Co-Occurring Evidence Discovery for COPD Patients using Natural Language Processing

Christopher Baechle[1], Ankur Agarwal[1], Ravi Behara[1], Xingquan Zhu[1]
[1]Florida Atlantic University, Boca Raton, Florida

Abstract— **Chronic Obstructive Pulmonary Disease (COPD) is a chronic lung disease that affects airflow to the lungs. Discovering the co-occurrence of COPD with other diseases and symptoms is invaluable to medical staff. Building co-occurrence indexes and finding causal relationships with COPD can be difficult because often times disease prevalence within a population influences results. A method which can better separate occurrence within COPD patients from population prevalence would be desirable. Natural Language Processing (NLP) methods are used to examine 64,371 deidentified clinical notes and discover associations between COPD and medical terms. A co-occurrence score is presented which can penalize scores based on term prevalence. The maximum improvements in recall for symptoms and diseases were 0.212 and 0.130. The maximum improvements in precision for symptoms and diseases were 0.303 and 0.333.**

Keywords—Data Mining; Knowledge Discovery and Management; Decision Support Systems;

I. INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is a family of diseases associated with reduced airflow to the lungs. Over time, patients will experience decreasing airflow as well as increasing inflammation of the tissues that line the airway. The National Institutes of Health (NIH) estimates that approximately 24 million Americans have COPD, with many not even being aware [1]. Though the disease primarily affects smokers, COPD may also occur in those genetically predisposed or due to air pollution. COPD has no known cure.

COPD is characterized by chronic cough with sputum production and increasing shortness of breath [2]. This means that COPD often co-occurs with other lung diseases and diseases caused by smoking. However, many diseases that co-occur with COPD are not contained within the family of COPD diseases. For example, hypertension often co-occurs with COPD because smoking increases the risk for both diseases [3]. Other diseases such as asthma may also affect the lungs and have a high co-occurrence with COPD.

Currently, no standard set of ground-truth terms exists for evaluating the performance of COPD co-occurrence analysis. The contribution of our work is (1) proposed methodology and manual creation of an expert reviewed dictionary and (2) proposition of new mathematical formulas for finding COPD related terms. After the ground truth dictionary has been created, it is evaluated using precision and recall against traditional methods for finding disease and term co-occurrence.

A. Natural Language Processing

The past several years have seen an increase in the electronic storage of patient records using Electronic Health Records (EHR). Data is typically stored in two formats: structured and unstructured. Structured data is stored in a form which can be directly queried and results returned as a normalized data structure. Structured data includes ICD-10 codes and patient demographics.

Clinical notes written by medical professionals during the treatment and discharge of patients are considered unstructured data. These notes often contain more information than ICD-10 codes because they are dictated for the purpose comprehensive patient documentation rather than billing purposes [4]. Medical professionals can quickly dictate large amounts of unstructured information without the conversion losses a structured system would incur. However, these notes have the disadvantage that they cannot be easily queried. The field which processes such unstructured data, or natural language, is known as Natural Language Processing (NLP).

Our research attempts to create a computational framework for the discovery of co-occurring diseases and symptoms in COPD patients. COPD was chosen because it is tangential to many lung diseases. Clinical notes are used as the primary data source due to a potentially high yield of information. Several NLP techniques are employed in this framework in an effort to maximize the information captured within these notes. Although an increasing number of researchers are using NLP with clinical notes as a data source [5], [6] , few have explored COPD clinical notes [7].

B. Clinical Natural Language Processing Tools

The Clinical Text Analysis and Extraction System (cTAKES) represents the latest advancements in clinical NLP. The project began as a cooperative effort between IBM and the Mayo Clinic for the annotation of diseases, medications, laboratory, and anatomical locations in clinical notes [8]. cTAKES is built on top of IBM's Unstructured Information Management Architecture (UIMA). Key to UIMA are the concepts of annotators and the Common Analysis System (CAS). Annotators are code written by system users which analyzes documents and attempts to record structure.

Although cTAKES and UIMA provide useful features, both tools are designed to be used document-at-a-time. This limits the use in document aggregation. Analyzing the frequency of disease occurrence in a document corpus would not be possible with a UIMA annotator and any annotations which use frequency counts would need system extensions.

978-1-5090-4179-4/17/$31.00 ©2017 IEEE

Our research makes use of Apache cTAKES and has written the code necessary to annotate document aggregations.

The Unified Medical Language System (UMLS) is a set of medical dictionaries maintained by the National Library of Medicine (NLM). Many diseases contain variations in spelling, abbreviations, and acronyms for the same disease or symptom. UMLS offers mappings between these variations to a common ID known as a Concept ID (CID). cTAKES offers the normalized form of medical terms using UMLS CIDs.

Zeng et al. have created a system to assist in the detection of co-morbidities in clinical notes [7]. This system primarily uses the Health Information Text Extraction System (HITEx) to assist in the finding of co-morbidities. The existence of COPD and another disease in a clinical note is considered a comorbidity. This methodology is common in the determination of co-occurring diseases. However, this methodology may not be ideal as diseases which occur with high prevalence in a general population will statistically also co-occur with high frequency independent of COPD status. Ideally, penalizing diseases and symptoms which occur with high frequency in a general population would allow a more accurate picture. While such penalizations have been greatly researched in the Information Retrieval (IR) community [9], few have attempted to adapt these methods to clinical NLP [10].

II. Methodology

The data used for this study is comprised of 64,371 deidentified patient discharge summaries. 8.94% of these contain COPD as either a primary diagnosis or contributing factor. The average clinical note contains 10.8 disease/disorder or symptom mention. Discharge summaries span six years of collection.

A. Co-Occurrence Evidence Discovery (COED)

As previously mentioned, our work requires extensions to cTAKES which allows for processing entire document sets rather than the document-at-a-time architecture currently employed. Co-Occurrence Evidence Discovery (COED) represents our research and implementation of these extensions. Fig. 1 describes an overview of our architecture.

Aggregator – After cTAKES annotation is complete, a notification message is sent to the aggregator. The aggregator stores annotations in a MongoDB instance for fast dictionary lookups. cTAKES can take several seconds to annotate a single clinical note. To reduce time consuming duplication of work, new notes can be added without re-annotation of existing notes. Only those components subsequent to aggregation must then be re-run.

Analyzer – A hash table containing all annotated clinical notes is then sent to the analyzer, which looks for co-occurring terms. The analyzer subsequently maps co-occurring terms with their corresponding UMLS definitions to then be delivered to the scoring mechanism.

Score – The scoring mechanism then scores each term using equations and parameters outlined in the next section.

Ranker – Scores are then ranked and recombined with UMLS definitions for user accessible output.

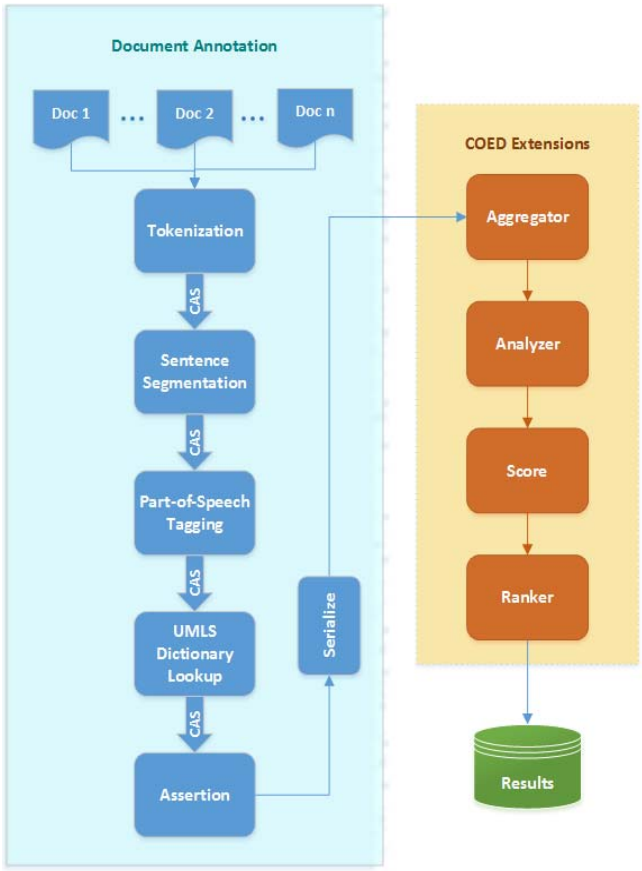

Fig. 1. Block diagram of COED system.

B. Co-occurrence Score

Co-occurrence of diseases and symptoms with COPD is traditionally calculated as follows and serves as our baseline co-occurrence method.

$$f_{COPD}(t, D_{COPD}) = \frac{|\{d \in D_{COPD} : t \in d\}|}{|D_{COPD}|} \quad (1)$$

Where D_{COPD} is the set of documents containing COPD as a diagnosis and t is the term which co-occurrence is to be calculated. This measurement however, prefers terms which occur frequently in the corpus. For example, research shows arthritis to have a great deal of co-occurrence with COPD [11]. However, arthritis tends to have a great deal of co-occurrence with many diseases as it occurs in 1 in 5 American adults [12]. The primary causes of both diseases are different and risk factors largely independent. Terms which appear often in the document corpus should therefore be penalized, as shown in eq. (3).

$$f_{all}(t, D) = \frac{|\{d \in D : t \in d\}|}{|D|} \quad (2)$$

$$f(t, D, D_{COPD}) = \frac{f_{COPD}(t, D_{COPD})}{f_{all}(t, D)} \quad (3)$$

As the frequency of the term increases in the corpus of documents, the co-occurrence is penalized. This can be helpful in the discovery of terms unique to COPD. However, this will also give a great amount of co-occurrence weight to rare diseases only found in COPD patients. In many cases, a more desirable result would be a lower weighting of COPD specific terms. Adding a parameter for the penalization of rare terms follows.

$$f(t, D, D_{COPD}) = \frac{f_{COPD}(t, D_{COPD})^{\lambda}}{f_{all}(t, D)} \quad (4)$$

Many variants of this score are possible. The variant primarily used in this research looks at COPD vs non-COPD documents instead of COPD vs all documents. $D_{\overline{COPD}}$ is defined as the set of documents which do not contain COPD as a primary or contributing diagnosis. $\lambda = 2$ is used for experimentation.

$$f(t, D_{\overline{COPD}}, D_{COPD}) = \frac{f_{COPD}(t, D_{COPD})^{\lambda}}{f_{\overline{COPD}}(t, D_{\overline{COPD}})} \quad (5)$$

C. Evaluation

In order to analyze the performance of retrieved results, a ground truth dictionary of terms was created. 107 diseases and 46 symptoms were chosen using evidence based approaches. Terms which were directly related to COPD such as bronchitis and cough were chosen. Terms which contain weak associations with common risk factors were not chosen. Smoking is known to exacerbate many diseases such as kidney disease by hardening arteries and reducing blood flow to organs. However, smoking is not the primary cause of kidney disease therefore kidney disease not chosen. Table I contains a sample of disease and symptom ground truth terms.

TABLE I. SELECTION OF GROUND TRUTH TERMS

Disease/Disorders	**Symptoms**
Chronic lung disease	Distressed breathing
Bullous emphysema	Wheezing
Pulmonary congestion	Smoking
Bronchitis	Chest pains
Acute respiratory failure	Cough
Asthmas	Reflux
Gastro esophageal reflux	Crackle
Carcinoma of lung	Clubbing (morphologic abnormality)
Pneumonia	Carbon dioxide, increased level
Congestive Heart Failure	Deficiencies, Oxygen

Precision and recall were used at the primary performance metrics. Relevant terms are those defined in the ground truth dictionary and retrieved terms are those found by using both baseline and COED methods. The number of relevant terms is fixed for each category of medical terms. However, the number of retrieved terms is varied where $10 \leq n \leq relevent\ terms$ and $n \in \mathbb{Z}$. Precision and recall are defined in eq. (6) and eq. (7).

$$precision = \frac{|\{relevent\ terms\} \cap \{retrieved\ terms\}|}{|\{retrieved\ terms\}|} \quad (6)$$

$$recall = \frac{|\{relevent\ terms\} \cap \{retrieved\ terms\}|}{|\{relevent\ terms\}|} \quad (7)$$

III. RESULTS

A. Diseases/Disorders

A sample of the highest scoring terms is shown in Table II. The baseline method shows diseases which have a high population prevalence (such as diabetes), to occur higher in the baseline method than COED. Additionally, respiratory failure is a more appropriate highest rank term than hypertension. Precision and recall additionally are higher in comparison to the baseline method as shown in Fig. 2 and Fig. 3.

TABLE II. TOP 10 RESULTS FOR DISEASES & DISORDERS

Baseline	**COED**
Hypertension	Respiratory failure
Diabetes mellitus	Hypertension
Coronary disease	Pneumonia
Heart fibrillation	Congestive heart failure
Arteriopathic disease	Diabetes mellitus
Congestive heart failure	Chronic respiratory failure
Pneumonia	Acute respiratory distress
Respiratory failure	Acute chronic respiratory failure
Anemia	Chronic respiratory insufficiency
Kidney disease	Heart Fibrillation

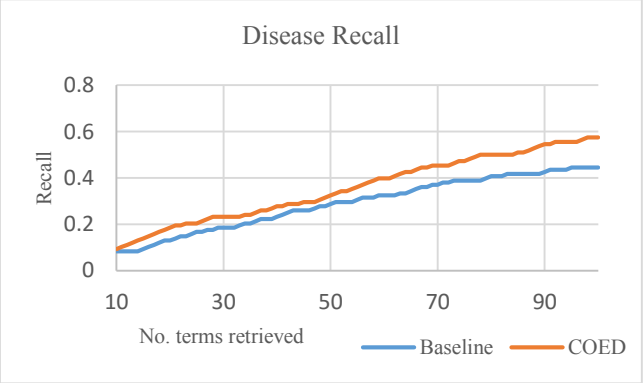

Fig. 2. Baseline vs COED disease recall.

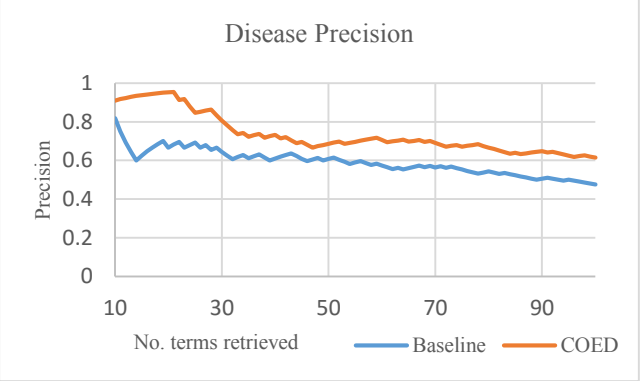

Fig. 3. Baseline vs COED disease precision.

B. *Symptoms*

A sample of the highest scoring terms is shown in Table III. The baseline method returns the top scoring term as pain while COED returns a breathing condition. Additionally, COED returns smoking, a direct known cause of COPD, in the top results. Precision and recall additionally are higher in comparison to the baseline method as shown in Fig. 4 and Fig. 5.

TABLE III. TOP 10 RESULTS FOR SYMPTOMS

Baseline	COED
Pain NOS	Dyspneas
Dyspneas	Oxygen supply
MG body	Wheezings
Normal skin	Pain NOS
Chest pains	MG body
Cough	Respiratory insufficiency
Allergies	Smoker
Arterial tension	Decreased air entry
Edema	Cough
Atrial fibrillations	Normal skin

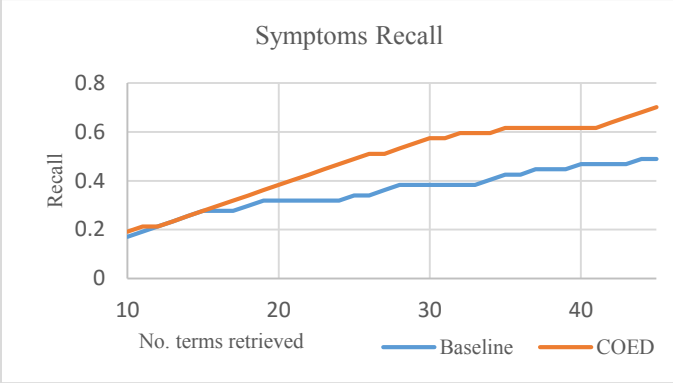

Fig. 4. Baseline vs COED symptom recall.

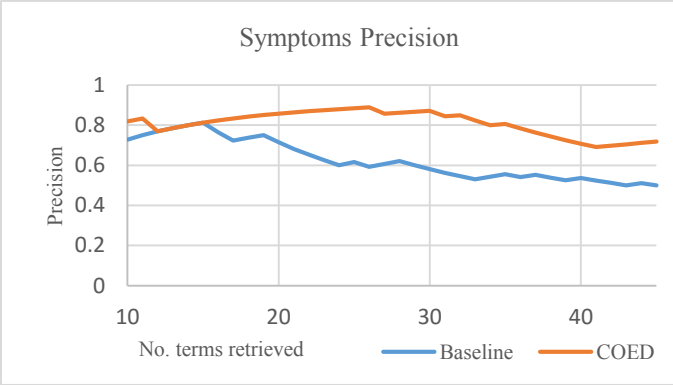

Fig. 5. Baseline vs COED symptom precision.

IV. Conclusion

As shown in the results, penalizing terms which are highly frequent in the corpus results in better precision and recall performance. Penalizing frequently occurring terms gives a better picture of the diseases and symptoms co-occurring with COPD. Using a mathematical and computational approach rather than purely expert driven approach, large dictionaries of COPD related terms can be assembled in a short amount of time. Additionally, localized data may return slightly different results based on patient population. This allows dictionaries to be created on a per-hospital basis rather than nationally, which may not account for localized concerns.

Future work intends to expand this methodology to other diseases to increase confidence in results. Many diseases do not contain ground truth dictionaries for the purposes of information retrieval analysis and must be created using similar methodology. Finally, we intend to integrate the software into an EHR system directly for analytical feedback to medical professionals about their patient population. This can serve as a decision support system to assist medical staff in developing patient treatment procedures.

V. Acknowledgments

This work was supported by NSF grants IIP-1444949 and IIP-1624497.

VI. References

[1] American Lung Association, "COPD Fact Sheet," 2014. [Online]. Available: http://bit.ly/1rOoy1i. [Accessed: 05-Aug-2016].

[2] T. L. Petty, "The history of COPD Early historical landmarks," Int. J. COPD, vol. 1, pp. 3–14, 2006.

[3] A. Marengoni, D. Rizzuto, H. X. Wang, B. Winblad, and L. Fratiglioni, "Patterns of chronic multimorbidity in the elderly population," J. Am. Geriatr. Soc., vol. 57, no. 2, pp. 225–230, 2009.

[4] S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson, "Data from clinical notes: a perspective on the tension between structure and flexible documentation.," J. Am. Med. Inform. Assoc., vol. 18, no. 2, pp. 181–186, 2011.

[5] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?," J. Biomed. Inform., vol. 42, no. 5, pp. 760–772, 2009.

[6] Y. Wu, J. C. Denny, S. T. Rosenbloom, R. A. Miller, D. A. Giuse, and H. Xu, "A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries.," AMIA Annu. Symp. Proc., vol. 2012, pp. 997–1003, 2012.

[7] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system.," BMC Med. Inform. Decis. Mak., vol. 6, p. 30, 2006.

[8] G. K. Savova, J. J. Masanz, P. V Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," J. Am. Med. Informatics Assoc., vol. 17, no. 5, pp. 507–513, 2010.

[9] J. Ramos, J. Eden, and R. Edu, "Using TF-IDF to Determine Word Relevance in Document Queries," Processing, 2003.

[10] S. T. Wu, H. Liu, D. Li, C. Tao, M. A. Musen, C. G. Chute, and N. H. Shah, "Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis.," J. Am. Med. Inform. Assoc., vol. 19, no. e1, pp. e149-56, 2012.

[11] WebMD, "COPD Comorbid Conditions: Heart Disease, Osteoporosis, and More." [Online]. Available: http://wb.md/2dGwUqq. [Accessed: 01-Aug-2016].

[12] CDC, "Addressing the Nation's Most Common Cause of Disability At A Glance 2015," 2015. [Online]. Available: http://bit.ly/1FKbR7i. [Accessed: 01-Aug-2016].