CrossMark

# RESEARCH Open Access

# Big data driven co-occurring evidence discovery in chronic obstructive pulmonary disease patients

Christopher Baechle\*, Ankur Agarwal and Xingguan Zhu

\*Correspondence: cbaechle@fau.edu Department of Computer & Electrical Engineering and Computer Science, College of Engineering, Florida Atlantic University, Boca Raton, FL, USA

#### **Abstract**

**Background:** 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, symptoms, and medications 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. Large hospital systems may potentially have tens of millions of patient records spanning decades of collection and a big data approach that is scalable is desirable. The presented method, Co-Occurring Evidence Discovery (COED), presents a methodology and framework to address these issues.

**Methods:** Natural Language Processing methods are used to examine 64,371 deidentified clinical notes and discover associations between COPD and medical terms. Apache cTAKES is leveraged to annotate and structure clinical notes. Several extensions to cTAKES have been written to parallelize the annotation of large sets of clinical notes. A co-occurrence score is presented which can penalize scores based on term prevalence, as well as a baseline method traditionally used for finding co-occurrence. These scoring systems are implemented using Apache Spark. Dictionaries of ground truth terms for diseases, medications, and symptoms have been created using clinical domain knowledge. COED and baseline methods are compared using precision, recall, and F1 score.

**Results:** The highest scoring diseases using COED are lung and respiratory diseases. In contrast, baseline methods for co-occurrence rank diseases with high population prevalence highest. Medications and symptoms evaluated with COED share similar results. When evaluated against ground truth dictionaries, the maximum improvements in recall for symptoms, diseases, and medications were 0.212, 0.130, and 0.174. The maximum improvements in precision for symptoms, diseases, and medications were 0.303, 0.333, and 0.180. Median increase in F1 score for symptoms, diseases, and medications were 38.1%, 23.0%, and 17.1%. A paired t-test was performed and F1 score increases were found to be statistically significant, where p < 0.01.

**Conclusion:** 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, symptoms, and medications 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.

**Keywords:** Big data, Decision support system, Data mining, Health informatics



#### 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.

Historically, many medical professionals diagnosed COPD as chronic bronchitis or emphysema. More recently, diseases characterized by chronic cough with sputum production and increasing shortness of breath are encompassed in the blanket diagnosis of COPD [2]. This means that COPD often co-occurs with related lung diseases. 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. Additionally, many medications not specifically created for COPD treatment are highly correlated with COPD. Aspirin has been shown to help in the treatment of COPD patients and is often prescribed by medical professionals [4].

The discovery of co-occurring diseases, symptoms, and medications can be useful to researchers and medical professionals. Researchers have shown interest in developing clinical guidelines which consider multimorbidities [5, 6]. Indexes exist which measure the likelihood of patient death based upon which diseases are present [7]. Building these indexes using a computational approach by algorithmically discovering co-occurring diseases, symptoms, and medications would greatly expand their accuracy and coverage.

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 and big data computational framework 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.

# **Background**

#### Clinical data sources

The past several years has 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. ICD-10 codes are commonly used for billing purposes and may not include a complete picture of the patient. Additionally, these billing codes may be assigned based on financial reimbursement incentives rather than accuracy [8].

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 [9]. 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).

### **Clinical NLP**

The medical domain has been an early adopter of NLP. Unlike many domains which use general NLP techniques, clinical NLP systems attempt to incorporate domain knowledge in order to increase the quality of extracted information. A common practice in NLP is stemming, which attempts to find the root of a word. This technique can potentially increase the quality of extracted data, but general techniques such as the Porter stemmer have limited use in the clinical domain [10]. Variants of common English terms often contain no more than an additional prefix or suffix. Clinical term variants may be medication brand names of the same generic drug or diseases which may have different common names internationally. Additional concerns exist such as extracting numeric values and units of laboratory results, identification of family histories vs patient history, and negation of findings. These concerns have led to the development of domain specific clinical NLP systems spanning back many decades.

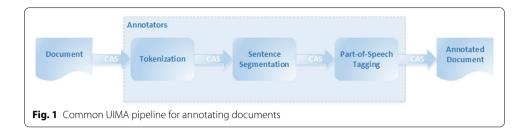
# LSP-MLP

The Linguistic String Project—Medical Language Processor (LSP—MLP) is the oldest traceable NLP system directed at information extraction from clinical notes. LSP began in 1965 for the purposes of developing an English language parser that could process scientific literature [11]. Researchers attempted to structure text in a way that could be easily queried and was an early question—answer system. The NIH funded an expansion of LSP to be applied to clinical notes. This research resulted in the MLP system. MLP was aimed at supplementing LSP with domain specific knowledge to increase the quality of information. LSP—MLP does not appear to be under active development or maintenance and has not seen any related research papers published in over a decade.

Many of the NLP techniques used by LSP–MLP have been superseded by researchers outside of the clinical domain. Modern approaches typically use statistical methods for the identification of parts of speech, sentence boundaries, and other structure. LSP attempts to structure language using grammars. Modern systems additionally make use of controlled medical vocabularies which are developed independently by organizations such as the National Library of Medicine (NLM). Additionally, LSP–MLP is fragile because it depends on the syntactic structure of the text rather than semantic meaning.

### MedLEE

The Medical Language Extraction and Encoding System (MedLEE) originated as a system to structure radiology reports. MedLEE eventually evolved as a general purpose clinical NLP system [12]. MedLEE was created without the help of additional NLP libraries and frameworks. Modern clinical NLP systems often use Commercial off the shelf (COTS) NLP tools such as UIMA, GATE, and OpenNLP to prevent duplication of efforts. MedLEE has improved in recent years to include medical vocabularies from the National Library of Medicine's Unified Medical Language System (UMLS). UMLS contains medical terms from many data sources in a standardized format.



#### Apache cTAKES

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 to annotate diseases, medications, laboratory, and anatomical locations in clinical notes [13]. 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. Figure 1 shows a typical UIMA pipeline.

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. Annotations which use frequency counts would need system extensions. Our research makes use of Apache cTAKES and has written the code necessary to annotate document aggregations.

# HITEx

The Health Information Text Extraction (HITEx) is an information extraction system aimed at general purpose processing of medical texts. The system departs from previous works as it uses a component based architecture based on the GATE framework [14]. The use of GATE allowed researchers to focus solely on domain specific concerns rather than low level tasks such as tokenization and sentence segmentation. This project has stalled and not published new research since 2006.

The highest layer of HITEx is the UMLS concept mapper which maps medical terms to UMLS concepts. This subsystem uses both exact string matching and fuzzy matching through truncation and normalization. The system has been used successfully as a hybrid system combining ICD-9 structured data and unstructured clinical notes. In addition to basic NLP tasks, HITEx contains modules capable of discerning the patient's primary diagnosis and smoking status.

Zeng et al. [14] have created a system to assist in the detection of co-morbidities in clinical notes. This system primarily uses 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. 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, medications, 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 [15], few have attempted to adapt these methods to clinical NLP [16].

# Clinical big data

Data mining of clinical data has been well studied since the emergence of the field. However, big data approaches have been far less studied. A review by Herland et al. [17] documents several big data systems. Current clinical big data systems are often built for the purpose of supervised machine learning tasks. In practice, many medical professionals use Clinical Decision Support Systems (CDSS). These allow the practitioner to make conclusions rather than relying on algorithmic classification. Many classification algorithms provide evidence which is difficult to interpret. Medical professionals may be uncomfortable diagnosing patients without clearly interpretable results. Aggregated data, summary statistics, and similar patients are examples of common information presented to medical professionals using a CDSS. Big data approaches in CDSS have seen little research to date.

Our research attempts to create a computational framework for the discovery of cooccurring diseases, medications, 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. With the emergence of electronic patient record databases, many large systems containing big data are now available. Examples of these are the Veterans Affairs (VA) hospital system and England's National Health Service (NHS). Our methodology uses a big data approach to finding co-occurring evidence and is validated using a dataset containing approximately 64,000 instances. Due to rarity of access to databases as large as the VA system, this dataset was the largest available to our research group. However, the methodology was designed with big data techniques as the foundation. It can be employed by organizations such as the VA hospital system without scalability issues. Although an increasing number of researchers are using NLP with clinical notes as a data source [18, 19], few have explored COPD clinical notes [14] and there is no documented evidence in Google Scholar of this methodology applied to big data.

# Methodology

The Apache Hadoop ecosystem is leveraged for COED. Hadoop Distributed Filesystem (HDFS) is used for the storage and distribution of deidentified patient discharge summaries. Apache Spark is utilized for MapReduce operations and the pyspark python interface is used for programming. Documents are represented as Resilient Distributed Datasets (RDD). Apache cTAKES is used for the extraction of medical terms from unstructured clinical notes. cTAKES offers several UIMA pipelines and UMLS fast-dictionary-lookup is used as the primary pipeline. Disease, medication, and symptom annotations are stored, excluding annotations marked "history" and those that have been negated. UMLS Concept IDs (CID) are extracted from each annotation and used as the primary term identifier.

The data used for this study is comprised of 64,371 deidentified patient discharge summaries. 0.0894 of these contain COPD as either a primary diagnosis or contributing

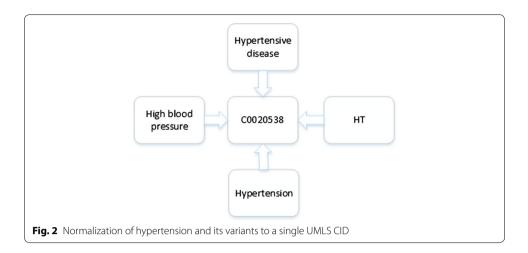
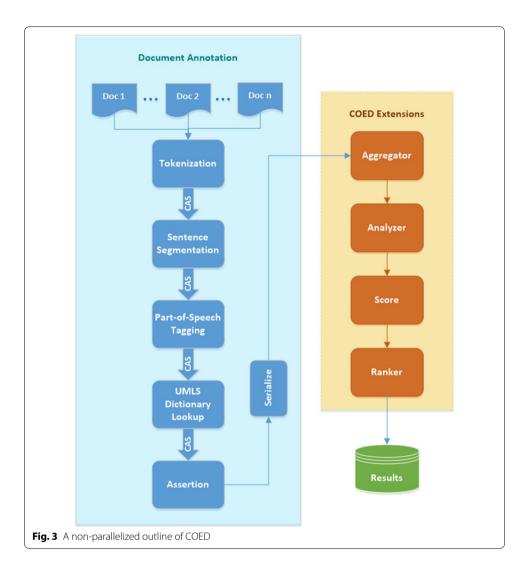


Table 1 Summary of dictionaries used by cTAKES

Dictionary	Description
Snomed-CT [20]	Snomed-CT is a set of clinical terms maintained by the International Health Terminology Standards Development Organization (IHTSDO). The development of Snomed-CT dates back to 1965 and is known for its comprehensive coverage of clinical terms. SNOMED-CT consists of concepts, descriptions, and relationships and can be used for semantic processing
NCI Thesaurus [21]	The National Cancer Institute (NCI) Thesaurus was created to assist in research systems made available by NCI. It covers clinical terminology regarding cancers, findings, drugs, therapies, anatomy, genes, and many other cancer research related terms. NCI thesaurus offers a partial model as to how these subjects relate to each other and aims to provide a common system for cancer researchers to communicate
MeSH [22]	Medical Subject Headings (MeSH) is an NLM controlled vocabulary used for indexing articles on NIH's pubmed. Additionally, relationships between terms are provided which can act as a thesaurus
ICD-9 [23, 24]	International Classification of Diseases (ICD) is a coding system designed for classification of diseases and disorders. ICD is maintained by the World Health Organization (WHO) and ICD-9 is the ninth revision of the system. In the United States, ICD-9 has seen popular usage in medical billing. The system has been adopted by many organizations, including the centers for disease control for reporting mortality and morbidity statistics [25]

factor. The average clinical note contains 20.6 disease, symptom, or medication mentions. Discharge summaries span 6 years of collection. Diseases and medications may have several spelling variations and abbreviations in common usage. Figure 2 shows several variations for hypertensive disease normalized to a common CID. Normalization to a common ID is necessary as variations in the same disease will lower aggregate counts. Splitting aggregate counts of diseases will result in diseases with lower than appropriate representation in search rankings.

The UMLS dictionaries used in this research are Snomed-CT, NCI Thesaurus, MeSH, and ICD-9. Table 1 outlines a brief description of each dictionary. Many terms are identified in multiple dictionaries and there exists a large amount of redundancy in coverage. UMLS CIDs are valid across dictionaries and it is possible to normalize a term discovered in multiple dictionaries to single CID. The use of multiple dictionaries assists in expanding coverage of common abbreviations and variants in spelling.

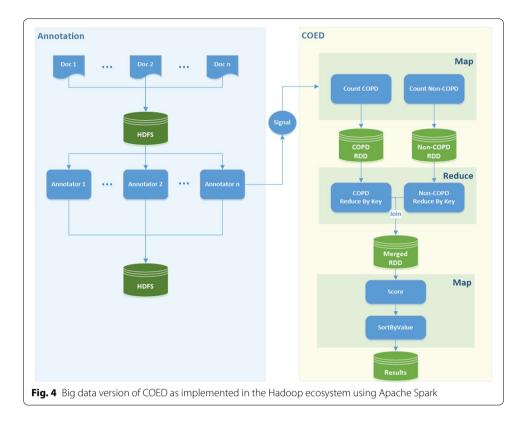


# Co-occurrence evidence discovery framework

An overview of COED is described in Fig. 3. When a document arrives for annotation, it passes through the Apache cTAKES pipeline. The pipeline begins with generalized NLP tasks such as tokenization before reaching clinical NLP tasks such as UMLS dictionary lookup. After the document is annotated, it is serialized and held until all document annotations for the dataset are complete. COED then runs corpus-at-a-time processing using the following components.

Aggregator—Gathers annotations into a single data file suitable for processing. In the Hadoop ecosystem, this is a tab separated file with one line per document.

Analyzer—Documents are counted and terms mapped to COPD and non-COPD lookup tables. Each document is considered a COPD document if COPD was the primary diagnosis. Terms contained within the same document are considered to be COPD terms and counts incremented within the lookup table. Documents that do not contain COPD as a primary diagnosis are mapped to the non-COPD lookup table in a similar fashion. Terms may exist in both tables. These term counts are later used with Eq. 5 to calculate COED scores.



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.

A big data prospective of COED is contained in Fig. 4. Raw documents are stored in HDFS. Extensions to cTAKES have been written which allow cTAKES to run as a daemon. The daemon waits for a signal to annotate a new document. This signal is delivered using ZeroMQ, a message queue system. Due to the high initial startup time of cTAKES (up to several minutes), this daemon is separately installed on slave servers and runs persistently. cTAKES annotators annotate a single document and store the resulting information in HDFS. Each annotator processes a single document at a time, but annotators may run independently because annotations exist independently of other documents.

After all annotations are complete, annotations are aggregated to a spark compatible file and a signal is sent to the second phase. COED runs as a series of map and reduce tasks. Word counts are performed for COPD and non-COPD documents. Results are stored in two separate RDDs and reduced by key using the *add* callback function. The two RDDs are then joined using the pyspark join() method creating a merged RDD of form [K,(V1,V2)]. Each term is then scored using Eq. 5 and then sorted by value using a custom sort function. No final reduce operation is required as the previous reduce has ensured distinct keys. Results are then outputted to a file for further analysis.

### Score

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

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

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 [26]. However, arthritis tends to have a great deal of co-occurrence with many diseases as it occurs in one in five American adults [27]. 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.

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

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

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

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

# **Evaluation**

In order to analyze the performance of retrieved results, a ground truth dictionary of terms was created. 107 diseases, 62 medications, and 46 symptoms were chosen using evidence based approaches. Many criteria were considered when selecting medical terms. Terms which were directly related to COPD such as bronchitis and cough were chosen. Additionally, terms which may not be directly associated with COPD, but have strong common risk factors, such as smoking 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 2 contains a sample of disease, medication, and symptom ground truth terms.

Table 2 Selection of ground-truth terms

Disease/disorders	Symptoms	Medications
Chronic lung disease	Distressed breathing	Spririva
Bullous emphysema	Wheezing	Advair
Pulmonary congestion	Smoking	Oxygen
Bronchitis	Chest pains	Albuterol
Acute respiratory failure	Cough	Combivent
Asthmas	Reflux	Prednisone
Gastro esophageal reflux	Crackle	Atrovent
Carcinoma of lung	Clubbing (morphologic abnormality)	Medrol
Pneumonia	Carbon dioxide, increased level	Duoneb
Congestive heart failure	Deficiencies, oxygen	Daliresp

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  $10 \le n \le r$  elevent terms and  $n \in \mathbb{Z}$ . Precision and recall are defined in Eqs. 6 and 7.

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

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

# Results

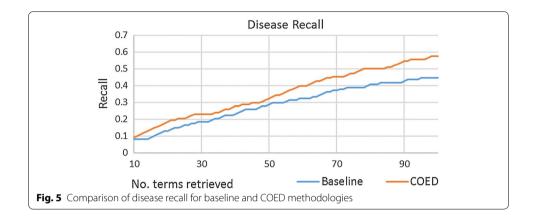
# Diseases and disorders

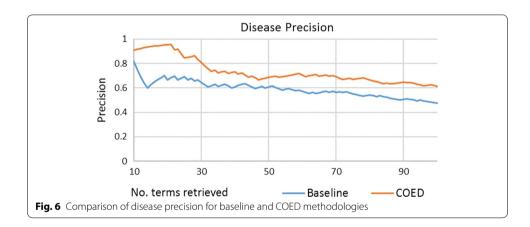
A sample of the highest scoring diseases is shown in Table 3. The baseline method shows diseases which have a high population prevalence (such as diabetes), to occur higher in the baseline method than COED. Additionally, respiratory failure is a more appropriate highest rank term than hypertension. Diseases with a high population prevalence may still rank high in COED. For example, diabetes ranks as the fifth highest term. However, the goal of COED is not to completely eliminate frequently occurring diseases from retrieval results. COED aims to rank them lower by penalizing their prevalence in the general population.

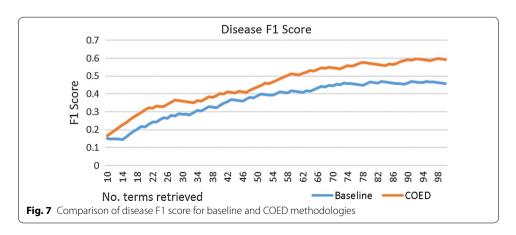
Table 3 Selection of top ten results for diseases and disorders

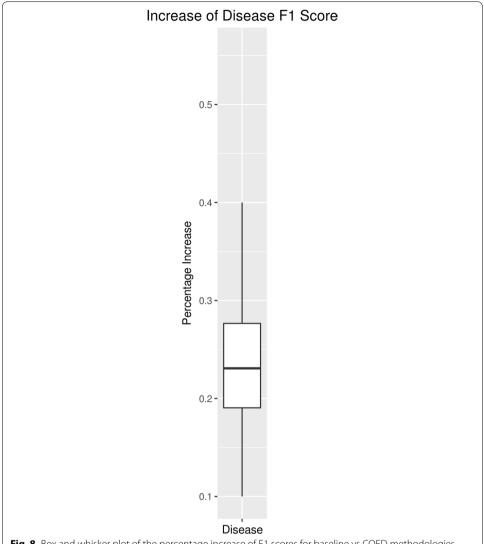
Rank	Baseline	COED
1	Hypertension	Respiratory failure
2	Diabetes mellitus	Hypertension
3	Coronary disease	Pneumonia
4	Heart fibrillation	Congestive heart failure
5	Arteriopathic disease	Diabetes mellitus
6	Congestive heart failure	Chronic respiratory failure
7	Pneumonia	Acute respiratory distress
8	Respiratory failure	Acute chronic respiratory failure
9	Anemia	Chronic respiratory insufficiency
10	Kidney disease	Heart fibrillation

Precision and recall were higher in comparison to the baseline method as shown in Figs. 5 and 6. As the number of terms increased, the difference in recall slightly increased. The difference in precision was greatest when the number of terms was low. Additionally, F1 score for COED was higher than baseline, as shown in Fig. 7. A paired t test was performed against F1 scores for COED and baseline. The resulting p value was p < 0.01, showing a statistically significant increase in F1 score for COED. Finally, the percentage increase of F1 score is shown in Fig. 8 as a box and whisker plot. The median increase was for COED F1 score was 23.0%, a considerable increase.









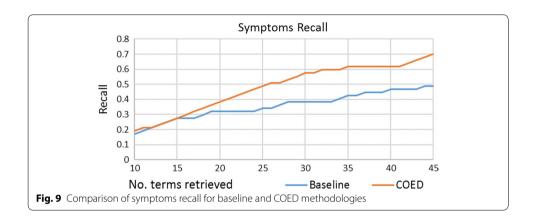
**Fig. 8** Box and whisker plot of the percentage increase of F1 scores for baseline vs COED methodologies. Median percentage increase for disease F1 score is 23.0%

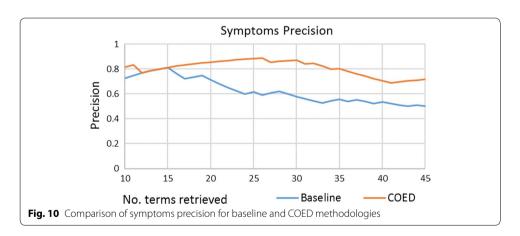
# **Symptoms**

A sample of the highest scoring symptoms is shown in Table 4. The baseline method returned the top scoring term as pain while COED returned a breathing condition. Additionally, COED returned smoking, a direct known cause of COPD, in the top results. Allergies are very common and appear in the baseline methodology but do not appear in the selection of COED results. Precision and recall were generally higher in comparison to the baseline method as shown in Figs. 9 and 10. However, in some instances, these two metrics were equal. The difference in precision and recall were generally largest for a high number of terms retrieved. F1 score results shown in Figs. 11 and 12 additionally show similar increases. The median F1 score for symptoms was 38.1%, another considerable increase. A paired t-test was performed against F1 score results and p < 0.01.

Table 4	Selection of	fton ten res	ults for symptoms	
Iable T	Jeiection o	LOD LEILIE3	uita ioi ayiiiptoiiis	

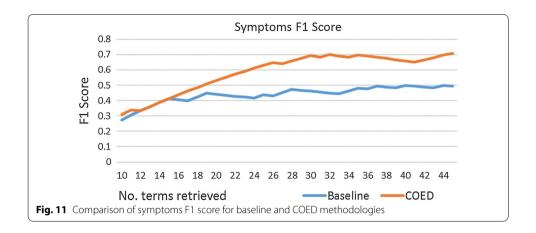
Rank	Baseline	COED
1	Pain NOS	Dyspneas
2	Dyspneas	Oxygen supply
3	MG body	Wheezings
4	Normal skin	Pain NOS
5	Chest pains	MG body
6	Cough	Respiratory insufficiency
7	Allergies	Smoker
8	Arterial tension	Decreased air entry
9	Edema	Cough
10	Atrial fibrillations	Normal skin

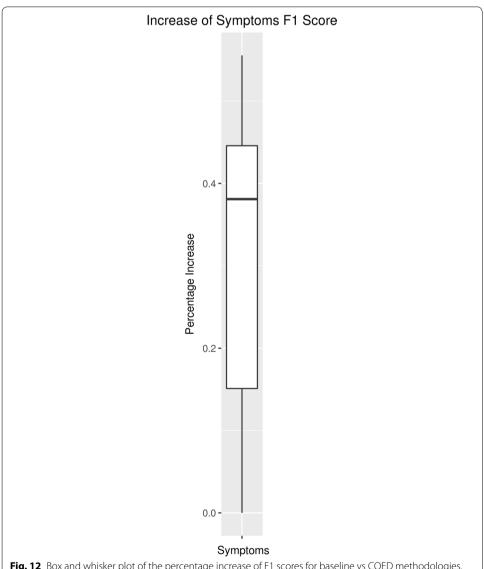




# Medications

A sample of the highest scoring medications is shown in Table 5. The baseline method has chosen Aspirin as the highest scoring medication. However, Aspirin is a very common medication and population prevalence causes it to be highly ranked. COED has chosen Spriva, a popular medicine to treat bronchospasms caused by COPD. In contrast to diseases and symptoms, COED and baseline methods are much more similar in precision and recall value, as shown in Figs. 13 and 14. F1 score is similar for less than 26 terms and median increase only 17.1%, as shown in Figs. 15 and 16. These increases are

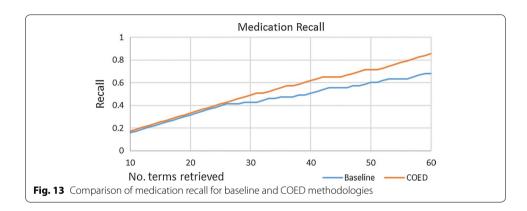


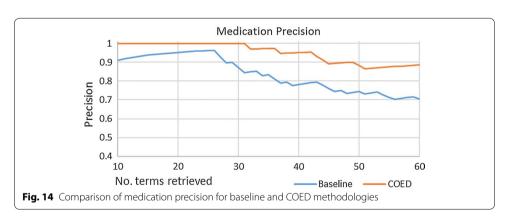


**Fig. 12** Box and whisker plot of the percentage increase of F1 scores for baseline vs COED methodologies. Median percentage increase for symptoms F1 score is 38.1%

Table 5 Selection of top ten results for medications

Rank	Baseline	COED
1	Aspirin	Spiriva
2	Albuterol	Advair
3	Oxygen	Oxygen
4	Advair	Albuterol
5	Prednisone	Combivent
6	Marevan	Prednisone
7	Lisinopril	Atrovent
8	Medrol	Medrol
9	Combivent	Duoneb
10	Spiriva	Daliresp

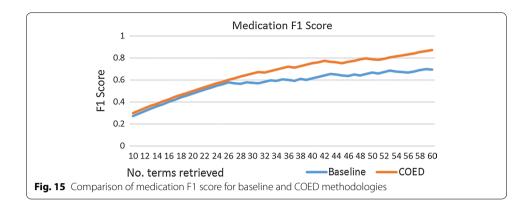


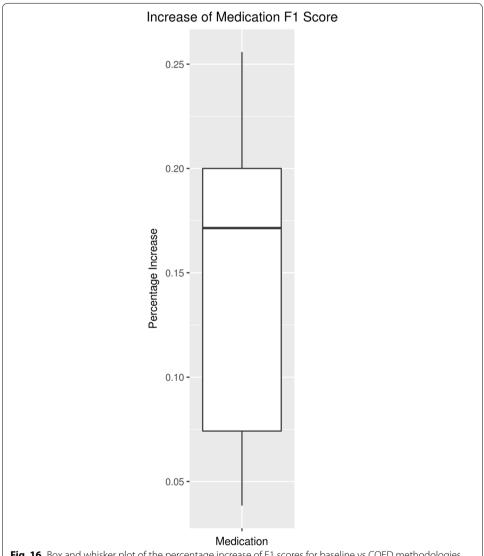


smaller than disease and symptoms. However, the increases were still found to be statistically significant at p < 0.01.

# **Conclusions**

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, symptoms, and medications co-occurring with COPD. Using a mathematical and computational approach rather than purely expert driven





**Fig. 16** Box and whisker plot of the percentage increase of F1 scores for baseline vs COED methodologies. Median percentage increase for medication F1 score is 17.1%

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.

#### Authors' contributions

CB carried out the conception, design, and implementation of this research as well as interpretation of results. AA and XZ made substantial contributions to the conception and design of this research as well as critically reviewing and interpreting results. CB carried out the drafting of manuscript. AA and XZ critically reviewed the manuscript. All authors read and approved the final manuscript.

#### Acknowledgements

Not applicable.

#### **Competing interests**

The authors declare that they have no competing interests.

#### Funding

This research was supported in part by NSF Grants IIP-1444949 and IIP-1624497.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 4 January 2017 Accepted: 15 March 2017

## Published online: 04 April 2017

# References

- 1. American Lung Association. COPD Fact Sheet, 2014. http://bit.ly/1rOoy1i. Accessed 05 Aug 2016.
- 2. Petty TL. The history of COPD early historical landmarks. Int J COPD. 2006;1:3–14.
- 3. Marengoni A, Rizzuto D, Wang HX, Winblad B, Fratiglioni L. Patterns of chronic multimorbidity in the elderly population. J Am Geriatr Soc. 2009;57(2):225–30.
- Aaron CP, Schwartz JE, Hoffman EA, Tracy R, Austin JHM, Smith LJ, Jacobs DR, Watson KE, Barr RG. Aspirin use and longitudinal progression of percent emphysema on CT: the MESA lung study. Am J Respiration Crit Care Med. 2015;191:A6354.
- Guthrie B, Payne K, Alderson P, McMurdo MET, Mercer SW. Adapting clinical guidelines to take account of multimorbidity. Br Med J. 2012;345:e6341.
- 6. Tinetti ME, Fried TR, Boyd CM, Badalà F, Nouri-mahdavi K, Raoof DA. Designing health care for the most common chronic condition—multimorbidity. JAMA. 2012;307(23):2493–4.
- 7. D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in outcome assessment: the Charlson comorbidity index. Methods Inf Med. 1993;32(5):382–7.
- 8. Danielsen RD, Simon AF, Pavlick R. The culture of cheating: from the classroom to the exam room. J Phys Assist Educ. 2006;17(1):23–9.
- 9. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011;18(2):181–6.
- 10. Porter MF. An algorithm for suffix stripping. Program. 1980;14(3):130-7.
- Sager N. Natural language information processing. Advanced Book Program. Boston: Addison-Wesley Publishing Company; 1981.
- 12. Friedman C.A Broad-coverage natural language processing system. Proceeding of the AMIA Symposium. American Medical Informatics Association. 2000; 270–4.
- Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13.
- Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006;6:30.

- Ramos J, Eden J, Edu R. Using TF-IDF to determine word relevance in document queries. Process Manag. 2003;24(5):513–23.
- Wu ST, Liu H, Li D, Tao C, Musen MA, Chute CG, Shah NH. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. J Am Med Inform Assoc. 2012;19(e1):e149–56.
- 17. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1(1):2.
- 18. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42(5):760–72.
- 19. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H. A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. AMIA Annu Symp Proc. 2012;2012:997–1003.
- 20. Ruch P, Gobeill J, Lovis C, Geissbühler A. Automatic medical encoding with SNOMED categories. BMC Med Inform Decis Mak. 2008;8(1):S6.
- 21. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40(1):30–43.
- 22. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88(3):265-6.
- 23. United States Department of Health and Human Services. The international classification of diseases. Geneva: World Health Organization; 1969.
- 24. Slee VN. The International classification of diseases: ninth revision (ICD-9). Ann Intern Med. 1978;88(3):424-6.
- 25. International Classification of Diseases, Ninth Revision (ICD-9). http://www.cdc.gov/nchs/icd/icd9.htm. Accessed 11 Jul 2016.
- WebMD. COPD Comorbid Conditions: heart disease, osteoporosis, and more. http://wb.md/2dGwUqq. Accessed 01
  Aug 2016.
- 27. CDC. Addressing the Nation's most common cause of disability at A Glance 2015. http://bit.ly/1FKbR7i. Accessed 01 Aug 2016.

# Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ▶ springeropen.com