MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231033

NLP-Assisted Differential Diagnosis of Chronic Obstructive Pulmonary Disease Exacerbation

Fatemeh SHAH-MOHAMMADI [a,1] and Joseph FINKELSTEIN [a]
[a] *Icahn School of Medicine at Mount Sinai, New York, NY, USA*

Abstract. Chronic Obstructive Pulmonary Disease (COPD) frequently coincides with other comorbidities such as congestive heart failure, hypertension, coronary artery disease, or atrial fibrillation. The exhibition of overlapping sets of symptoms associated with these conditions prevents early identification of an acute exacerbation upon admission to a hospital. Early identification of the underlying cause of exacerbation allows timely prescription of an optimal treatment plan as well as allows avoiding unnecessary clinical tests and specialist consultations. The aim of this study was to develop a predictive model for early identification of COPD exacerbation by using the clinical notes generated within 24 hours of admission to the hospital. The study cohort included patients with a prior diagnosis of COPD. Four predictive models have been developed, among which the support vector machine showed the best performance based on the resulting 80% F1 score.

Keywords. Chronic obstructive pulmonary disease (COPD), NLP, differential diagnosis

1. Introduction

Chronic obstructive pulmonary disease (COPD) is associated with increased mortality, morbidity, and high cost to health systems. COPD is expected to become the fifth leading cause of disability by 2030 [1]. Smoking history and exposure to environmental hazards, including air pollution, are considered the main COPD risk factors. COPD has been shown to be associated with significant comorbidities such as congestive heart failure, coronary artery disease, or atrial fibrillation [2]. Patients with poorly controlled COPD usually suffer from exacerbations presented as a sustained worsening of symptoms. Increased frequency of exacerbation episodes leads to faster lung function decline, quality of life reduction, and increase in the number of admissions into hospital and, accordingly, healthcare costs. When these patients are initially admitted to an emergency department (ED), they exhibit a host of symptoms associated with general respiratory distress as well as other symptoms whose presentation overlaps with a number of acute and chronic conditions. Early identification of the underlying cause of exacerbation allows timely prescription of an optimal treatment plan as well as allows avoiding unnecessary clinical tests and specialist consultations. Thus, conducting differential diagnosis in these patients early at admission is essential for understanding the underlying cause of the exacerbation and identifying further treatment course.

[1] Corresponding Author: Fatemeh Shah-Mohammadi, email: fateme.sh.mohammadi@gmail.com.

Patient medical records in hospitals include a significant amount of unstructured data, such as physician's notes, discharge summaries, and radiology reports. Since the free text is an important part of patient records, considering it in the predictive analysis is equally important. Due to the inherent value of the information present in these documents, and the fact that a manual review of free text records is a very time-consuming process, there is growing interest in developing Natural Language Processing (NLP) pipelines to extract such information from clinical texts. This task may be challenging due to the ambiguity and variations in language used for describing patients' conditions, history, and treatment. To describe a patient's condition, user-specific terminologies, abbreviations, and acronyms are often used. Many providers may exhibit differing styles and terminologies for defining a patient problem, encounter, or clinical course. Due to the variation and complexity of such unstructured information, a framework which can standardize the information by converting this unstructured data into a structured form is required.

The goal of this study was to investigate the feasibility of NLP in the construction of a classification model for differential diagnosis of COPD exacerbation using the clinical notes generated within 24 hours of admission to the hospital. To increase the accuracy of the model, model features (predictors) need to be selected carefully. Unstructured clinical notes are an important source of differential indicators characterizing the exacerbation. Various NLP tools exist to extract information (also referred as named entities) from clinical notes, including Clinical Text Analysis and Knowledge Extraction System (cTAKES) [3], MataMap/MetaMap [4], and Clinical Language Annotation, Modeling and Processing (CLAMP) [5]. While MetaMap and cTAKES are general-purpose NLP systems, CLAMP provides an integrated development environment with GUIs for users who need to build customized NLP pipelines for their individual applications. In this study, we used CLAMP to extract information from the clinical notes and use it as features in the development of the classification model.

2. Methods

The study cohort was formed by a manual review of patient charts and confirming the diagnosis of COPD. The cohort contained patients' socio-demographic information along with different notes documented during their admission. All patients in the cohort were admitted to the emergency department (ED) and then, depending on the severity of symptoms and clinical presentation, hospitalized for further treatment. A different number of notes have been generated from admission to discharge depending on a clinical course of a particular patient. The notes documented in ED included ED triage/intake notes, ED provider notes, ED progress notes, ED event notes, and ED disposition decisions, out of which the ED triage notes and ED provider notes were among the notes generated at the early stage of the admission. Table 1 shows the overall composition of the analytical dataset. The study patients were admitted to the ED due to a variety of reasons based on admitting diagnoses such as pulmonary embolism, congestive heart failure, COPD exacerbation, and shortness of breath (SOB). We used the discharge diagnosis from discharge summaries to categorize the patient admission as definitively resulting from a COPD exacerbation or not. A domain expert reviewed the discharge summaries and labeled the study patients accordingly. Thus, all study patients were assigned to one of two classes: the class labeled as 1 contained the notes for the

patients who were discharged with a final diagnosis of COPD exacerbation, while another class (labeled as 0) included the patients' notes with respiratory distress who were discharged due to a condition other than COPD.

Depending on the physical condition of the patient and the complexity of the visit, a different number of notes have been documented in ED. For building an early classification model to distinguish between the two classes, we only considered the very first notes documented right after admission to the ED, which comprised ED triage note and ED provider notes. The triage note described the reason for the patient's visit, including specific symptoms and incidents. The ED provider note was used for providing documentation of the patient assessment throughout the emergency department visit. It contained various information about the patients, including the very first vital signs observed in ED, such as blood pressure and respiratory rate.

Table 1. Composition of the analytical dataset.

Labels	Number of patients (n=80)	Percentage (%)
1- discharge diagnosis of COPD exacerbation	35	44%
0- discharged with a non-COPD diagnosis	45	56%

2.1. NLP pipeline

To build an NLP pipeline in this study CLAMP was used as an entity extraction tool. This tool follows pipeline-based architecture composed of multiple NLP components. These components are either machine learning-based or dictionary-based components. The list of CLAMP's components includes sentence boundary detection, tokenizer, part-of-speech tagger, section header identifier, abbreviation reorganization and disambiguation, named entity recognition, UMLS encoder, and rule engine. CLAMP is currently available in two versions: CLAMP-CMD (a command line NLP system) and CLAMP-GUI, which provides a GUI for building customized NLP pipelines. After selecting the components of a pipeline, users can click each component to customize its settings. For dictionary-based named entity recognition components, users can specify their own dictionary files. For machine learning–based named entity recognition, users can swap the default machine learning model with models trained on local data. Depending on the components used in structuring the pipeline, the output of CLAMP can contain the start, and the end point of the word (or sequence of words) detected as an entity within the text, the semantic tag associated with it, Concept Unique Identifier (CUI) number (along with RX-Norm code for the entities tagged as drug), assertion, and the actual text extracted as an entity. Negation and temporality are resolved by embedded CLAMP functionality.

2.2. Feature Extraction

The predictive features in this study were extracted from both structured data and unstructured notes. These features have been later used to train the predictive model. Predictive features comprised social determinants of health (SDH) and vital signs indicative of respiratory distress, such as respiratory rate (RR), oxygen saturation (SpO2), pulse rate, and systolic and diastolic pressure. SDH included sociodemographic variables and substance misuse profile previously shown as one of the important morbidity risk

factors in COPD patients. The vital signs, i.e. SpO2 level, RR, pulse rate, and blood pressure (systolic and diastolic pressure), were extracted from the provider note using a dedicated NLP pipeline. To extract these values, we developed a pipeline that used CLAMP's default clinical named entity recognition toolkit, passed for every patient's provider notes through the CLAMP internal engine, and then extracted the values for four vital signs by focusing on the entities tagged as "test" and "lab value". For example, since the entity "SpO2" is tagged as "test" and the value for SpO2 is tagged as "lab value" in CLAMP's output, the value for this vital sign can be easily extracted.

Substance misuse profile included smoking, alcohol, and drug abuse history. A separate pipeline was developed to identify the smoking, alcohol, and drug abuse status of the patients, leveraging rule-based components of the CLAMP-GUI. This pipeline utilized three tags as follows: "smoking status", "drug status," and "alcohol status" using a dictionary developed by manual review of all patients' provider notes and finding all verbiage and spelling variations of phrases that could be considered as an indication for patients' smoking, alcohol, and drug abuse status. We extracted information regarding these three metrics by passing patient's provider notes through the pipeline. Regarding the smoking status, patients were categorized into four groups: "former smoker", "current smoker", "never smoked," and "undetermined". For the alcohol and drug status, patients were categorized into three groups: "yes" (indicating that the patient consumes alcohol or has a drug misuse problem), "no" indicating the absence of these problems, and "undetermined" shows that the note doesn't contain any mention for these conditions. The accuracy and F1-score for this pipeline on smoking status were 75% and 75%, on alcohol status were 85% and 85%, and on drug status were 70% and 70%, respectively.

Patient's age, race, and gender have been extracted as structured data elements from EHR. For all patients, the triage notes were also processed by CLAMP to identify most common documented symptoms presented during admission across classes. For the exploratory data analysis, the summary statistics of all predictive features were generated. Random forest, naïve Bayes, support vector machine (SVM), and logistic regression machine learning models were deployed. We used standard scaling to normalize features. We used 5-fold cross-validation to train and evaluate the performance of the models. All analyses were performed in Anaconda Jupyter Notebook, using Python 3.8. The project has been approved by the institutional review board.

3. Results

Considering Triage notes, the most frequent reasons for the patient visit in both classes were shortness of breath (SOB) and COPD. Cough and wheezing were the next most common symptoms among the patients diagnosed with COPD exacerbation (class 1), and for the patients in class 0, chest pain and cough were the next frequent symptoms.

Summary statistics for the smoking, drug, and alcohol status is as follow: in both classes, 60% of patients were former smokers. In class 1, 30% of patients were current smokers, whereas in class 0 - only 23%. In class 1, only 9% consumed alcohol, whereas in the class 0 - 37%. The proportion of patients with substance misuse in class 1 was 24%, while it was 46% in class 0. The percentage of females in class 0 was 37%, while it was above 50% in class 1. The percentage of Asians and Whites in class 0 was almost twice higher than in class 1. The average age in class 1 and class 0 was 73 and 69 years old, respectively. Black/African Americans accounted for 43% and 36% in class 0 and class 1, respectively. We developed four predictive models based on naïve bayes, logistic

regression, random forest, and SVM machine learning methods. F1-score and accuracy were used as the evaluation metrics. Evaluation results are presented in Table 2. SVM resulted in the highest accuracy and F1-score (81% and 80%, respectively).

Table 2. Summary of evaluations.

Models	F1-Score	Accuracy
Naïve Bayes	0.46	0.47
Logistic Regression	0.52	0.48
Random Forest	0.74	0.76
Support Vector Machine (SVM)	0.80	0.81

4. Discussion

Analysis of the patient's triage notes revealed that chest wheezing, cough, and poor appetite were the most frequent symptoms among the patients that were eventually discharged with COPD exacerbation. These findings are in good agreement with the current clinical diagnostic strategies to differentiate COPD exacerbation from other patient comorbidities. The percentage of patients who were current smokers was higher among the patients diagnosed with COPD exacerbation which is well aligned with typical COPD presentation. Among the four developed machine learning models in this study, SVM resulted in the highest F1-score and accuracy, which is sufficient in order to be used in future diagnostic decision support tools.

5. Conclusions

This paper developed a machine learning model to expedite differential diagnosis of COPD exacerbation leveraging NLP to extract features from clinical notes documented right after the hospitalization. Utilizing predictive features, comprising social determinants of health and vital signs measured right after admission, four predictive models have been developed. The support vector machine-based model showed the best performance and resulted in 80% F1-score. We concluded that NLP-driven automated classification of admission notes is a promising vehicle for conducting differential diagnosis in COPD patients early at admission.

References

[1] Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 2006 Nov;3(11):e442, doi: 10.1371/journal.pmed.0030442.
[2] Wedzicha JA, Seemungal TA. COPD exacerbations: defining their cause and prevention. Lancet. 2007 Sep;370(9589):786-96, doi: 10.1016/S0140-6736(07)61382-8.
[3] Savova GK, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct.
[4] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36, doi: 10.1136/jamia.2009.002733.
[5] Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, Xu H. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc. 2018 Mar;25(3):331-6, doi: 10.1093/jamia/ocx132.