An NLP-based Cognitive System for Disease Status Identification in Electronic Health Records

Homa Alemzadeh, *Member, IEEE* and Murthy Devarakonda, *Fellow, IEEE*
IBM Research, Yorktown Heights, NY

***Abstract*—This paper presents a natural language processing (NLP) based cognitive decision support system that automatically identifies the status of a disease from the clinical notes of a patient record. The system relies on IBM Watson Patient Record NLP analytics and supervised or semi-supervised learning techniques. It uses unstructured text in clinical notes, data from the structured part of a patient record, and disease control targets from the clinical guidelines. We evaluated the system using de-identified patient records of 414 hypertensive patients from a multi-specialty hospital system in the U.S. The experimental results show that, using supervised learning methods, our system can achieve an average 0.86 F1-score in identifying disease status passages and average accuracy of 0.77 in classifying the status as controlled or not. To the best of our knowledge, this is the first system to automatically identify disease control status from clinical notes.**

I. Introduction

Electronic Health Records (EHRs) are the main source of information for assessment, diagnosis, and treatment of disease in clinical care. An EHR typically contains a patient's historical health data, collected over several years of patient care. This data includes both physicians' clinical notes written in unstructured text recording their observations, assessments, and plans, as well as structured data such as ordered medications, vital signs measurements, laboratory test results, and procedures conducted.

With vast amounts of data being recorded in a patient record, manual retrieval of relevant information for a specific clinical task is challenging, often causing cognitive overload and inefficiency for physicians [1]. Various systems have been developed to support physician decision making by automatically generating clinical summaries [1][2][3], problem lists [4], and treatment performance measures [5] based on EHR data. An important insight that can be automatically extracted from EHRs and provided to a physician is the status of active problems. Such information can be used to assess the effectiveness of the ongoing treatment and to decide follow up actions. A longitudinal chronology of disease status can also be used to conduct detailed epidemiological studies leading to better population health policies and treatment strategies.

Previous work on the prediction of disease status from EHRs mainly focused on using the structured data, identifying the presence or absence of disease [6], or detecting assertions (e.g., hypothetical or historical) [7] from clinical notes. Structured data alone is not adequate for identifying disease status because a physician's assessment that considers co-morbidities, age, and other conditions is the most important for patient care. An automated information extraction system for measuring congestive heart failure (CHF) treatment performance was reported in [5]. However, the system only extracted disease status measures, medications, and reasons for not receiving certain medications from clinical notes.

In this paper, we present an NLP-based cognitive system for automated extraction of disease status using *both unstructured and structured data* in a patient record. The system identifies whether the patient status complies with disease-specific control targets for each visit. We use the IBM Watson Patient Record NLP analytics to extract mentions of a disease and its related test results from the clinical notes. This information, along with the vitals and test results recorded in the structured data of the patient record, is then used as features in a machine learning model that classifies the status of the disease in each clinical note (i.e., for each visit) as "Unknown", "Controlled", or "Not Controlled".

For evaluating the system, we specifically focused on hypertension, as an example of a chronic condition, which affects about 70 million adults (29%) in the United States [8]. The data was collected from 414 de-identified EHRs of hypertensive patients in a major multi-provider U.S. hospital system. Our dataset included 5,035 candidate snippets of text, which potentially discussed hypertension status, extracted from over 55,000 clinical notes. Of these, 2,086 snippets were manually labeled and all the data was also automatically labeled based on rules applied to blood pressure measurements recorded in the structured data. In addition to supervised learning algorithms trained on manual and automated labels, we adapted a semi-supervised method called co-training to expand our set of manually labeled training data with additional 2,949 unlabeled examples. The experimental results show that, using supervised learning, our system achieves an average F1-score of 0.86 in identifying disease status mentions in the clinical notes and an accuracy of 0.77 in classifying the mentioned disease status as "Controlled" or "Not Controlled".

II. System Description

The disease status chronology is determined by creating a timeline of disease status, in which for each patient visit the status of disease is classified as "Unknown" ("NA"), "Controlled" ("C"), or "Not Controlled" ("NC"), based on the evidence in the patient record. Disease status chronology can provide a summary of disease progression over time by highlighting transitions between "Controlled" and "Uncontrolled" states. For example, multiple transitions between "Controlled" and "Uncontrolled" states can reveal a less stable condition compared to a patient who mostly stays in a "Controlled" state. This can be used as a measure of treatment performance, indicating which changes in the

This work was done when H. Alemzadeh was at IBM Research. She is now with the Department of Electrical and Computer Engineering at the University of Virginia, Charlottesville, VA 22904 (e-mail: alemzadeh@virginia.edu).

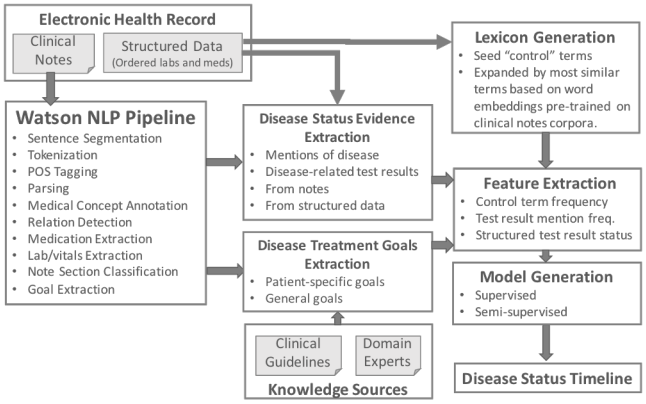


Figure 2. Overall system structure

(a)

```
401.9 HYPERTENSION NOS
Comment: not controlled today.
Plan: Increase Lisipnopril from 20 to 40 mg.
Is also on metoprolol.
He will get a home machine and monitor BP and
```

(b)

```
401.9 HYPERTENSION NOS
Comment: BP still running high 160/85.
```

(c)

	encounterID	vitalsDate	vitalsName	value
1	4181-5 (1 of 172)	2006-03-17 12:00:00	WEIGHT/SCALE	2720
2	4181-5 (2 of 172)	2006-03-19 19:00:00	BLOOD PRESSURE	146/84
3	4181-5 (3 of 172)	2006-03-19 19:00:00	PULSE	102
4	4181-5 (4 of 172)	2006-03-19 19:00:00	PULSE OXIMETRY	97
5	4181-5 (5 of 172)	2006-03-19 19:00:00	RESPIRATIONS	18
6	4181-5 (6 of 172)	2006-03-19 19:00:00	TEMPERATURE	100.8
7	4181-5 (7 of 172)	2006-03-19 23:18:00	BLOOD PRESSURE	112/74
8	4181-5 (8 of 172)	2006-03-19 23:18:00	PULSE	103

Figure 1. Three types of disease status evidence in an EHR: (a) mentions of disease in unstructured notes, (b) mentions of vital sign measurements in unstructured clinical notes, and (c) structured vital sign measurements

treatment plan have been most effective over time. Figure 1 shows the overall structure of our cognitive system for identifying disease status, and details are described next.

A. Disease Status Evidence Extraction

The disease status can be *potentially* inferred from one of the following sources in a patient record:

- Mentions of the disease in a clinical note, typically in the "Assessment and Plan" section, for example:
  "HTN: Continue Present Meds"
  "HTN is well controlled"
- Mentions of the disease-related laboratory tests or vital signs measurements in a clinical note, for example:
  "BP is still remaining high"
  "Started to use home BP machine"
- Related diagnostic test results or vital signs recorded in the structured data, for example:
  Blood Pressure 160/80

Examples from a patient record for these sources is shown in Figure 1. Note that some of the textual mentions, such as "started to use home BP machine" do not inform whether the disease is controlled or not.

We use Watson Patient Record NLP analytics components including sentence segmentation, medical concept annotation, note section classification, labs/vitals extraction, and relation detection, to identify the snippets of text related to disease status within the clinical notes. Specifically, we extract sentences containing the mentions of disease (e.g., HTN or hypertension) and related vitals and diagnostic test results (e.g., BP or blood pressure) in the "Assessment and Plan", "Chief Complaint", "History of Present Illness", and "Diagnosis and Finding" sections of "progress" notes. The piece of text starting from one sentence before to one sentence after the sentence containing the mention of disease or test result is recorded as a candidate snippet for identifying disease status.

The vital sign measurements and lab results are retrieved from the structured data sections of a patient record and the Watson relation detection component is used to identify those that are related to a specific disease (e.g., BP is a vital sign related to hypertension). We identify the vitals and test results related to a clinical note based on the common encounter IDs and dates on which the measurements were made and notes were written. Specifically, we search for the related vital signs measured within a period of three days before and after a clinical note date and identify the vitals with the same encounter ID or with the closest date distance to the note.

B. Disease Treatment Goals Extraction

We extract disease-specific control targets from the standard clinical guidelines (e.g., JNC VIII for hypertension [9]). These treatment goals are converted into a set of logic rules based on the patient demographic features (e.g., age, gender, ethnicity, pregnancy status) and diagnostic test results (e.g., blood pressure measurements). These rules can be executed by a reasoning engine in real-time to determine the disease status relative to the control targets solely based on the structured data.

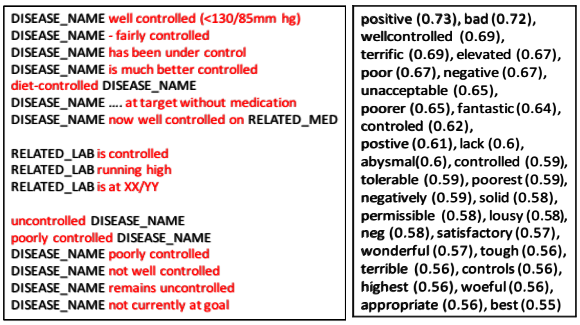
C. Feature Extraction

The extracted disease evidence and treatment goals are assembled into a set of features used by the machine learning algorithm. The following are the key features that were selected for constructing our model:

Control term frequency: Based on a review of 150 clinical notes (a very small subset of the total), we manually identified a few patterns that describe the disease status (see Figure 3). From these patterns, we developed pattern matching rules using regular expressions, parse tree pattern matching, mentions of the disease, and a dictionary of 11 'disease status' seed terms (e.g., controlled, uncontrolled, high, low, and suboptimal) to identify the disease status in candidate snippets. However, the accuracy of the rule-based approach was low (micro average F1-score of 0.38), as it was highly dependent on the set of terms and patterns included in our dictionary, but it formed the baseline for our experiments (see Table I).

We further expanded the disease status seed terms by finding similar terms using word embeddings (word2vec models [10]) pre-trained on three separate corpora: clinical notes from all EHRs from the hospital system; i2b2 database [11]; the Google News data [12]. The final lexicon of disease status consisted of 519 terms that appeared in a similar context to the seed terms with a similarity score of 0.40 or better. Figure 3(b) shows examples of the highest scored terms in the lexicon. One interesting observation is that misspellings of seed control terms (e.g., 'controled' and 'postive') are automatically detected and included in the lexicon.

The standard term frequency (TF) metric is used to compute a feature vector representing the frequency of each disease status term in a text snippet of a clinical note.



(a) (b)

Figure 3. (a) Disease status patterns in clinical notes, (b) Examples of 'disease control' terms and their scores

Test result mention frequency: The disease-related vital signs in a text snippet are identified using regular expression patterns (e.g., 'BP' or 'blood pressure' mentioned before a numerical value: 'BP 160/80'). The frequency of such patterns in a text snippet is encoded as a feature in the model.

Structured test result status: The status of disease control targets (e.g., blood pressure within limits or abnormal and high) are generated by the logic rules executed on the related vital sign measurements. This status ("NA", "C", or "NC") is used as a binary feature vector representing the structured test results status in one experiment, and as automatically generated labels for training the machine learning models in a separate experiment, as described next.

D. Ground Truth Generation

Typically, clinical text classification tasks are done using supervised learning algorithms trained on previously labeled data created by medical experts. So, we manually labeled a subset of candidate snippets and used this as the main ground truth for supervised training and testing of our classifiers. However, given the large number and size of notes in a patient record, manual labeling of data is not only expensive, but also prone to errors due to possible mistakes or inconsistencies among annotators. This often results in generation of data sets that are of limited size and thus difficulty in achieving satisfactory generalization for classifiers. To address this challenge, in addition to the manually generated labels, we also *automatically labeled* all the data based on the disease status inferred from blood pressure measurements in the structured data. This is done based on the assumption that there is agreement between the disease status evidence in the unstructured clinical notes and the structured sections of a patient record. However, when vital sign measurements (structured evidence on disease status) are present, there is no guarantee that any mention of the disease status is also present in the related clinical note. So, it makes sense to only use the automatically generated labels for cases where disease status evidence is known to be present in a clinical note (and in the structured data) and perform "C" vs. "NC" classification.

In addition, we adapted a semi-supervised training method called co-training to increase the size of the labeled training data. Co-training requires different classifiers trained on two different feature sets that provide independent but complementary views of the data. These two classifiers are both trained on a labeled data set and then train each other by iteratively labeling new samples from a large set of unlabeled data and adding the identically labeled new samples to the labeled training set [13].

E. Model Generation

We first experimented by creating a supervised classifier trained on manually generated labels with different models, including Naïve Bayes, Decision Trees, and Support Vector Machines (SVM), as well as ensemble methods such as AdaBoost and RandomForests. Our analysis using a 10-fold cross-validation scheme, showed that the SVM model with linear kernel achieved the best trade-off between model performance (F1-score) and complexity among the classifiers, so we used SVM in all of the following experiments. We also found that two levels of binary classifiers ("NA" vs. "C/NC", and "C" vs. "NC") achieves a better performance compared to one ternary classifier ("NA" vs. "C" vs. "NC"). We developed the following classifiers (also see Table I):

Classifier 1 – Unstructured features and manual labels: This classifier extracts features from the unstructured clinical notes and is trained and tested on manually generated labels.

Classifier 2 – Structured features and manual labels: This classifier uses the disease status based on the structured data as the only feature vector and is trained and tested on manually generated labels.

Classifier 3 – Unstructured features and automated labels: We used automatically generated labels for training the third classifier, which uses only the unstructured features. This binary classifier was only used to classify the samples that were known to have an indication of disease status (were not in "NA" class) to label them as a "C" or "NC" class.

Classifier 4 – Co-training of classifier 1 and classifier 2: This classifier used co-training by leveraging classifiers 1 and 2, which use distinctly different (i.e. unstructured and structured) features. We initialized classifiers 1 and 2 by training them on manually generated labels. In the first iteration, 50 random samples from unlabeled data were selected and labeled by the classifiers. Then, five "NC" and five "C" samples which were identically labeled by both classifiers were added to the labeled data set. In subsequent iterations, we randomly picked 10 samples from the unlabeled data and continued the co-training process until no more samples were left in the unlabeled data set.

III. RESULTS AND DISCUSSION

Our dataset was composed of 414 patient records in which hypertensive disorder was mentioned as one of the main problems in the problem list. The raw data included 55,000 clinical notes with an average of about 133 notes, collected over an average of 7.2 years, per patient. After extracting the disease evidence from the clinical notes, our dataset consisted of 5,035 candidate snippets, from which 2,086 were manually labeled and the rest (2,949) were used as an unlabeled set for semi-supervised learning. The labeled data set consisted of 743 "NA" samples, 931 "C" samples, and 412 "NC" samples.

Table I shows the performance of the classifiers in terms of average F1-score calculated over ten 10-fold cross-validation experiments (i.e., 100 runs). Except classifier 4, all other classifiers were evaluated by comparing their results against the manually generated ground-truth. For each classifier, we

TABLE I PERFORMANCE OF CLASSIFIERS

Classifier	Features	Ground Truth for Training	"C" vs. "NC"					"NA" vs. "C/NC"				
			"NC" Class			Accuracy		"C/NC" Class			Accuracy	
			Precision	Recall	F1	Avg.	Std.	Precision	Recall	F1	Avg.	Std.
Baseline	Rule-based	Manual labels (Unstructured data)	0.59	0.11	0.19	--		0.64	0.13	0.22	0.39	-
Classifier 1	Unstructured data		0.79	0.33	0.46	0.77	0.04	0.91	0.82	0.86	0.83	0.03
Classifier 2	Structured data		0.55	0.66	0.60	0.73	0.04	0.64	1.00	0.78	0.64	0.03
Classifiers 1+2	Unstructured and Structured data		0.78	0.35	0.47	0.76	0.03	0.91	0.82	0.86	0.83	0.02
Classifier 3	Unstructured data	Automated (Structured data)	0.42	0.72	0.52	0.54	0.08					
Classifier 4	Unstructured data	Co-training (Classifiers 1 and 2)	0.83	0.30	0.43	0.75	0.02					

report the precision, recall, and F1-score for the positive class (requiring intervention) and the overall accuracy, which is the fraction of samples labeled correctly out of all of the test samples. As shown in Table I, classifier 1 achieves the best F1-score of 0.86 (average) in identifying snippets that contain any indication of disease status ("C/NC"). This classifier can classify "C" and "NC" classes with an average accuracy of 0.77, but it can identify the "NC" class with an F1-score of 0.46 only. In both cases, we see much improvement versus the rule-based method (with F1-scores of 0.22 and 0.19).

The results for classifier 2 show that by only using the evidence from structured data, the F1-score for identifying disease status snippets ("C" or "NC") drops to 0.78, but we can achieve 14% higher performance for identifying the "NC" classes (0.60). The lower F1-score for "NA" vs. "C/NC" classification might be due to disagreement between mentions of disease in the notes and measurements collected in the structured data. For example, we found that for 461 candidate snippets, there were blood pressure measurements in structured data from which the hypertension status could be inferred, but no mention of disease status was present in the candidate snippet. Further, using both structured and unstructured features (classifiers 1 + 2), achieved similar results as classifier 1.

The classifiers 3 and 4 were only applied to identifying "C" vs. "NC" states (1,343 samples). By using the automated labels generated based on structured vital signs, in classifier 3 we observed a decrease in accuracy (0.54) compared to using manual labels in classifier 1, but the recall score improved compared to all other classifiers. This could be partly due to the disagreement between manual and automated labels. The overall Cohen-kappa-score for agreement between these labels was 0.53. This might be due to the fact that mentions of disease status in the clinical notes are concluded by the physician based on the observations on the overall state of the patient, including structured measurements, prescribed medications, patient-specific goals for the disease, other active conditions, or even social factors. For example, we observed many candidate snippets where the hypertension targets (blood pressure thresholds) have been set differently from the standard thresholds in the clinical guidelines. However, the agreement between labels was more consistent for "NC" (0.66) than "C" (0.42) samples, which might be the reason for higher recall and F1 scores in classifiers 2 and 3.

In the co-training method, we first used classifier 1 to label and filter "NA" samples in the unlabeled data set. Then, using co-training we increased the size of the labeled data set (from 1343 to 2300) by adding 957 new samples (834 "C" and 123 "NC") that were consistently labeled by both classifiers 1 and 2. However, the F1-score when the classifier 1 was trained on co-trained labeled data did not change compared to when it was trained on the original manually labeled data set. Our analysis of the learning curve of classifier 1 showed that the model training was saturated by the 1343 training samples, and so the new labels did not provide further improvement.

IV. CONCLUSION

We presented an NLP-based cognitive system for automated extraction of disease status from electronic health records. A chronology of disease status instances can provide disease progression over time and a measure of treatment performance to assist physicians in clinical decision making. Our system extracts disease-related evidence from both unstructured clinical notes and structured patient record data to infer the disease status at each visit. We investigated the use of semi-supervised techniques for generating machine learning models to decrease the need for manual ground truth generation. Automated label generation based on structured data decreased the overall accuracy of our system because of disagreements between structured and unstructured data. Co-training using unstructured and structured models showed potential feasibility for automatically increasing the size of training data in clinical text classification tasks, but did not improve the performance. Further investigation into methods for automated generation of labeled data and feature selection are the subject of future work.

REFERENCES

[1] J. Feblowitz, et al., "Summarization of clinical information: A conceptual model," *Journal of Biomedical Informatics*, vol.44, no. 4, pp. 688-699, 2011.

[2] R. Pivovarov and N. Elhadad, "Automated methods for the summarization of electronic health records," *JAMIA,* vol. 22, no. 5, pp. 938-947, 2015.

[3] M. Devarakonda, et al., "Problem-oriented patient record summary: an early report on a Watson application," *IEEE 16th Int. Conf. on e-Health Networking, Applications and Services (Healthcom),* pp. 281-286, 2014.

[4] M. Devarakonda and C. Tsou, "Automated Problem List Generation from Electronic Medical Records in IBM Watson," *Innovative Applications of Artificial Intelligence (IAAI- 15)*, pp. 3942-3947, 2015.

[5] S.M. Meystre, et al., "Congestive heart failure information extraction framework for automated treatment performance measures assessment," *JAMIA*, ocw097, Jul. 2016.

[6] H. Yang, et al., "A text mining approach to the prediction of disease status from clinical discharge summaries," *JAMIA*, vol.16, no.4, pp.596-600, 2009.

[7] R. Kirk and S. M. Harabagiu, "A flexible framework for deriving assertions from electronic medical records," *JAMIA*, vol. 18, no. 5, pp. 568-573, 2011.

[8] "High Blood Pressure Facts," Centers for Disease Control and Prevention; http://www.cdc.gov/bloodpressure/facts.htm.

[9] P. A. James, et al., "2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8)," *JAMA,* vol. 311, no.5, pp. 507-20, 2014.

[10] T. Mikolov, et al., "Efficient Estimation of Word Representations in Vector Space," *Proc. of Workshop at ICLR*, *arXiv preprint arXiv:1301.3781*, 2013.

[11] "NLP Research Data Sets," Informatics for Integrating Biology and the Bedside (i2b2); https://www.i2b2.org/NLP/DataSets/Main.php.

[12] "Word2vec," Google, https://code.google.com/archive/p/word2vec/.

[13] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *ACM Conf. on Computational Learning Theory*, pp. 92-100, 1998.