NLP-specific applications in mind [16]. Generative models such as HMMs must define a joint probability over observation and label sequences, which is intractable for many sequence labeling NLP tasks. HMMs also make strong independence assumptions that are often not valid for real data. Discriminative models such as MEMMs are more tractable, but MEMMs can often "ignore" observations in favor of a transition between states that it believes is particularly likely. CRFs

were designed as an alternative to both of these models specifically for NLP sequence labelling tasks such as part-of-speech tagging. The differences between CRFs, HMMs, and MEMMs are displayed graphically in figure 3-1.

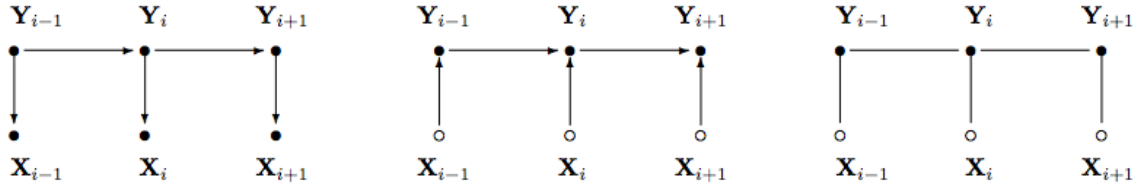

Figure 3-1: Graphical structures of simple HMMs (left), MEMMs (center), and chain structured CRFs (right) for sequence label prediction. An outlined circle indicates that a variable is not generated by the model, while a solid circle is generated by the model [16].

Lafferty, McCallum, and Pereira define a CRF on observation variables $\mathbf{X}$ (e.g. words and their additional features) and random variables $\mathbf{Y}$ (e.g. part-of-speech word tags):

> Let $G = (V, E)$ be a graph such that $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$, so that $\mathbf{Y}$ is indexed by the vertices of G. Then $(\mathbf{X}, \mathbf{Y})$ is a conditional random field in case, when conditioned on $\mathbf{X}$, the random variables $\mathbf{Y}_v$ obey the Markov property with respect to the graph: $p(\mathbf{Y}_v|\mathbf{X}, \mathbf{Y}_w, w \neq v) = p(\mathbf{Y}_v|\mathbf{X}, \mathbf{Y}_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G [16].

3.2 Dataset and methods

3.2.1 Dataset and cohort

Our dataset for this task came partially from the Research Patient Data Registry (RPDR) from Partners hospitals in the Boston area and partially from Dana-Farber Cancer Institute (DFCI). The dataset includes all physician notes from 4732 stage I-III breast cancer patients who received at least one of the following chemotherapy drugs between May 1996

and May 2015: doxorubicin, docetaxel, paclitaxel, cyclophosphamide, trastuzumab, pertuzumab, methotrexate, and fluorouracil. Demographic information as well as the percentage of patients who received each chemotherapy drug is available in table 3.1.

Table 3.1: Breast cancer patient characteristics (n=4732). * indicates that data was available only for RPDR dataset patients (n=311).

Demographics	
Age, mean (SD)	51.37 (11.16)
Non-Hispanic white*, %	79.7
Chemotherapy drugs, %	
Doxorubicin	74.0
Docetaxel	12.3
Paclitaxel	57.0
Cyclophosphamide	82.4
Trastuzumab	23.9
Pertuzumab	1.8
Methotrexate	2.2
Fluorouracil	3.6

While our goal was to extract side effects of chemotherapy drugs from free text notes, it is not clearly feasible even for an oncologist to distinguish between a general patient symptom and a chemotherapy drug side effect. Distinguishing side effects is difficult first because all possible side effects for a drug are not always known, and often the only reliable source of side effects is clinical trials. Second, this task is difficult because even the attending physician might not be able to definitively attribute a symptom, which can be a side effect of a chemotherapy drug, to the drug itself and not to another source. To try to identify chemotherapy side effects, we looked only at notes in our dataset that were dated between the start of chemotherapy treatment and 90 days after the end of treatment. However, for the notes in this time range, our goal was to label every symptom in the notes, not just symptoms that were definitively chemotherapy drug side effects.

Physicians document symptoms primarily in discharge summaries and longitudinal notes in the RPDR dataset and in clinical notes in the DFCI dataset. These free text notes were broken up into individual sentences using the Punkt Sentence Tokenizer, part of the Natural Language Toolkit (nltk) for Python [2, 15]. There were a total of 264,881 sentences between the start of chemotherapy treatment and 90 days after the end of treatment across all chemotherapy drugs. The number of sentences for each drug is available in table 3.2.

Table 3.2: Number of sentences per drug in discharge summaries and longitudinal notes between the start of chemotherapy treatment and 90 days after the end of treatment.

Drug	**# Sentences**
Doxorubicin	39,661
Docetaxel	18,897
Paclitaxel	103,564
Cyclophosphamide	37,874
Trastuzumab	35,921
Methotrexate	3675
Fluorouracil	25,289

3.2.2 Features and labels

Approximately 10,000 sentences were manually labeled by two physicians. Each word was labeled as one of "negative", "neutral", or "positive". Since we were able to manually label only a fraction of the total available sentences, we chose to label sentences from a single chemotherapy drug to maximize consistency of the symptoms. For this, we chose paclitaxel because it had the most sentences of any drug (as shown in table 3.2). Thus, we could label more than 90,000 additional sentences for paclitaxel alone if we wanted to expand the labeled dataset in the future. The number of "negative", "neutral", and "positive" labels in the labeled dataset is in table 3.3.

Table 3.3: Number of each label in the symptom extraction dataset.

Negative	1191
Neutral	196,512
Positive	1909

For each word in a sentence, we used the following features: a bias term, the lowercased word, the last three characters of the word, the last two characters of the word, whether the word was uppercase, whether the word was title-cased, and whether the word was a number. For each word, we also included the lowercased previous word, whether the previous word was uppercase, and whether the previous word was title-cased, or a "beginning of sentence" tag instead if the word was at the beginning of the sentence. Similarly, we included the same features about the next word in the sentence, or an "end of sentence" tag if the word was at the end of the sentence.

3.2.3 Classifier training and tuning

We held out 20% of the labeled dataset for the test set. Of the remaining 80% of the labeled dataset, we used 80% as training data and 20% as validation data to select CRF hyperparameters and to select the final features listed in the previous section. Once the final model had been selected, we trained a CRF model using the entire 80% of the labeled dataset used for training and validation, and tested performance on the held out 20%. A flow chart of all of our methods is available in figure 3-2.

3.3 Results and discussion

The performance of our final model on the held out test dataset is given in table 3.4. Precision, recall, and F1 scores are given individually for each

of the labels (negative, neutral, and positive). The overall accuracy was 99.2%. We note that as shown in table 3.3, the vast majority of labels were "neutral", and thus the neutral label has very high precision, recall, and F1 score, and our overall accuracy is very high. We are primarily interested in the precision/recall for the positive and negative labels. Note that some examples of mispredictions are given in table A.1.

Table 3.4: CRF model performance on held out test sentences.

	Precision	**Recall**	**F1**
Negative	0.86	0.69	0.77
Neutral	0.99	1.00	1.00
Positive	0.82	0.56	0.66

We also present a learning curve analysis in figure 3-3, which shows how much our model's F1 score improves from having more training data. As expected, we see that the positive and negative label F1 scores increase significantly from the first few hundreds of training sentences, and then increase more slowly as additional data is added. The neutral label F1 score is always near 1.0, but has a visible increase in F1 score over the first few hundred sentences as well. And finally, we see that the F1 score for negative labels becomes distinctly higher than the F1 score for positive labels after the first 5000 or so training sentences. We hypothesize that negative labels are generally easier for the model to extract from the text because physicians document negative symptoms in more specific terms than they document positive symptoms (e.g. by using phrases such as "patient denies having").

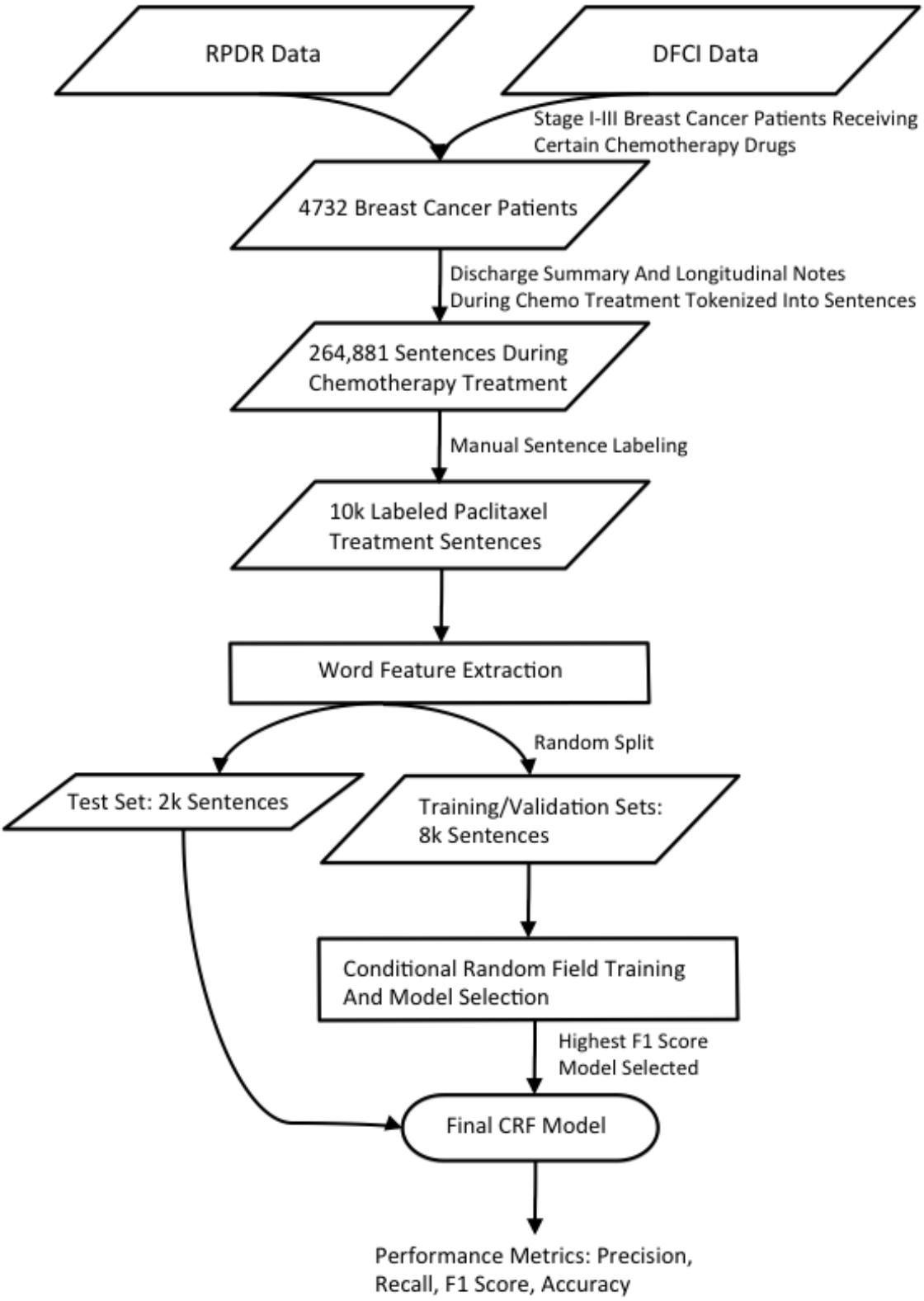

Figure 3-2: Flow chart of symptom extraction methods.

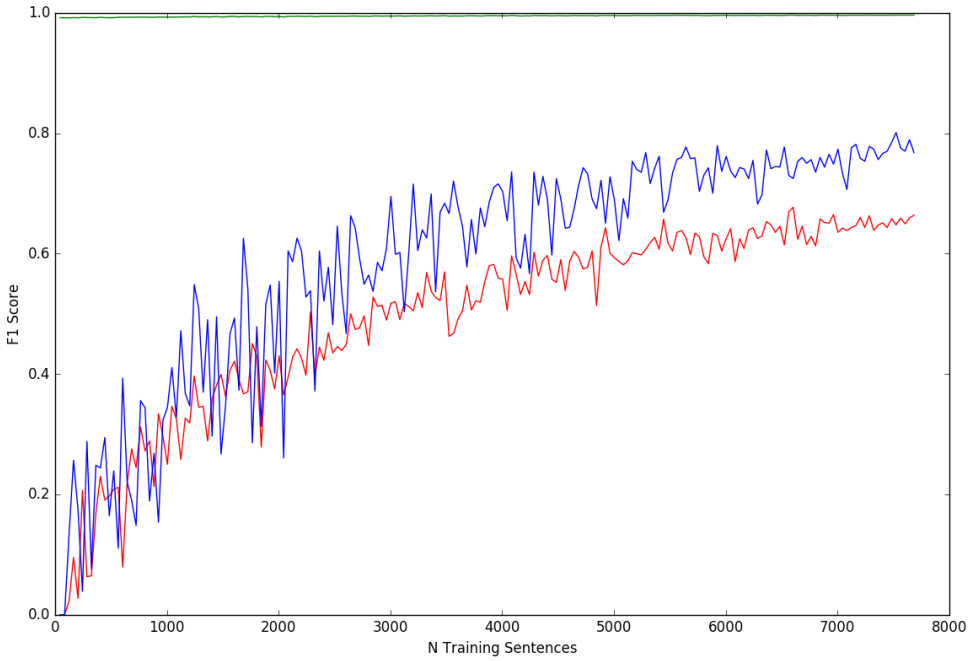

Figure 3-3: CRF learning curve analysis for positive (red), neutral (green), and negative (blue) labels.

Chapter 4

Conclusions and Future Work

In this chapter, we discuss the implications of our results for both the CRT outcome prediction task and the breast cancer symptom extraction task. We then discuss the limitations of our current results, what additional work might be needed to put our methods to clinical use, and finally, we discuss potential future applications of our work.

4.1 CRT outcome prediction conclusions and limitations

While we were able to slightly improve accuracy of CRT non-response prediction compared to the current baseline, our achieved recall is likely not high enough to be useful to physicians. Most prior studies, including randomized controlled trials, to identify individual predictors of CRT response have not identified any significant predictors [7, 6, 17]. Our results further suggest that predicting CRT response is a difficult task.

Beyond limitations due to our model's performance, this work has limitations due to the dataset we used. First, our model was designed to predict CRT non-responders from the set of patients who were already

recommended for CRT by their physicians. This model could not be used, for example, to predict which heart failure patients should receive CRT. Second, we used only RPDR data from Partners hospitals in the Boston area. Thus, our model might not be fully generalizable to other hospitals' EHRs. Even applying our model to other EHR systems would require additional effort to work with different data storage methods.

4.2 CRT outcome prediction future work

In most machine learning tasks, having more training data points is almost always better. In this task, we had only 990 patients. Thus, better performance might be achievable with more patients. At an average of 41,578 CRT procedures per year, a significantly larger dataset is technically attainable, although that data is likely distributed across a large number of EHR (and non-EHR) systems. Accessing a larger portion of this data could both improve model performance and create a more generalizable model.

If a CRT outcome prediction model were to achieve acceptable performance, doctors recommending CRT to a patient based on the current guidelines could use the model to flag patients who are unlikely to respond to CRT. The model flagging a patient could lead the doctor to manually review that patient's medical history more carefully, and then make the final decision about whether to assign CRT. Preventing even a small number of unnecessary CRT procedures would save significant amounts of money and would lead to better outcomes and quality of life for the patients involved.

4.3 Symptom extraction conclusions and limitations

Our model's performance predicting positive and negative symptom labels was high enough for some useful applications, such as creating a large database of the relative frequencies of symptoms experienced by breast cancer patients during chemotherapy. However, the F1 scores for positive and negative symptoms were not close to 100%. Thus, manual post-processing of the predicted labels would be required for applications that require very accurate labels.

Putting our model's performance into context is not exactly straightforward. There is no similar baseline to compare the results to. Additionally, we do not currently know the "accuracy" of manual labelling. During manual labelling, there were many cases where the annotating physician was unsure of the correct label. We could determine the accuracy of human labelling by having multiple physicians label the same notes and calculating the frequency with which they agree on a label. As of now, we suspect that a significant number of the model's misclassifications stem from the inherent uncertainty of the task and of the manual annotators.

In this thesis, we sought to clearly define what does and does not constitute a symptom. However, in the free-text, there are situations that, for example: 1) clearly indicate a symptom, but do so in a narrative manner, 2) indicate a historical symptom, 3) present a "laundry-list" of symptoms which might have been copied from previous notes, or 4) are clearly discussing a symptom, but are not clearly indicating that the patient has that symptom. Setting more strict guidelines for labelling depending on the target application of the model might yield better performance. E.g. we could label only declaratively stated, clearly present-time symptoms if we wanted a model that might miss narratively stated or historical symp-

toms.

As discussed previously, one limitation of our methods is that we cannot specifically distinguish chemotherapy drug side effects from other patient symptoms. This limitation holds not just for our machine learning model, but for the problem formulation itself; i.e., it is often not possible for a doctor to specifically attribute the cause of a symptom to a certain drug. Thus, in this thesis, we framed our results as extracting symptoms experienced by breast cancer patients during chemotherapy instead of specifically as extracting chemotherapy drug side effects. Another limitation of these results is that we only predict whether or not a word indicates a symptom. We do not predict any notion of what that symptom is. Even common symptoms such as nausea have many synonyms and other ways of being expressed in the free-text. Thus, obtaining, for example, the number of patients who experience nausea during chemotherapy, would require manually labelling whether each predicted symptom indicates nausea.

Finally, as in the CRT outcome prediction task, our dataset for symptom extraction was limited to RPDR data from Partners hospitals in the Boston area and data from DFCI. In fact, most data was from DFCI. Thus, our model might not be generalizable to other hospitals where the ways of documenting symptoms in the free-text could be different.

4.4 Symptom extraction future work

While our learning curve analysis in figure 3-3 showed diminishing returns to F1 score with additional training data, additional labeled data might increase model performance. In this thesis, we labeled only 10,000 sentences because labeling was time consuming and required a trained physician. From the dataset used in this thesis alone, we had 264,881 sentences available to label. A much larger training dataset could allow

the model to learn more of the infrequent ways of expressing symptoms and especially some of the more narrative symptom indications.

There are well-known issues with adverse side effect documentation during clinical trials, including for chemotherapy drugs [11, 13, 28]. We could apply the current model to all of our 264,881 sentences in order to get rough estimates of the frequency with which breast cancer patients experience symptoms during the course of chemotherapy. We could, for example, model the frequency of symptoms over time throughout the course of treatment. And, with additional manual effort to categorize the types of predicted symptoms, we could obtain how frequently breast cancer patients experience each individual symptom during chemotherapy.

The same methods for symptom extraction during chemotherapy could be applied to other tasks. For many other tasks, the only difference would be the manually labeled dataset. Some examples of other tasks include symptom extraction for other diseases and extracting patient goals of care discussions.

Finally, with additional post-processing to verify and correct our model's predicted labels, these methods could be applied to many other tasks requiring highly accurate data. For example, our model could be used to build a patient cohort experiencing some set of symptoms (in cases where a rule-based method of finding patients with those symptoms might be ineffective). Our model's output, predicted positive and negative symptoms, could also be used as features to build additional machine learning models. For example, we could predict the probability of certain outcomes such as hospitalization or death given a patient's current symptoms, and doctors could then take preventative measures early on.

Appendix A

Tables

Table A.1: Example sentences with words mislabeled by the CRF model. Negative labeled words are blue and bold, while positive labeled words are red and italic.

Actual	she denies **fever chills** but has *rigors*
Predicted	she denies **fever chills** but has rigors
Actual	ct dye caused a high *fever* and she has a *rash and blistering*
Predicted	ct dye caused a high *fever* and she has a rash and *blistering*
Actual	she also endorses a *chest heaviness* that is related to the *cough*
Predicted	she also endorses a chest heaviness that is related to the *cough*
Actual	no h o **diarrhea** thus cdi unlikely
Predicted	no h o diarrhea thus cdi unlikely
Actual	she also notes that she has *not eaten all day*
Predicted	she also notes that she has not eaten all day
Actual	she has a *cough* with green sputum and *dysuria*
Predicted	she has a *cough* with green sputum and dysuria
Actual	pt states that her **pain** is well controlled
Predicted	pt states that her *pain* is well controlled
Actual	her **pain** was well controlled with po pain medications
Predicted	her *pain* was well controlled with po pain medications

Bibliography

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[3] Eugene M Breydo, Julia T Chu, and Alexander Turchin. Identification of inactive medications in narrative medical text. In *AMIA Annual Symposium Proceedings*, volume 2008, page 66. American Medical Informatics Association, 2008.

[4] Yong-Mei Cha. Cardiac resynchronization therapy. *Cardiac Pacing and ICDs, 6*, pages 374–412, 2014.

[5] Neal A Chatterjee and Jagmeet P Singh. Cardiac resynchronization therapy: past, present, and future. *Heart failure clinics*, 11(2):287–303, 2015.

[6] Eugene S Chung, Angel R Leon, Luigi Tavazzi, Jing-Ping Sun, Petros Nihoyannopoulos, John Merlino, William T Abraham, Stefano Ghio, Christophe Leclercq, Jeroen J Bax, et al. Results of the predictors of response to crt (prospect) trial. *Circulation*, 117(20):2608–2616, 2008.

[7] Eugene S Chung, Angel R Leon, Luigi Tavazzi, Jing-Ping Sun, Petros Nihoyannopoulos, John Merlino, William T Abraham, Stefano Ghio, Christophe Leclercq, Jeroen J Bax, et al. Results of the predictors of response to crt (prospect) trial. *Circulation*, 117(20):2608–2616, 2008.

[8] Shortliffe EH. Computer programs to support clinical decision making. *JAMA*, 258(1):61–66, 1987.

[9] Arthur M Feldman, Gregory de Lissovoy, Michael R Bristow, Leslie A Saxon, Teresa De Marco, David A Kass, John Boehmer, Steven Singh, David J Whellan, Peter Carson, et al. Cost effectiveness of cardiac resynchronization therapy in the comparison of medical therapy, pacing, and defibrillation in heart failure (companion) trial. *Journal of the American College of Cardiology*, 46(12):2311–2321, 2005.

[10] Daniel J Friedman, Gaurav A Upadhyay, Alefiyah Rajabali, Robert K Altman, Mary Orencole, Kimberly A Parks, Stephanie A Moore, Mi Young Park, Michael H Picard, Jeremy N Ruskin, et al. Progressive ventricular dysfunction among nonresponders to cardiac resynchronization therapy: Baseline predictors and associated clinical outcomes. *Heart Rhythm*, 11(11):1991–1998, 2014.

[11] Erik K Fromme, Kristine M Eilers, Motomi Mori, Yi-Ching Hsieh, and Tomasz M Beer. How accurate is clinician reporting of chemotherapy adverse effects? a comparison with patient-reported symptoms from the quality-of-life questionnaire c30. *Journal of Clinical Oncology*, 22(17):3485–3490, 2004.

[12] Michael J. Hassett, A. James O'Malley, Juliana R. Pakes, Joseph P. Newhouse, and Craig C. Earle. Frequency and cost of chemotherapy-related serious adverse effects in a population sample of women with breast cancer. *JNCI: Journal of the National Cancer Institute*, 98(16):1108, 2006.

[13] John PA Ioannidis and Joseph Lau. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *Jama*, 285(4):437–443, 2001.

[14] Henry J., Pylypchuk Y., Searcy T., and Patel V. Adoption of electronic health record systems among u.s. non-federal acute care hospitals: 2008-2015. ONC Data Brief No. 35, 2016.

[15] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

[16] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

[17] Guillaume Lecoq, Christophe Leclercq, Emmanuelle Leray, Christophe Crocq, Christine Alonso, Christian de Place, Philippe Mabo, and Claude Daubert. Clinical and electrocardiographic predictors of a positive response to cardiac resynchronization therapy in advanced heart failure. *European heart journal*, 26(11):1094–1100, 2005.

[18] Cecilia Linde, William T Abraham, Michael R Gold, J Claude Daubert, Anthony SL Tang, James B Young, Lou Sherfesee, J Harrison Hudnall, Dedra H Fagan, and John G Cleland. Predictors of short-term clinical response to cardiac resynchronization therapy. *European Journal of Heart Failure*, 2017.

[19] Cecilia Linde, William T Abraham, Michael R Gold, Martin St John Sutton, Stefano Ghio, Claude Daubert, REVERSE (REsynchronization reVErses Remodeling in Systolic left vEntricular dysfunction)

Study Group, et al. Randomized trial of cardiac resynchronization in mildly symptomatic heart failure patients and in asymptomatic patients with left ventricular dysfunction and previous heart failure symptoms. *Journal of the American College of Cardiology*, 52(23):1834–1843, 2008.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[21] Kimberly D Miller, Rebecca L Siegel, Chun Chieh Lin, Angela B Mariotto, Joan L Kramer, Julia H Rowland, Kevin D Stein, Rick Alteri, and Ahmedin Jemal. Cancer treatment and survivorship statistics, 2016. *CA: a cancer journal for clinicians*, 66(4):271–289, 2016.

[22] Arthur J Moss, W Jackson Hall, David S Cannom, Helmut Klein, Mary W Brown, James P Daubert, NA Mark Estes III, Elyse Foster, Henry Greenberg, Steven L Higgins, et al. Cardiac-resynchronization therapy for the prevention of heart-failure events. *New England Journal of Medicine*, 361(14):1329–1338, 2009.

[23] Dariush Mozaffarian, E Benjamin, A Go, Donna K Arnett, Michael J Blaha, Mary Cushman, Sandeep R Das, MD Sarah de Ferranti, Jean-Pierre Després, Heather J Fullerton, et al. Aha statistical update. *Heart Dis. stroke*, 132, 2015.

[24] Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics*, pages 643–674. Springer, 2014.

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

[26] John P Pestian, Christopher Brew, Paweł Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and Włodzisław Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, 2007.

[27] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

[28] Peter M Rothwell. External validity of randomised controlled trials:?to whom do the results of this trial apply?? *The Lancet*, 365(9453):82–93, 2005.

[29] Charles Safran, Meryl Bloomrosen, W. Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C. Tang, and Don E. Detmer. Toward a national framework for the secondary use of health data: An american medical informatics association white paper. *Journal of the American Medical Informatics Association*, 14(1):1, 2007.

[30] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66(1):7–30, 2016.

[31] Arun Raghav Mahankali Sridhar, Vivek Yarlagadda, Sravanthi Parasa, Yeruva Madhu Reddy, Dhavalkumar Patel, Dhanunjaya Lakkireddy, Bruce L Wilkoff, and Buddhadeb Dawn. Cardiac resynchronization therapy. *Circulation: Arrhythmia and Electrophysiology*, 9(3):e003108, 2016.

[32] Anthony SL Tang, George A Wells, Mario Talajic, Malcolm O Arnold, Robert Sheldon, Stuart Connolly, Stefan H Hohnloser, Graham Nichol, David H Birnie, John L Sapp, et al. Cardiac-resynchronization therapy for mild-to-moderate heart failure. *New England Journal of Medicine*, 363(25):2385–2395, 2010.

[33] Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, 20(5):828–835, 2013.

[34] Murdoch TB and Detsky AS. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352, 2013.

[35] Cynthia M Tracy, Andrew E Epstein, Dawood Darbar, John P DiMarco, Sandra B Dunbar, NA Mark Estes, T Bruce Ferguson, Stephen C Hammill, Pamela E Karasik, Mark S Link, et al. 2012 accf/aha/hrs focused update of the 2008 guidelines for device-based therapy of cardiac rhythm abnormalities. *Circulation*, 126(14):1784–1800, 2012.

[36] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008.

[37] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 2007.

[38] Adam Yala, Regina Barzilay, Laura Salama, Molly Griffin, Grace Sollender, Aditya Bardia, Constance Lehman, Julliette M Buckley, Suzanne B Coopey, Fernanda Polubriaginof, et al. Using machine learning to parse breast pathology reports. *Breast cancer research and treatment*, 161(2):203–211, 2017.

[39] Cheuk-Man Yu, Elaine Chau, John E Sanderson, Katherine Fan, Man-Oi Tang, Wing-Hong Fung, Hong Lin, Shun-Ling Kong, Yui-

Ming Lam, Michael RS Hill, et al. Tissue doppler echocardiographic evidence of reverse remodeling and improved synchronicity by simultaneously delaying regional contraction after biventricular pacing therapy in heart failure. *Circulation*, 105(4):438–445, 2002.