Predicting the Effectiveness of Cardiac Resynchronization Therapy Using Natural Language Processing

Austin Freel[1], Josh Haimson[1], Michael Traub[1]
Charlotta Lindvall, MD, PhD.[2]
[1]Massachusetts Institute of Technology
[2]Massachusetts General Hospital

Cardiac Resynchronization Therapy (CRT) has become a standard therapy for a subset of patients suffering from Heart Failure (HF). While CRT is successful in the majority of cases, a significant minority of patients express a neutral or negative response to the therapy without well-understood cause. Since much of the relevant information in clinical data is in narrative text form, current analyses are limited to small information-rich datasets where researchers can read clinical notes on each patient or large information-poor datasets where only structured information on each patient is analyzed. In this paper, we use state-of-the-art natural language processing (NLP) techniques to analyze a large dataset of CRT patients. We find that including the free-text information through the use of NLP techniques both improved prediction accuracy over current clinical performance by $\sim$9% and allowed our model to discover latent clinical variables of the problem. We estimate that if the model generalizes, a 9% reduction in CRT prescriptions could save the US Healthcare system as much as $926MM while simultaneously reducing any adverse side-effects of CRT for many patients who are falsely prescribed the therapy. We hope these results will motivate further research into previously unknown predictors of successful CRT outcomes and demonstrate the benefits of using NLP models in clinical settings.

Code used in this paper can be found here: https://github.mit.edu/jhaimson/6806_final_proj

Introduction

Cardiac Resynchronization Therapy and the Challenges of Evidence-Based Medicine

Cardiac Resynchronization Therapy (CRT) has become an increasingly popular therapy for a subset of patients with Heart Failure. While CRT is effective in the majority of cases, an estimated one-third of patients express a neutral or negative response to the therapy without well-understood cause (Chatterjee and Singh, 2015). Since the procedure is both expensive and invasive, much effort is being placed into understanding the primary predictors of success.

The current clinical guidelines for CRT can be seen in the decision tree shown in Figure 1. The guidelines make a recommendation based on 7 concrete variables: NYHA class, LVEF, QRS, LBBB, Sinus Rhythm, Ischemic Cardiomyopathy and the presence of comorbidities. The guidelines also leave significant room for interpretation in their recommendations, with recommendations as vague as "CRT might be reasonable," (Tracy et al., 2012).

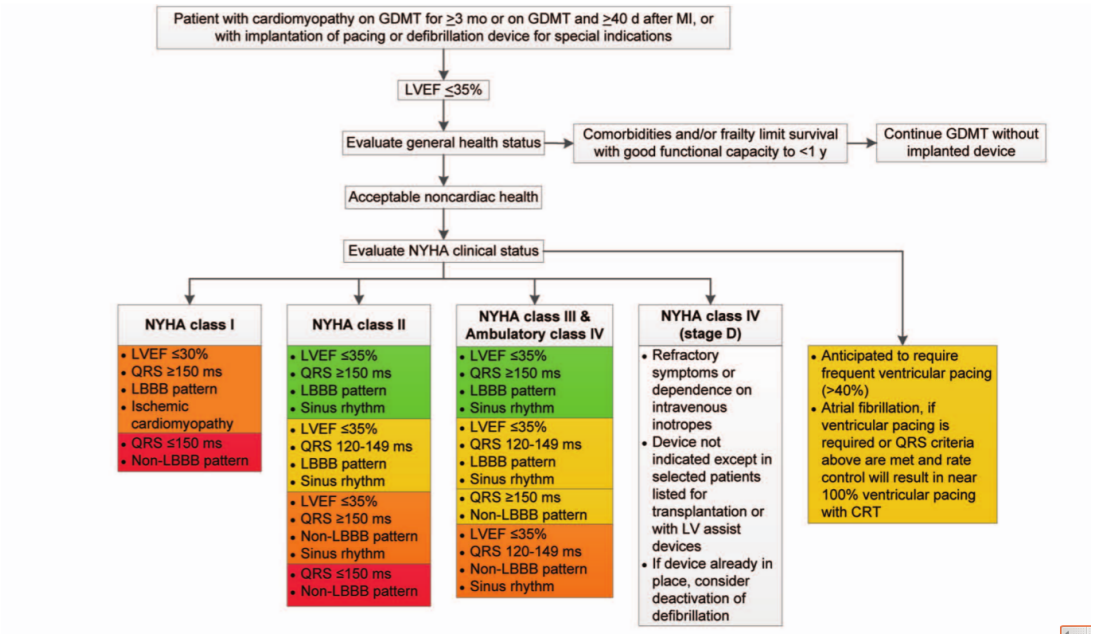

Figure 1: The current clinical guidelines for CRT. The guidelines are determined by color, where green means "CRT is recommended," yellow means "CRT is reasonable," orange means "CRT might be reasonable," and red means "CRT is not recommended." (Tracy et al., 2012)

One of the core challenges of understanding the effectiveness of a treatment in a clinical setting is the inherent messiness of real-world clinical data. Although most healthcare systems have adopted electronic medical records and systems such as ICD codes have been adopted for epidemiological/billing purposes, these systems tend to be siloed from each other in non-

interoperable formats and much of the clinically relevant data contained in a medical record is locked in narrative text notes. This means that analyses of clinical medical data are usually limited to either a) a cohort size small enough that researchers can manually read every note, or b) the limited number of variables that are stored as structured data in a medical record.

An example of such an analysis for CRT can be seen in Friedman et al. (2014). In the study, the authors analyze 32 variables (various demographics, lab values, comorbidities and medications) in a cohort of 328 patients who received CRT. While they successfully determine a number of factors in the success of CRT, they are limited in the size of the cohort they can analyze because they had to extract these variables manually. Scaling the size of this cohort in order to find trends that are only visible in larger populations quickly becomes an intractable problem, and switching to analyzing structured data only will miss important pieces of information like key lab values, symptom expression, family history, and social history.

Natural Language Processing (NLP) techniques can help overcome these challenges by enabling the computational analysis of narrative text notes in large patient cohorts.

Research Goals

The goals for this research were twofold:

1. Build a model which can predict the success of CRT using both free text and structured information
2. Determine if such a model can find predictors of success that have previously been overlooked

Methods

MGH Dataset

In this section we will describe our data set with respect to both its features and labels, and evaluate how our sample may be biased and what this means for our research.

Description of Data

Our dataset consists of the complete medical histories of 907 patients who received CRT at Massachusetts General Hospital (MGH). This includes both structured data for each patient (lab values, diagnosis codes, etc) and unstructured notes from their doctors. Some of the information that was especially useful and relevant is summarized in Table 1.

Data type	Document	Information Contained
Structured	Encounters	Inpatient/Outpatient, Duration of stay
	Diagnosis Codes	ICD-9 Code
	Labs	Lab type, Value, High/Low indicator
Unstructured notes	Cardiology reports	Relevant lab values (EF, QRS), Clinical characters (LBBB, sinus rhythm), Cardiologist notes
	Longitudinal Medical Record (LMR)	Summary of lab values, Symptoms, Family history, Social history

Table 1: Summary of the types and variety of information contained in the data set for each patient.

Patients	907
% Responders	52.6%
Structured Documents	3100K
Note Documents	245K
Structured Fields	44M
Note Sentences	26M

Table 2: Summary of the size of our dataset with respect to the number of patients and information contained in notes and structured documents.

Below are example note excerpts from two patients taken from notes in their LMR records. These illustrate how these notes contain information about the patient history, symptoms, medical characteristics, and the sentiment of their doctor.

> "A very pleasant 68-year-old gentleman with a history of ischemic cardiomyopathy presented with class III symptoms of heart failure, has had an upgrade of his device to biventricular implantable cardioverter-defibrillators, currently in sinus rhythm."

> "This is a 54-year-old woman with end stage heart failure secondary to Chagas disease. Her main symptoms are shortness of breath, chest discomfort, anxiety, and existential distress."

We find that we have over ten times as many structured documents as notes, or free-text documents. However, when looking at the number of structured entry fields and the number of sentences to find a better comparison for units of information, we see that over one-third of the information in a patient's medical records is contained in notes. This is summarized in Table 2 and Figure 2. This heuristic along with the fact that the most clinically relevant information for CRT is stored in notes indicate the importance of better utilizing this aspect of a patient's

medical history.

Definition of CRT Success

Prior research has defined the success of CRT measures by the magnitude of the change of left ventricle ejection fraction (LVEF or EF) from before the procedure until a fixed time period after the procedure, (Friedman et al., 2014). Upon recommendation from cardiologists at MGH, we classify patients with a change in EF after 12 months $\geq 8\%$ as *responders* and patients with less than this as *non-responders*. However, EF values were not directly provided as structured data in the data set, so we first had to extract them from the cardiology notes. To extract these ejection fraction measurements, we defined a regular expression that would extract these values from free text notes and label them with the date of the note they were extracted from. With these dated values, we determined a patient's classification using the value before the date and the value measured closest to 12 months after the procedure. We have checked against manually extracted values that this procedure works reliably, though small errors may remain in the label extraction process which may add artificial noise to our data.

Biases of Our Cohort

The most important characteristic that all of our patients have in common is that they were recommended by a clinician to receive CRT. This means that our data set contains only true and false positives with respect to the prediction methods used by doctors. Another way to say this is that if doctors already possessed perfect knowledge of their patients and the factors that determine success of the procedure, then the data should contain nearly 100% responders. Therefore, we interpret the fact that our data contains 52.6% responders as a current medical prediction baseline for our machine learning models.

Baselines

In order to set a benchmark to compare our NLP results against, we created two baselines. The first baseline is a hard-coded decision tree model that matches the decision tree shown in Figure 1, which clinicians use today to decide whether or not CRT is appropriate for a given patient. Of the four possible output categories of the decision tree, which range from *CRT is not recommended* to *CRT is recommended*, we deem the red and orange categories as *should not have CRT* and the yellow and green categories as *should have CRT*. Although this is a generalization, it is the best we can do given the inherent vagueness of the outputs and the fact that physicians often differ from one another in regards to what to do in the orange and yellow categories.

It is worth noting that our hard-coded model is not exactly the same as the decision tree shown in Figure 1, since we could not extract all input values, and since the validity of our extractions is still being verified. However, we ensured the validity of a portion of them that had

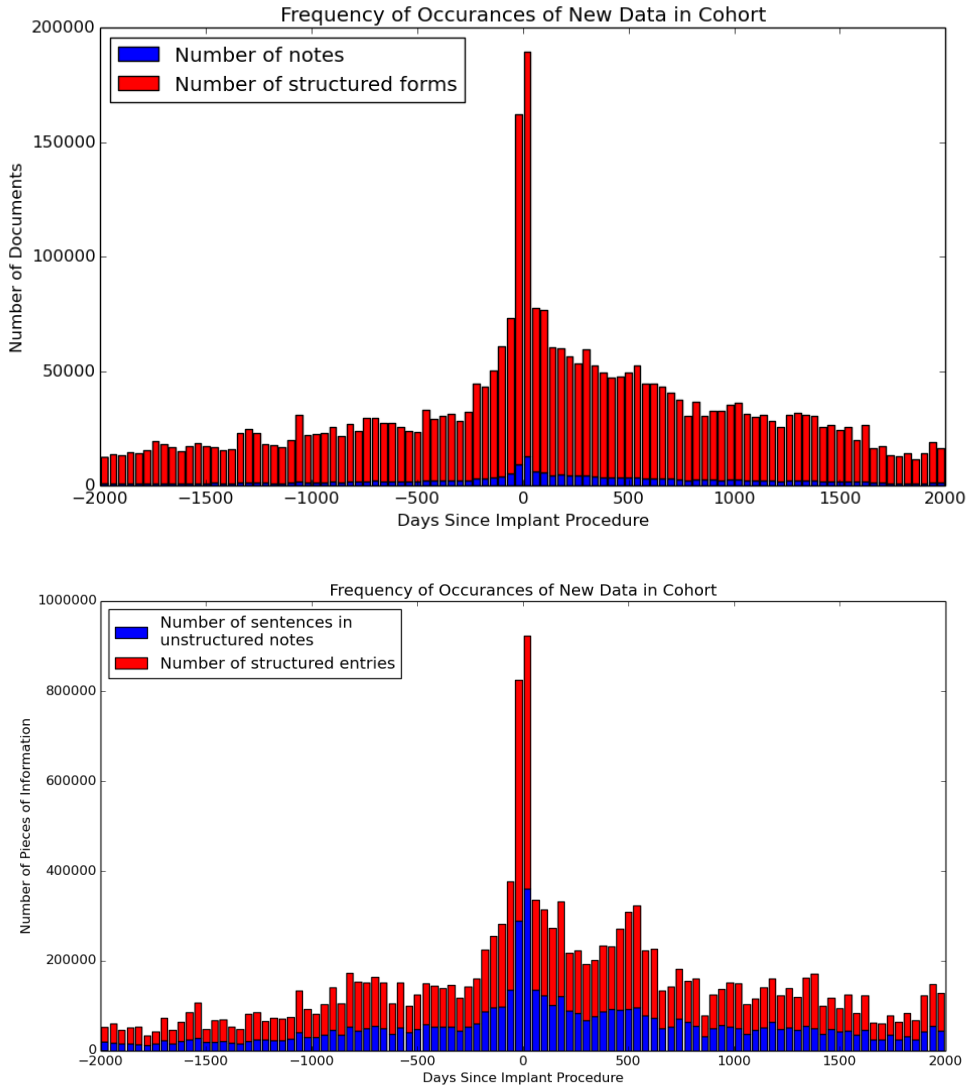

Figure 2: Above: plot of the frequency of structured and note documents for each relative date with respect to the date of the CRT procedure for the patient. Below: The same plot except we compare the fields in the structured data and the number of sentences in the notes. This second plot better captures the relative information content of these different sources.

corresponding annotations, and so it is believed that our decision tree model is quite accurate. It is also worth noting that all values extracted are from before the procedure date. We remain consistent with this throughout all of our tests so that we are sure to be predicting without any post-procedure knowledge.

The second baseline is a structured-data-only baseline which is designed to model the fields that current research focuses on. It uses a concatenation of the features we have engineered from the Diagnoses, Encounters, and Labs fields of the patients. The specifics of each of these feature vectors are discussed below.

Structured Data Models

Diagnoses

Over the course of a patient's medical history, each diagnosis they receive is codified as a structured field which can be analyzed for epidemiological purposes. The format of this structured field is defined by the World Health Organization in a format known as International Classification of Diseases (ICD). While the US healthcare system has recently switched to the ICD-10 standard, the data in the MGH CRT dataset is from before that switch and is using the ICD-9 standard.

In creating a feature vector derived from these ICD-9 codes, we wanted to be able to capture a feature space that could adequately represent the space of possible comorbidities and past medical experiences that a patient could have. In order to do this, we used the Clinical Classifications Software (CCS) published by the Healthcare Cost and Utilization Project (HCUP) to transform each individual ICD-9 code into a hierarchy of disease. For example, the ICD-9 code for Endocardial fibroelastosis is 425.3, but using CCS we can convert this into the code 7.2.2, where 7 represents "Disease of the Circulatory System," 7.2 represents "Heart Disease," and 7.2.2 represents "Cardiomyopathy, "(Agency for Healthcare Research and Quality: Healthcare Cost and Utilization Project, 2008).

Once we have this hierarchical code, we create a "one-hot" vector for each level in the hierarchy, allowing us to simultaneously represent a disease like Endocardial fibroelastosis as "Disease of the Circulatory System," "Heart Disease," "Cardiomyopathy," and "Endocardial fibroelastosis."

Lastly, once we have such a vector for each diagnosis in a patient's medical history, we sum these vectors to get a diagnosis vector representing the entire patient's diagnosis history.

Encounters

Every patient has a record for each encounter he or she has had with a hospital, which could be anything from a quick checkup to spending days or weeks in the hospital. It was determined that the most relevant fields recorded at each encounter are (a) Inpatient vs. Outpatient, which is a boolean of whether the patient had to be checked in overnight, (b) Length of Stay, which would be 0 if this was an Outpatient encounter, and (c) Number of Extra Diagnoses, which is

a number of diagnoses recorded on that encounter beyond the primary diagnosis that was the cause of the encounter.

Through testing, it was determined that looking at the last N pre-procedure encounters was the optimum strategy, and the specific feature vector we ended up using looked at the last 5 encounters. The vector had three features per encounter, corresponding to the three main data points mentioned above, as well as three extra features corresponding to averages of these points, namely the inpatient ratio, average length of stay, and average number of extra diagnoses.

Labs

Every patient also has a history of lab records. It was determined that by far the most relevant field is, unsurprisingly, the lab value. Since our data set is fairly small, we first decided that instead of using the actual lab value we would instead use flags that indicated whether the lab result was normal (N) or abnormally low (L) or high (H), thus reducing the range of possible results. A handful of feature vectors were built to try and understand the labs and use them as predictors. Specifically, they looked at features such as total counts for each lab, total L and H counts for each lab, and the latest flag (N/L/H) for each lab.

The reasoning behind these features is that two patients who frequently receive the same lab tests probably have similar diagnoses, since these tests are used to assess the severity of diagnoses and progress of treatments. Looking also at L and H flags then helps capture relative performance on these tests, and looking at the most recent tests reflects the fact that lab values leading up to the procedure are most relevant.

As we developed our models further, we also attempted to use the lab values, both numerical and categorical, as inputs rather than the flags.

Natural Language Models

Clinical Value Extraction (CVE)

Having made use of most of the structured data, we then turned to understanding the free text. The first strategy was to use regular expressions to parse the notes and pick out values that were already known to be important predictors, namely the same values that were inputs to the clinical decision tree (NYHA Class, LVEF, QRS, LBBB, Sinus Rhythm). However, instead of using a pre-defined decision tree, we instead used these values as inputs into different machine learning models and let the machine determine the correct decision lines.

Bag of Words (BOW)

We then turned to bag-of-words techniques that assumed no previous knowledge of the problem and instead looked at the free text as a whole. Since the number of notes for each patient was so large, we attempted to reduce the number of notes we looked at in order to reduce both complexity and noise. To achieve this, we only looked at the N notes leading up to the procedure

for each patient. We also only looked at Cardiology Reports and Longitudinal Medical Records, as these were deemed most important and information-rich. Together, these simplifications left us with a smaller and denser dataset to work with.

The first technique we tried was tf-idf, knowing that it has the power to differentiate between important words in a document and words that just are quite common over the whole corpus. Without satisfying results, we then turned to n-grams, trying unigram, bigram, and trigram models. To help reduce the noise generated from highly infrequent words, which is quite large according to Zipf's law, we also set a threshold on the number of documents a word must appear in to be acknowledged by the model. Through testing of the different approaches and performing grid searches of the parameter space, we found that the bigram model performed the best, specifically with a document frequency threshold of 5%. These basic models actually ended up performing quite well, and their results not only increased predictive accuracy considerably, they also pointed to new latent predictive features in the free text clinician notes.

Paragraph Vectors

Paragraph vectors are a state-of-the-art technique for representing arbitrary length sequences of words in a fixed dimensional space. They are particularly well-suited for our dataset because we have large amounts of text per patient, but a small amount of patients on which to predict CRT success (small from a machine learning perspective). While other state-of-the-art neural language models such as Long Short-Term Memory (LSTM) models or Gated Recurrent Unit (GRU) models also perform well at classifying arbitrary lengths of text, they require large amounts of labeled samples for training. Paragraph vectors avoid this problem by training feature representations in an unsupervised manner.

As shown in Figure 3, paragraph vector representations are created by training a neural classifier to predict a given word in a sentence given n-dimensional feature representations of the words in a context window around the given word (Mikolov et al., 2013) and an n-dimensional paragraph vector which stays constant across all context windows in a paragraph. A paragraph in this context refers to any arbitrary length sequence of words, ranging in size from one sentence to an entire document. In training, all parameters are updated via stochastic gradient descent to maximize the log-probability of a paragraph. In testing, the algorithm infers the paragraph-vector by fixing the word-vector representations and again maximizing the log probability of the paragraph, (Le and Mikolov, 2014).

In our usage of paragraph vectors, we used the concatenation of Distributed Memory Paragraph Vectors (PV-DM) and Distributed Bag of Words Paragraph Vectors (PV-DBOW) generated for a given paragraph to achieve a feature representation equivalent to the highest performing representations in Le and Mikolov (2014). We chose to treat each document as a "paragraph," obtaining a 600-dimensional feature vector for each document (300 dimensional PV-DM concatenated with a 300 dimensional PV-DBOW). We trained two of these models seperately, one for Cardiology notes and one for Longitudinal Medical Record (LMR) notes. When using these models to represent free text information from a patient's medical records, we

concatenate the last k Cardiology and LMR notes before the procedure date into a $2 * k * 600$-dimensional feature vector.

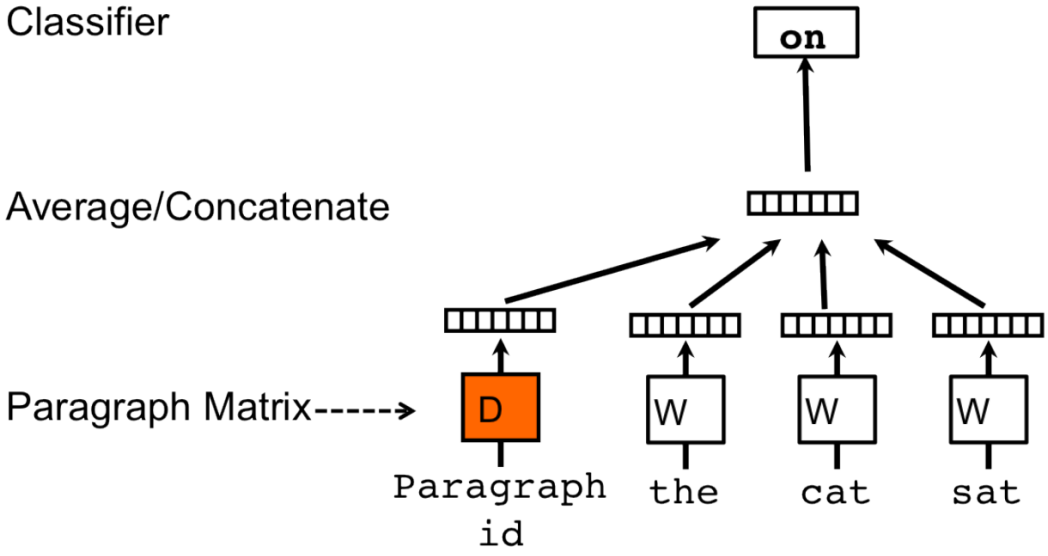

Figure 3: Paragraph vector representations are created by training a neural classifier to predict a given word in a sentence given n-dimensional feature representations of the words in a context window around the given word (Mikolov et al., 2013) and an n-dimensional paragraph vector which stays constant across all context windows in a paragraph, (Le and Mikolov, 2014).

Classifiers

When making predictions based on the feature vectors described above, we explored a number of standard classification techniques. Specifically, we tested Maximum Entropy, Support Vector Machine, Adaboost and Decision Tree Classifiers. We did not test neural network based methods as we did not have enough sample data for a neural network to efficiently converge. From initial testing, there were minimal performance differences between the various classifiers. Therefore, we chose to work with the Adaboost classifier for the interpretability of its learned parameters.

Results

NLP Models Improved Prediction Accuracy by $\sim$9%

Various combinations of the features and classifiers described above were run on our dataset of 907 CRT patients from MGH. The results of our models are reported in Table 3. All reported results have a train:test split ratio of 2:1 with 5-fold cross-validation. For each of the models, we performed a small gridsearch over their hyperparameters, and the best performing configuration of the models are reported. It's important to note that these configurations have not been determined by an exhaustive search through the space of model hyperparameters, so there may still be room for performance improvements.

Features Description	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Clinical Performance	.526	.526	1*	.689*
Tracy et al. (2012) Decision Tree	.538	.590	.300	.397
Structured Data Only (ICD9, Encounters, Labs)	.546	.565	.543	.551
Clinical Value Extraction (CVE)	.593	.626	.593	.608
CVE + Structured Data	.578	.604	.601	.602
CVE + BOW Bigrams	.612	.610	.705	.652
CVE + Structured Data + BOW Bigrams	.583	.603	.621	.609
CVE + Structured Data + Paragraph Vectors	.585	.600	.592	.594

Table 3: Optimal mean scores for each set of features input into an AdaBoost classifier, found with a grid-search over the number of weak learners. Each configuration was run according to a 2:1 train:test split and 5-fold cross-validation split. This corresponds to a 90% confidence bound of $\pm.021$ for all values. *Clinical performance has artificially high Recall and F1 scores because our dataset only contains patients that were prescribed CRT, meaning there were no negative clinical predictions.

NLP Models Highlight Relevant Clinical Predictors

When looking at the learned parameters for our best performing model, CVE + BOW, we find that the highest weighted features contain interesting predictors, which include both expected symptoms and new findings. The fact that symptoms were amongst the best predictors not only validates our model, it also helps point out which symptoms are important to look at as predictors. In the bigram model, the parameters included symptoms such as "ventricular arrhythmia" and "back pain". The trigram model was able to pick up on even more symptoms, such as "elevated la pressure" and "anteroseptal myocardial infarction", just to name a few.

Perhaps more interesting than the symptoms were the unexpected findings. For example, highly predictive bigrams included phrases like "placed he" and "his visit" that explicitly referenced gender and thereby showed gender to be a good predictor. Others tapped into concepts

like family history, such as the bigram "father died". And still others were more abstract, expressing forms of sentiment, such as the bigram "pleased to". The trigrams even began to pick up on more temporally sensitive phrases, such as "was well until" and "started on heparin". These results are exciting not only because they show the ability of the model to hone in on complex and interesting features, but also because they show the model's ability to highlight new, previously overlooked predictors. Moving forward, we hope to share and discuss these results with researchers and clinicians, validate features extracted by our models, and then incorporate them into newer, better models.

Discussion

Predicting the effectiveness of CRT for a given patient is still not a solved problem. In fact, given the data in our corpus, it is far from solved, with only 52.6% of our patients having benefited from the treatment. Since our data only contains true and false positives, we have had to reframe the problem we are trying to solve. Namely, we want to reduce the number of false positives and help physicians answer the question "Does this patient really need CRT?". Since CRT is so expensive, any reduction of false positives would lead to a lot of saved money and reduced burden on the patients. Thus far, our models have reduced the false positive prediction rate in our corpus from 48% to 39%. Assuming about 80,000 patients receive CRT a year and each one costs around $130,000 USD, this 9% decrease in false positives could translate to $936 million in savings per year. We hope to see our accuracy grow even higher in the near future as we combine our results from the structured and unstructured data and more exhaustively explore the space of hyperparameters.

Most clinicians today also only look at a small set of known predictors in evaluating a patient for CRT. However, even through the use of a simple bigram bag-of-words model, we were able to show that there are likely many more key predictors that can and should be looked at by clinicians. As we continue to improve our models, we hope to cooperate with researchers to validate the new predictors we have found and try to incorporate them into newer, more accurate models.

Overall, we were able to use both basic and state of the art natural language processing techniques to increase the predictive accuracy of our models and elucidate latent predictors contained in the free text of clinician notes. We look forward to continuing to work on this problem and hope to show that in the sometimes chaotic realm that is medical data, NLP has the ability to make sense of data, deepen our understanding of a problem, and thereby help patients receive the help they need.

References and Notes

Agency for Healthcare Research and Quality: Healthcare Cost and Utilization Project. Clinical classifications software (ccs) for icd-9-cm, 2008.

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

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.

Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

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.

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: a report of the american college of cardiology foundation/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 60(14):1297–1313, 2012.