Journal of Artificial Intelligence Research 72 (2021) 429-474 Submitted 01/2021; published 10/2021

NLP Methods for Extraction of Symptoms from Unstructured Data for Use in Prognostic COVID-19 Analytic Models

Greg M. Silverman gms@umn.edu
Himanshu S. Sahoo sahoo009@umn.edu
Nicholas E. Ingraham ingra107@umn.edu
Monica Lupei lupei001@umn.edu
University of Minnesota, Minneapolis, USA 55455

Michael A. Puskarich mike-em@umn.edu
Hennepin County Medical Center, Minneapolis, USA 55415
University of Minnesota, Minneapolis, USA 55455

Michael Usher mgusher@umn.edu
James Dries james.dries32@gmail.com
Raymond L. Finzel finze006@umn.edu
University of Minnesota, Minneapolis, USA 55455

Eric Murray eri81112@fairview.org
M Health Fairview, Saint Paul, USA 55104

John Sartori jsartori@umn.edu
Gyorgy Simon simo0342@umn.edu
Rui Zhang zhan1386@umn.edu
University of Minnesota, Minneapolis, USA 55455

Genevieve B. Melton gmelton@umn.edu
Christopher J. Tignanelli ctignane@umn.edu
M Health Fairview, Saint Paul, USA 55104
University of Minnesota, Minneapolis, USA 55455

Serguei VS Pakhomov pakh0002@umn.edu
University of Minnesota, Minneapolis, USA 55455

Abstract

Statistical modeling of outcomes based on a patient's presenting symptoms (symptomatology) can help deliver high quality care and allocate essential resources, which is especially important during the COVID-19 pandemic. Patient symptoms are typically found in unstructured notes, and thus not readily available for clinical decision making. In an attempt to fill this gap, this study compared two methods for symptom extraction from Emergency Department (ED) admission notes. Both methods utilized a lexicon derived by expanding The Center for Disease Control and Prevention's (CDC) Symptoms of Coronavirus list. The first method utilized a *word2vec* model to expand the lexicon using a dictionary mapping to the Unified Medical Language System (UMLS). The second method utilized the

©2021 AI Access Foundation. All rights reserved.

expanded lexicon as a rule-based gazetteer and the UMLS. These methods were evaluated against a manually annotated reference (f1-score of 0.87 for UMLS-based ensemble; and 0.85 for rule-based gazetteer with UMLS). Through analyses of associations of extracted symptoms used as features against various outcomes, salient risks among the population of COVID-19 patients, including increased risk of in-hospital mortality (OR 1.85, p-value $<$ 0.001), were identified for patients presenting with dyspnea. Disparities between English and non-English speaking patients were also identified, the most salient being a concerning finding of opposing risk signals between fatigue and in-hospital mortality (non-English: OR 1.95, p-value = 0.02; English: OR 0.63, p-value = 0.01). While use of symptomatology for modeling of outcomes is not unique, unlike previous studies this study showed that models built using symptoms with the outcome of in-hospital mortality were not significantly different from models using data collected during an in-patient encounter (AUC of 0.9 with 95% CI of [0.88, 0.91] using only vital signs; AUC of 0.87 with 95% CI of [0.85, 0.88] using only symptoms). These findings indicate that prognostic models based on symptomatology could aid in extending COVID-19 patient care through telemedicine, replacing the need for in-person options. The methods presented in this study have potential for use in development of symptomatology-based models for other diseases, including for the study of Post-Acute Sequelae of COVID-19 (PASC).

1. Introduction

Due to the strain and episodic shortage of clinical resources during the COVID-19 pandemic, in some instances clinicians have had to triage patients based on available resources (e.g., ICU beds, ventilators, personal protective equipment, dialysis, etc.) to manage the influx of new cases. To address this challenge, clinicians need accurate tools to allocate essential medical resources to patients who need them most, primarily those at high risk for mortality. Prognostic modeling can help identify patients at higher risk for mortality and better inform clinicians' treatment decisions, thus potentially improving outcomes by allocating resources efficiently.

Patients seeking clinical treatment for COVID-19 usually exhibit symptoms and are concerned about their prognosis and need for medical support. As it pertains to COVID-19, understanding future prognosis is far more critical than the initial diagnosis. When properly constructed, prognostic models can improve patient outcomes by guiding care decisions (Croft et al., 2015; Wynants et al., 2020). Statistical and deep learning prognostic models offer potential answers to the limitations of current outcome predictions in clinical settings. A study from the Wuhan, Hubei, and Guangdong provinces in China showed that a deep learning model could identify COVID-19 patients upon admission who are at greater risk for critical illness. In this study, Liang et al. (2020) constructed a Deep Learning Survival Cox model that performed statistically better than a standard LASSO Cox model using ten variables. However, Liang et al.'s (2020) study limitations included a lack of important predictors, such as lab results and vital signs, and it did not perform as strongly in predicting hospitalization compared to the other models. In another study, Wollenstein-Betech et al. (2020) illustrated that logistic regression and support vector machines (SVMs) accurately predicted hospitalization, mortality, need for ICU, and use of ventilator, with respective area under the receiver operating characteristic curve (AUC) of 0.75, 0.70, 0.64, and 0.86 for each outcome. Their study included models that tracked the disease progression by measuring variables such as the development of pneumonia, which significantly increased the accuracy of the models for ICU admittance, mortality, and ventilator use. This potentially offers real-time prognostic implications.

Despite these advances in prognostic models, there is room for improvement in patient outcomes. Furthermore, all these models utilize structured data, including labs, medications, comorbidities, etc. In this study, we outline methodologies that utilize Natural Language Processing (NLP) for clinical concept mapping to a well-defined lexicon and the Unified Medical Language System (UMLS) to extract information about patient symptomatology exhibited at the time of Emergency Department (ED) presentation. While the presented methodologies may be generalizable to a variety of acute and chronic conditions, we explore the potential incremental benefit of clinical NLP concept extraction specifically to the use-case of symptomatology for acute COVID-19.

2. Background & Significance

The application of NLP-derived symptomatology in clinical practice can provide valuable insight into disease progression for improving treatment decisions. Examples of how symptomatology can improve clinical outcomes are presented below, with specific attention to the COVID-19 pandemic. In order to be successful, however, there are several obstacles that need to be addressed.

2.1 Symptomatology

There are significant gaps in NLP's broad use for various clinical and research applications. In primary care, general practitioners may have difficulty diagnosing illness from a set of common symptoms due to significant overlap between syndromes. For example, fatigue is a common symptom among many diseases and in isolation may not provide significant predictive power towards one diagnosis over another. When evaluated with other presenting symptoms, an overall diagnosis and prognosis of a particular disease becomes much clearer. Additional testing is then required to progress further towards the likely diagnosis and prognosis. However, with advances in computing power and machine-learning algorithms, as well as the ubiquity of the electronic health record (EHR), certain constellations of symptoms may carry significant predictive power that can be overlooked by clinicians; especially non-specific symptoms such as fatigue. Here, evidence-based patient management could be valuable. On this frontline in health care, a renewed emphasis on symptom epidemiology has potential to improve treatment decisions by focusing on evidence-based outcomes, rather than diagnosis (Rosendal et al., 2015). This strategy could simplify the process of differential diagnosis in primary care.

Use of symptoms mapped to UMLS concepts has been studied in the non-COVID-19 domain for use in predictive modeling of colorectal cancer. Hoogendoorn et al. (2016) showed that model performance was significantly better when using UMLS mapped concepts as predictors when mined from general practitioner and consultation notes. In another study, Stephens et al. (2020) developed predictive models for influenza that used UMLS-driven NLP methods to extract symptoms from unstructured notes from ambulatory care visits. This strategy significantly improved data quality by reducing the frequency of false negatives for predicted outcomes and reducing non-random missing data. In this case, the UMLS-driven NLP model outperformed the basic LASSO regression model with an AUC of 0.71. In the area of surgical complication detection, models incorporating NLP-derived surgical site infection text indicators accelerated the manual process of detecting surgical site infection for data abstractors (Skube et al., 2020). These examples demonstrate the capacity of symptomatology-based modeling algorithms to improve diagnosis and prognosis.

2.1.1 COVID-19

Numerous studies have examined the manifestations of COVID-19 symptomatology. In a meta-analysis of 14 studies across data from various countries, researchers found there were 32 symptoms across multiple organ systems. These symptoms differ from those released by organizations, including WHO, United States CDC, NHS, China CDC, Institut Pasteur, and Mayo Clinic (Champika et al., 2020). Given the wide range of reported symptoms, clinicians would benefit from knowing which symptoms indicate the greatest risk of severe infection or mortality.

In the United Kingdom the study by Abdulaal et al. (2020) utilized a neural network to create a prognostic model for mortality of COVID-19 patients upon admission. The model specifically used predictors of demographics, comorbidities, smoking history, and symptomatology to predict patient mortality with 86.25% accuracy and an AUC of 90.12%. Relatedly, a study of COVID-19 symptomatology in Wuhan found that fatigue and expectoration showed linear associations with COVID-19 severity, but the study did not have a large sample size and did not examine associations with outcomes such as mortality (Li et al., 2020).

Our study aims to improve and expand on the use of symptomatology in prognostic models with focus on associations between predictors. To achieve this, we evaluated two methods for automated extraction of symptoms from Emergency Department (ED) admission notes: The first used ensembling of UMLS-based NLP classifiers; while the second used a rule-based lexical gazetteer (hereafter "Hybrid COVID Symptoms Gazetteer") that included terms from the UMLS.

2.1.2 Implications

Use of NLP methods for mining of a patient's presenting symptoms has other practical implications. At various times and locations during the pandemic, restrictions on movement, lack of personal protective equipment, or clinic closures made in-person clinical visits difficult for many patients. During and since these challenges, telemedicine has become an increasingly vital tool for patient care, particularly for COVID-19. With a prognostic model based on symptomatology, COVID-19 patients could safely isolate at home and also receive informed treatment from clinicians, replacing the need for other in-person options that might jeopardize the well-being of health care workers and patients.

A meta-analysis by Hincapié et al. (2020) consisting of 43 studies found while some health care systems have implementation issues, the potential benefits of telemedicine expansion are enormous, especially in areas where there are insufficient clinical resources. Another study by Andrews et al. (2020) found there was high satisfaction with telemedicine for both patients and health care workers, and that both groups were open to continuing telemedicine delivery methods after the pandemic. Lastly, telemedicine has the potential to redefine a billable medical service if telemedical consultations prove adept at treating

COVID-19 patients. Improvement of remote medical services has potential beyond COVID-19, especially for communities lacking a robust health care infrastructure.

2.1.3 Obstacles

An important consideration is that patient symptoms are not always part of the structured EHR, and thus not readily available for near real-time analysis (Pakhomov et al., 2008). This can be addressed through use of appropriate NLP methods for extraction of symptoms from unstructured clinical text within the EHR, such as ED admission or outpatient (OP) notes.

Another major issue involves inconsistent documentation of symptoms, especially for non-English speakers. Among COVID-19 patients, not speaking English has been found to be independently associated with worse outcomes (Ingraham et al., 2021). A major goal of this study is to show the presented NLP methods have the potential to give insight into this issue by providing data necessary for examining mechanisms behind this association.

2.2 NLP Methods

An overview of NLP methods used in this study is presented below. These methods provide a flexible framework for mining terms and concepts, such as symptoms, that may not be easily accessible to clinicians, with the goal of improving patient care. These methods are described throughout Section 3 with application to the symptomatology of COVID-19.

2.2.1 UMLS and Information Extraction

The UMLS Metathesaurus was developed by the United States' National Library of Medicine (NLM, 2009) as a knowledge-driven resource to identify biomedical concepts that associate multiple terms from multiple source vocabularies into a single concept unique identifier (CUI). The UMLS contains the Semantic Network for mapping CUIs into a set of broad subject categories, or Semantic Groups and Types. The Semantic Network provides consistent categorization of UMLS concepts into high-level groupings (e.g., findings, disorders, procedures, etc.) (Bodenreider, 2020). The UMLS consists of over 4 million distinct concepts organized by 15 semantic groups, 133 semantic types, and 54 semantic relationships (He et al., 2017).

UMLS is designed to allow for a broad range of applications, including standardization of the information extraction (IE) process. Use of UMLS for IE originally focused on identifying biomedical concepts in biomedical literature and later broadened to clinical text (Aronson, 2001; Aronson & Lang, 2010). To utilize UMLS for IE, groups of words in unstructured text that correspond to UMLS concepts must be identified using Named Entity Recognition (NER). Once identified, candidate named entities (NEs) can be mapped to UMLS CUIs and then extracted for conceptual indexing (Reátegui & Ratté, 2018).

Use of NER with mapping to UMLS concepts has been investigated in many clinical studies that leveraged ensembling techniques for combining multiple NLP classifiers to improve overall performance. Finley et al. (2017) assessed medical acronym disambiguation with UMLS lookup using a majority vote ensemble to improve overall performance, and Kuo et al. (2016) compared ensembling methods for extraction using categorized UMLS concepts for improved identification of patient cohorts. Lastly, Bompelli et al. (2020) showed ensembles of output from NLP annotation systems can provide a modest performance increase over individual NLP annotation systems for tasks involving UMLS mapping that were novel for those systems.

2.2.2 Lexical Gazetteer

A gazetteer is a dictionary of terms derived from a given lexicon (a.k.a., vocabulary). Gazetteers that use a defined set of rules (a.k.a., rule-base) for lookup of terms within text have been successfully used in many domains as an alternative to more sophisticated NLP annotation systems for classification, NER and IE. For example, Nguyen and Patrick (2014) demonstrated the use of a gazetteer for classification of radiology reports into reportable and non-reportable cancer cases. Liu et al. (2013) successfully used a gazetteer to select cohorts of heart failure and peripheral arterial disease patients from unstructured clinical notes. Gazetteer lexicons are highly targeted through construction by domain experts, especially when combined with appropriate lexical rules (Elkin et al., 2008) and have been shown to work very well with continuous maintenance (Couto, Campos, & Lamurias, 2017). Furthermore, gazetteer lexicons provide a simpler alternative to UMLS-based NLP pipelines through ease of adaptation and modification of linguistic constructs for improvement in matching of relevant terms (Meystre & Haug, 2005). Gazetteers and their rule-bases can easily be deployed together as a standalone tool using containerization technologies such as Docker, or the rule-base alone can be deployed as part of an existing infrastructure such as that developed by the Open Health NLP (OHNLP) consortium for the National COVID Cohort Collaborative (N3C) based on the work of Wen et al. (2019).

2.2.3 Word Embeddings & Query Expansion

Word embeddings use data-driven techniques to produce distributed semantic representations of words through training over a large corpus of text (Mikolov et al., 2013a). In Mikolov et al.'s (2013b) *word2vec* implementation, words within a corpus are assigned to vector values that represent semantic representations between words. This allows *word2vec* to be used as a tool for identifying other words within the corpus that closely match the vector values of the target set of words. A common measure for assessing similarity between vector embeddings is the cosine distance, where the cosine of the angle between two vectors is computed using linear algebra.[1] Other similarity measures include the Jaccard and Dice measures (Ljubesic et al., 2008). For this study we used the cosine distance measure.

Query expansion is a procedure used in IE, in which semantically similar terms to the original target set of words are identified using models such as *word2vec*. Query expansion helps better identify relevant documents within a corpus, thus increasing overall document retrieval coverage (viz., sensitivity) (Aklouche et al., 2018; Bursi et al., 2006). In two separate studies Silverman et al. (2019) and Tignanelli et al. (2020) showed that simple ensemble methods produced marked increases in both precision and recall compared to individual NLP annotation systems for determining if appropriate treatment was provided in prehospital care. This was achieved by using a predefined clinical lexicon for query expansion using a *word2vec* model trained by Pakhomov et al. (2016). Similarly, Fan et al. (2019) trained a *word2vec* model on clinical notes to expand dietary supplement vocabulary

[1] This determines if the vectors are pointing in the same direction.

by finding corresponding misspellings and brand names, which helped retrieve more relevant documents.

2.3 Our Contribution

This study compared two methods used at our institution for extraction of symptoms derived from a lexicon based on the guidelines listed for Symptoms of Coronavirus published by the United States' Center for Disease Control's division of National Center for Immunization and Respiratory Disease (CDC, 2020). The first used standard ”out-of-the-box” clinical NLP annotation systems to map extracted symptoms to a dictionary of UMLS concepts derived from this lexicon. The second, developed by Sahoo and Silverman (2020) specifically for this study, used the Hybrid COVID Symptoms Gazetteer in tandem with the UMLS. All NLP annotation systems used in this study performed roughly the same in terms of extraction performance in our baseline experiments. However, results indicate the Hybrid COVID Symptoms Gazetteer significantly outperforms the other NLP annotation systems in terms of document processing run-time. Unlike other rule-based gazetteers existing in the literature, the Hybrid COVID Symptoms Gazetteer presented in this study was unique in that it also leveraged the UMLS.

To demonstrate the utility of these methods, extracted symptoms were employed to create features for use in prognostic modeling and analysis of associations. The methodologies presented in this study allow for a more consistent and efficient categorization of symptoms through automation in comparison to manual extraction methods that use clinical abstractors as described by Abdulaal et al. (2020) and others. Furthermore, the Hybrid COVID Symptoms Gazetteer leveraged in this study has potential for use in real-time clinical decision support (CDS). At the time this study was conducted, these methods had not been applied to mining patient symptoms for acute COVID-19, nor have they been presented in detail to allow ease of replication.

This study shows that use of extracted symptoms in clinical analytics can be easily used as a tool to assess patient risk. Furthermore, results of this study indicate patient symptomatology has potential to serve as a proxy for tests involving direct patient contact. Unlike previous studies, the goal of this study is not to show how various models can be improved through use of symptoms, but how use of reported symptoms can be very effective for general clinical decision making.

3. Data & Methods

For this study, we used a NLP pipeline that employed two methods for extraction of presenting symptoms from unstructured ED admission notes using a lexicon derived from the CDC Symptoms of Coronavirus. The pipeline utilized the lexicon, an ensemble of ”out-of-the-box” clinical NLP annotation systems, and a Hybrid COVID Symptoms Gazetteer. Extraction performance across all NLP annotation systems used in this study was compared to a manually annotated reference set of notes (hereafter referred to as ”manually annotated reference”) to assess gaps in the lexicon. Run-time performance between ”out-of-the-box” clinical NLP annotation systems and the Hybrid COVID Symptoms Gazetteer was also assessed. Extracted symptoms were then used as features for examining associations with various clinical outcomes. These methods were rapidly implemented by Silverman et al. (2020) at our institution to address the need for quick extraction methods for obtaining patient presenting symptoms to augment data for a registry of COVID-19 positive patients.

3.1 Clinical Data

Data used in this study were collected from M Health Fairview, which is an integrated healthcare delivery system affiliated with the University of Minnesota, composed of 12 hospitals and services in the ambulatory and post-acute settings. Annually, there are over 368,000 ED visits within the network. Between 9 and 30% of those cases are annually admitted as inpatients.

Between March 1, 2020 and December 17, 2020 there were a total of 19,924 ED visits for 10,110 unique patients confirmed as COVID-19 positive within the M Health Fairview network. All patients that did not opt out of research and were seen in the ED within a window of up to 14 days prior to the return of a positive COVID-19 test result were included in this study. Patients diagnosed as OPs and then presented to an ED following their diagnosis were excluded from this study, since our interest was in evaluating presenting symptoms.[2] For patients with multiple ED visits within this window, we used the first (i.e., index) visit. Only data for patients older than 18 years of age were considered, thus reducing the cohort to 5,006 patients. Use of these data for this study was approved by our Institutional Review Board. Table 1 highlights demographics for the population in this study.

Demographics		**Frequency**
Gender	Male	2,449 (48.92%)
	Female	2,557 (51.08%)
Race	White	2,605 (52.04%)
	Black	942 (18.82%)
	Asian	477 (9.53%)
	Hispanic	427 (8.53%)
	Declined	472 (9.43%)
	Other	83 (1.66%)
Language	English Speaking	3,678 (73.47%)
	Non English Speaking	1,328 (26.53%)
Age (Years)	18-40	1,404 (28.05%)
	41-54	1,025 (20.48%)
	55-64	777 (15.52%)
	65-79	1,084 (21.65%)
	80 or more	716 (14.3%)

Table 1: Demographics of patients participating in the study.

[2] This study was conducted prior to the implementation of rapid Polymerase Chain Reaction (PCR) testing, with results, on average, available 24-72 hours after the test was given in the ED. Thus, these criteria ensured that all patients presenting initially to the ED with COVID-19 were included.

3.1.1 Discrete and Unstructured Data

The following discrete data were made available through the M Health Fairview EHR: demographics, labs, vitals taken at the ED visit, home medications taken for at least 3 months prior to the ED visit, and comorbidities identified using ICD-10 codes, (for architecture, see "Files" in Figure 1). A set of automated scripts performed the following tasks (for architecture, see "Pre-processing/ETL" in Figure 1): data extraction, transformation and loading (ETL) of the data to the COVID-19 patient registry (for architecture, see "COVID-19 Patient Registry" in Figure 1). Unstructured admission notes from the ED were also made available. Notes were processed using the methods described below in Sections 3.3 and 3.4 to extract the symptoms used in the analyses described in Section 3.6 and then added to the COVID-19 patient registry. The COVID-19 patient registry was part of the M Health COVID-19 AI Pipeline, as described by Silverman et al. (2020).

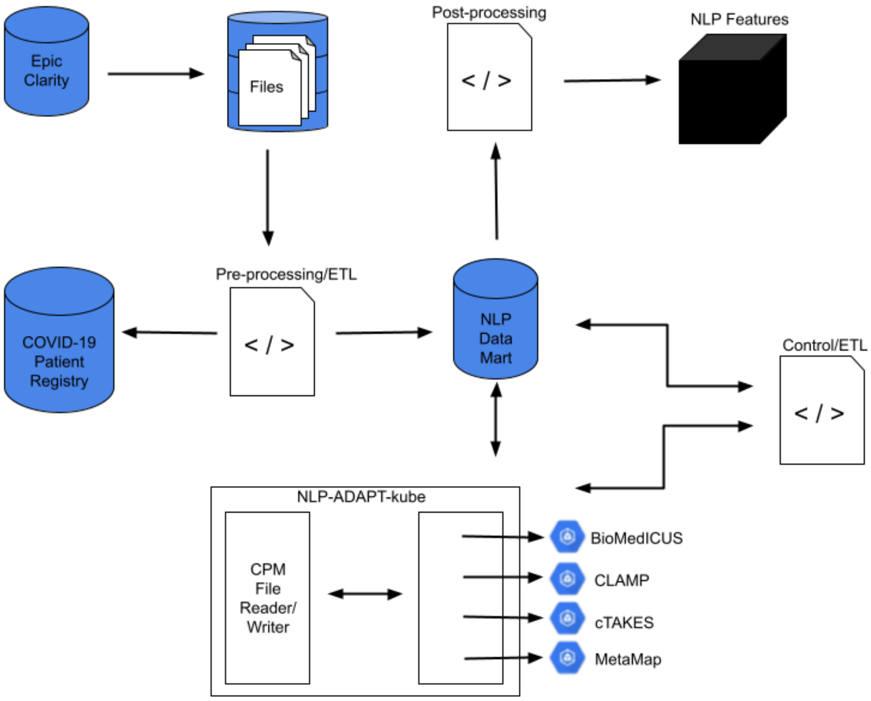


Figure 1: High level view of the architecture of the NLP portion of the M Health COVID-19 AI pipeline.

3.2 CDC Symptoms of Coronavirus

The CDC's (2020) guidelines listed for "People with COVID-19 have had a wide range of symptoms reported – ranging from mild symptoms to severe illness" (hereafter referred to as "CDC's guidelines") were used to create the lexicon employed in this study as described below. As of September 1, 2020, these included:

- ◦ Cough
- ◦ Shortness of breath or difficulty breathing (dyspnea)
- ◦ Fatigue
- ◦ Muscle or body aches (aches)
- ◦ Headache
- ◦ New loss of taste or smell
- ◦ Sore throat
- ◦ Congestion or runny nose
- ◦ Nausea or vomiting
- ◦ Diarrhea
- ◦ Fever or chills

3.2.1 Lexicon of Covid-19 Symptoms

A specialized lexicon of 164 terms based on the CDC's guidelines was created using equivalent medical terminology, abbreviations, synonyms, allied symptoms, alternate spellings and misspellings. This lexicon was iteratively derived by four board certified clinicians (two critical care [CT, ML], one emergency [MP], and one pulmonary fellow [NI]). Terms in this lexicon (hereafter referred to as "164 derived COVID-19 symptoms") were used to extract the symptoms used in our study. Each of the derived 164 terms were clustered under one of the main 11 symptoms in the CDC's guidelines (as outlined in Section 3.2). Please see Online Appendix 1 for the 164 derived COVID-19 symptoms and their clustering strategy (hereafter referred to as "acute CDC symptoms").

3.2.2 Query Expansion of 164 Derived Covid-19 Symptoms

For this study, the *word2vec* model trained by Pakhomov et al. (2016) on a corpus of clinical notes from encounters at M Health Fairview between 2010 and 2014, was used for expansion of the 164 derived COVID-19 symptoms described in Section 3.2.1. This *word2vec* model was trained with embeddings having up to four word sequences by using the word2phrase tool, allowing for query expansion of phrases.

Terms in the list of 164 derived COVID-19 symptoms were mapped to the UMLS using MetaMap to get the CUI and its preferred term.[3]

[3] The following MetaMap options were used for this task:

The UMLS preferred term was used to query for the top 100 semantically similar terms to the preferred term using the Gensim most_similar method (Řehůřek, 2016) to interface with the *word2vec* model.

The query expansion strategy used in this study is described below in Figure 2.

```
For a given parent symptom:
    If parent symptom has UMLS mapping:
        If semantically similar terms exist for parent symptom:
            if cosine distance >= 0.75 AND terms have UMLS mapping:
                Then select terms
            Else:
            If cosine distance < 0.75 and terms have UMLS mapping:
                Then select top two terms with UMLS mapping
        Else:
            If semantically similar terms do not exist for parent symptom:
                Then select the parent term
        Else:
            pass
    Else:
        pass
    next parent symptom
```

Figure 2: Hierarchical rules for query expansion.

The final set of mapped UMLS concepts associated with symptoms was further reviewed by three board certified clinicians (ML, NI, MP) to ensure concepts were appropriately clustered by acute CDC symptoms. This final set of concepts was made available as a dictionary of CUIs for use in the UMLS-based portion of the NLP Pipeline described below in Section 3.3. For a complete catalogue of the UMLS expansion of the clustered acute CDC symptoms used to create this dictionary (hereafter referred to as "UMLS-based symptom cluster") please see Online Appendix 2, and for an example of a query expansion using these rules, please see Appendix B.

3.3 NLP Pipeline: UMLS-based Concept Extraction

The NLP portion of the M Health COVID-19 AI Pipeline (described in Section 3.1.1; hereafter referred to as "NLP Pipeline") was used for annotating and extracting acute CDC symptoms from ED admission notes (for overall architecture, see Figure 1). Two extraction methods were explored: The first described below used symptoms mapped to CUIs using the UMLS-based symptom cluster as described above in Section 3.2.2. The second, described

(a) Evaluation score threshold of 900;

(b) Limited to the set of semantic types: *fndg*, *sosy*, *dsyn*, *patf*;

(c) Query term processing only;

(d) Use of concept gaps; and

(e) Ignore word order (Lang, 2016).

below in Section 3.4, extracted acute CDC symptoms (described in Section 3.2.1) using the rule-base of the Hybrid COVID Symptoms Gazetteer by including additional UMLS terms.

3.3.1 NLP Pipeline: NLP-ADAPT-KUBE

ED admission notes were annotated for CUIs using the NLP Artifact Discovery and Preparation Toolkit for Kubernetes (NLP-ADAPT-KUBE). NLP-ADAPT-KUBE was developed by The Natural Language Processing/Information Extraction Program (the NLP/IE Program) at the University of Minnesota to automate document annotation at medium to large scale (Finzel & Silverman, 2019). NLP-ADAPT-KUBE uses Docker containerization for ease of repeatable deployment and automation of processes, and Kubernetes to control processes across multiple compute nodes (Boettiger, 2015; Dikaleh et al., 2017). For this study, we used the Unstructured Information Management Architecture (UIMA) Collection Processing Management (CPM) file reader implementation of NLP-ADAPT-KUBE, which was designed for medium-scale annotation tasks. This implementation allows each UIMA annotation processing engine to read text files from disk, and to write processed annotations back to disk as UIMA XML Metadata Interchange Common Analysis System (CAS XMI) formatted files (Ferrucci & Lally, 2004).

NLP-ADAPT-KUBE includes the following UIMA-based annotation systems (formerly referred to as ”out-of-the-box” clinical NLP annotation systems): the BioMedical Information Collection and Understanding System (BioMedICUS); the Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP); the Clinical Text Analysis and Knowledge Extraction System (cTAKES); and MetaMap (with UIMA adapter) (Savova et al., 2010; Aronson, 2001; Soysal et al., 2018; Knoll, 2019) (for architecture, see ”NLP-ADAPT-KUBE” in Figure 1).

3.3.2 UMLS Concept Extraction

Methods for mapping of concepts representing COVID-19 symptoms to the UMLS as extracted from ED notes by the UIMA-based annotation systems are presented below, along with a description of the extraction process. Lastly, an overview of how extracted symptoms were transformed into features for use in our statistical analyses is given.

3.3.3 Semantic Type Mapping

When UIMA-based annotation systems in NLP-ADAPT-KUBE provided their own notion of “semantic type,” output was constrained using these system-specific categories. In cases where individual systems provided UMLS Semantic Types but no bespoke semantic groups, we relied on the Semantic Type Mappings and Semantic Groups developed by NLM (2018). This strategy, as illustrated in Figure 3, enabled a comprehensive hierarchical mapping of semantic types to the semantic group Disorders for use in this study. At the top level is the UMLS semantic group Disorders that best aligned with the semantic types available in individual systems found within NLP-ADAPT-KUBE.

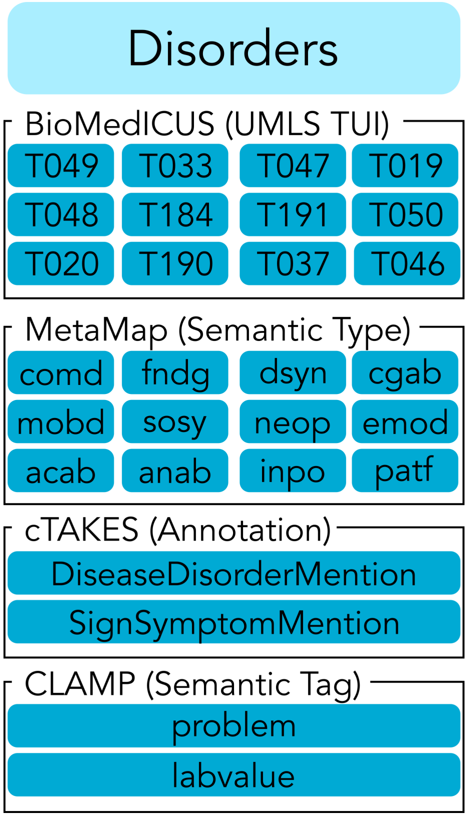


Figure 3: Hierarchical mapping of semantic types to Disorders semantic group to allow for UIMA-based annotation system extraction.

3.3.4 Extraction Overview

ED admission notes were pre-processed to convert all text to ASCII encoding as required by MetaMap and then added to the NLP Data Mart (for architecture, see ”NLP Data Mart” in Figure 1). For this study, default pipelines for all UIMA-based annotation systems were used. UMLS concept annotations were made for both positive and negative mentions at the sentence-level for each ED admission note (see Table 2 for the UIMA annotation types used by each individual UIMA-based annotation system for both positive and negative mentions). Once ED admission notes were annotated and output written back to disk, a custom script developed by Silverman (2020b) that utilized the library dkpro-cassis (developed by Klie and Castilho (2020)) was used to extract and transform each CAS XMI annotation object to Python lists, which were then loaded into a PostgreSQL database (for architecture, see ”NLP Data Mart” and ”Control/ETL” in Figure 1).

Annotations written to the NLP Data Mart were then filtered by the semantic group Disorders using the semantic type mapping described in Section 3.3.3 (for architecture, see ”Post-processing” in Figure 1) and processed to disambiguate UMLS concepts when a given span of text had multiple candidate concepts assigned to it.[4,5] Finally, all disambiguated concepts for a patient ED encounter were aggregated by the particular UIMA-based annotation system, CUI, and negation status (i.e., polarity).

System	Positive Mention	Negative Mention
BioMedICUS	biomedicus.v2.UmlsConcept	biomedicus.v2.Negated
CLAMP	edu.uth.clamp.ClampNameEntityUIMA (where assertion = 'present')	edu.uth.clamp.ClampRelationUIMA (where assertion = 'absent')
cTAKES	org.apache.ctakes.DiseaseDisorderMention & org.apache.ctakes.SignSymptomMention (where polarity = 1)	org.apache.ctakes.DiseaseDisorderMention & org.apache.ctakes.SignSymptomMention (where polarity = -1)
MetaMap	org.metamap.uima.ts.Candidate	org.metamap.uima.ts.Negation

Table 2: UIMA annotation types used in this study; with both positive and negative mentions for each UIMA-based annotation system.

3.3.5 Ensembling of Annotation Output

The NLP-Ensemble-Explorer framework developed by Silverman (2020a) was used with annotations produced by the UMLS-based portion of the NLP pipeline to create an ensemble of annotation output. NLP-Ensemble-Explorer was developed for evaluating NLP annotation system performance for NER and IE.[6,7]

In this study, we used a Boolean combination of UIMA-based annotation system output from NLP-ADAPT-KUBE as an ensemble (hereafter referred to as "Boolean ensemble") to extract UMLS concepts. NLP-Ensemble-Explorer uses the logical $\vee$ operator to represent a union set operation (or $\cup$); and the logical $\wedge$ operator to represent an intersection set operation (or $\cap$).[8] Boolean expressions in NLP-Ensemble-Explorer are represented as a binary tree and evaluated using the parse tree algorithms provided by Miller and Ranum (2013). For an example evaluation of a Boolean combination parse tree please see Appendix C.

For this study, the Boolean ensemble (((BioMedICUS $\wedge$ cTAKES) $\wedge$ MetaMap)$\vee$ CLAMP) was used to merge UMLS concepts extracted by individual UIMA-based annotation systems from ED notes (for architecture, see "Post-processing" in Figure 1). This ensemble was previously optimized for NER with respect to the f1-score in experiments conducted by the NLP/IE Program on a similar manually annotated reference set of notes

[4] The following hierarchical rules were implemented in our pipeline for concept disambiguation: a) Select the candidate concept associated with the longest overlapping span; b) If multiple candidate concepts were assigned to text spanning the same length, then use the highest likelihood score assigned to a candidate concept (*NB*: the only UIMA-based annotation system with no likelihood scoring was cTAKES); c) If there is no longest span of text and no tie-breaker likelihood score, concepts were randomly shuffled with the top concept assigned to the span of text.

[5] *NB*: For the UIMA-based annotation systems used in this study, only MetaMap had word sense disambiguation available for filtering out ambiguous mappings to the UMLS (NLM, 2020)

[6] NLP-Ensemble-Explorer works with any NLP annotation system architecture.

[7] Source code is available here: `https://github.com/nlpie/nlp-ensemble-explorer/blob/polarity/ensemble_explorer/extract_ensemble.py`.

[8] The equivalence of logical and set theoretic operators is given by the rules of logical conjunction and disjunction of sets (Plisko, 2014a, 2014b)

on the Disorders semantic group from M Health Fairview (precision: 0.44, recall: 0.61, f1-score: 0.51) and for the 2010 i2b2 Veteran's Affairs challenge set as described by Uzuner, South, and Duvall (2011) (i2b2) (precision: 0.83, recall: 0.97, f1-score: 0.90). This ensemble also performed almost as well as the top performing ensemble combinations for NER on the Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ) corpus described by Albright et al. (2013) (precision: 0.53, recall: 0.84, f1-score: 0.66).

3.3.6 UMLS-Based Features Derived From Extracted Acute CDC Symptoms

CUIs representing acute CDC symptoms obtained from the ensembled output described in Section 3.3.5 were abstracted from the sentence to the document-level (where each document represents an ED admission note). Each note was labeled accordingly with its polarity to account for whether that CUI had at least one positive (+1) or at least one negative (-1) mention within that note - independent of the specific sentence-level mention. Labeling of each note was thus treated as a multi-label classification task.

The UMLS-based symptom cluster described in Section 3.2.2 was used as a dictionary to map CUIs within a UMLS-based symptom cluster to the set of extracted symptoms associated with each note. Rules used to create features from the acute CDC symptoms were meant to capture the possibility that there were both positive and negative mentions of different symptoms within a given UMLS-based symptom cluster. Thus, if a positive mention of a CUI within a UMLS-based symptom cluster was found within a note in a particular sentence-level mention, a value of 1 would be assigned to that feature (with '_p' as a suffix). Similarly, if a negative mention within the same symptom cluster was found within that note in another sentence-level mention, a value of 1 would be assigned to that feature (with '_n' as a suffix). If there were no positive or negative mentions for all CUIs within a symptom cluster a value of 0 would be assigned appropriately to the '_p' and '_n' features. See Table 3 for example feature output.

cdc_fever_p	**cdc_fever_n**	**cdc_cough_p**	**cdc_cough_n**	**ed_encounter_date**
0	0	1	1	1/3/2020
1	0	0	0	8/20/2020
1	0	0	1	12/1/2020

Table 3: Example UMLS-based features.

The final set of extracted acute CDC symptoms with negation status was made available as features (hereby referred to as "UMLS-based features") for use in the COVID-19 Patient Registry for use in research, analytical applications and CDS (for architecture, see "NLP Features" in Figure 1 and "UMLS-based Features" in Figure 4).

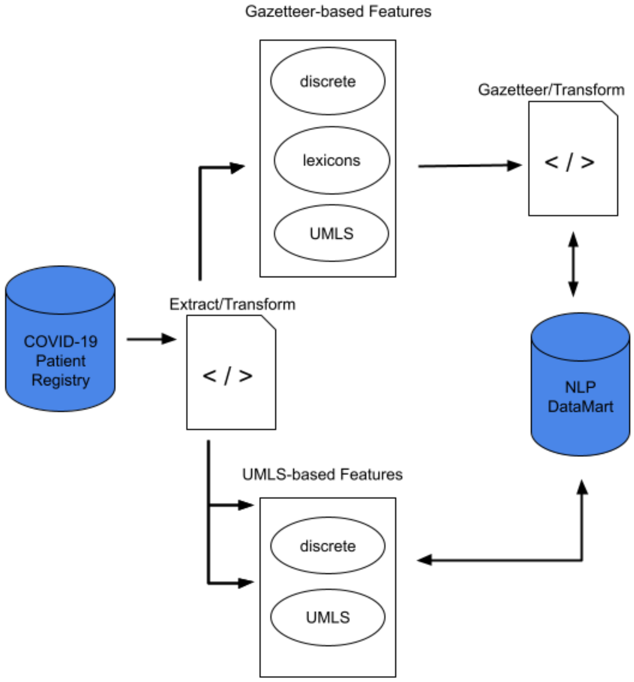


Figure 4: Architecture of NLP feature creation.

3.4 NLP Pipeline Architecture: Hybrid COVID Symptoms Gazetteer

In contrast to the UMLS-based portion of the NLP pipeline, we explored a hybrid NLP annotation system that used gazetteer rule-based lookup with the lexicon of 164 derived COVID-19 symptoms in conjunction with the UMLS.

3.4.1 Hybrid COVID Symptoms Gazetteer

The gazetteer used for this study was the lexicon of 164 derived COVID-19 symptoms. The gazetteer lookup developed by Sahoo and Silverman (2020) uses spaCy's PhraseMatcher class to add the lexicon of 164 derived COVID-19 symptoms to a blank spaCy model (Gillißen, 2020). A manifest of ED admission notes was used to define a set of files (i.e., ED notes) that were read in by a spaCy matcher instance to return any mention of symptoms in the lexicon. In addition, the span of the text containing the symptom mention was also returned. Output was then lemmatized word by word to convert the text to its base form (e.g., the base form of "vomiting" is "vomit").

The scispaCy *en_core_sci_lg* model, described by Neumann et al. (2019), was used for phrase detection through use of a knowledge base. To improve the phrase detection performance of the gazetteer, the scispaCy EntityLinker was added to the pipeline. *EntityLinker* links to a UMLS knowledge base to provide NE disambiguation. EntityLinker performed a string overlap-based search on NEs, comparing them with concepts in the UMLS knowledge base using an approximate nearest neighbor search. Terms in the lexicon that did not map to the UMLS knowledge base were then added through a rule-based lookup using EntityRuler.[9]

A few examples of rule based lookup, including coverage in lexicographic variations are given here:

- pattern: [‘LEMMA’: ‘can’, ‘LEMMA’: ‘not’, ‘LEMMA’: ‘breathe’], variations covered: can’t breathe, cannot breathe
- pattern: [‘LEMMA’: ‘high’, ‘LEMMA’: ‘temperature’], variations covered: high temperature, higher temperature

Pizarro’s (2021) implementation of NegEx (negspaCy) for detection of negating terms was the final stage of the Hybrid Symptom Gazetteer’s pipeline. A set of features (hereby referred to as ”gazetteer-based features”) was then created using rules similar to those described above in Section 3.3.6. (For architecture, see ”NLP Features” in Figure 1 and ”Gazetteer-based Features” in Figure 4.)

3.5 Extraction Performance and Run-Time Analysis

A description of the creation of the expert-curated manually annotation reference set of notes used in this study, along with their overall characteristics is discussed Section 3.5.1. Performance metrics for comparison of acute CDC symptom extraction for all NLP annotation systems used in this study with the expert-curated manually annotated reference are described in Section 3.5.2. Details of a run-time analysis for the UIMA-based annotation systems and the Hybrid COVID Symptoms Gazetteer are also provided in Section 3.5.3.

3.5.1 Expert-curated Manually Annotated Reference

A small manually annotated reference consisting of 46 randomly-selected ED admission notes from March 3rd through December 18th was created to assess baseline performance and correctness of our classification to acute CDC symptom labels. Labels for acute CDC symptoms were assigned across all notes using the rules outlined in Section 3.3.6 for capturing the possibility that there both positive and negative mentions within an acute CDC symptom cluster. Manual review of the notes was conducted by a board certified surgical critical care trauma attending surgeon with informatics training (CT) at our institution. The annotator had experience treating well over 250 COVID-19 positive patients and was blinded to the results of each NLP annotation system used in this study.

3.5.2 Extraction Performance Comparison of NLP Annotation Systems

Annotations produced by each NLP annotation system used in this study were compared against the manually annotated reference. For each acute CDC symptom and its negation

[9] Source code is available here: `https://github.com/nlpie/covid_symptom_gazetteer/tree/hybrid`

status, the measure of accuracy (percentage of correctly labeled instances) was calculated. We also calculated the weighted micro-average for positive predictive value (precision), sensitivity (recall) and harmonic mean (f1-score) for each symptom and all symptoms together.

3.5.3 Run-Time Comparison of NLP Annotation Systems

Each UIMA-based annotation system and the Hybrid COVID Symptoms Gazetteer processed the same set of 3,000 ED notes, which were randomly selected from the pool of 19,924 total ED notes described in Section 3.1, to give an estimate of NLP annotation system run-time. Performance tests were serially executed in a Kubernetes workflow to ensure equal access to operating system resources. Each NLP annotation system ran as a single Docker container/Kubernetes pod on the same Azure VMWare with configurations listed in Appendix A.

3.6 Statistical Analysis

Associations of the UMLS-based features (described in Section 3.3.6) with clinical outcomes were assessed using logistic regression for binary classification and negative binomial regression for outcomes involving counts. The dependent variables of interest were in-hospital mortality, hospital admission, ventilation, respiratory complications, liver complications, development of venous thromboembolism (VTE), development of atrial fibrillation, infectious complications, and hospital readmission. Additionally, a binary composite outcome was developed and coded as positive if a patient had an all-cause in-hospital mortality, required ICU admission or mechanical ventilation, or required a hospital length of stay greater than 7 days.

All models were risk-adjusted to account for patient-level baseline demographics (age, sex at birth, race/ethnicity, English vs non-English speaking), the Elixhauser comorbidity index, as described by Moore et al. (2017) and implemented by Stagg (2015), and the most aberrant vital signs within the first 24 hours of hospital admission. Imputation was deemed unnecessary given the low rate of missing data ($< 4\%$) (see Table D.1 in Appendix D for prevalence of missing data).

Odds ratios were used to measure associations between the UMLS-based features and the dependent outcomes, as described above. Odds ratios (OR) greater than one were associated with an increased risk of developing the outcome, while OR less than one were associated with a decreased risk of developing the outcome. A significance value of $p < 0.05$ was used. To ascertain documentation practices in the ED, this same methodology was applied to assess differences in associations between English and non-English speakers for the outcomes of in-hospital mortality, hospital admission, and the composite outcome, as described above.

Lastly, model outcome differences was examined using multiple logistic regression with vital signs alone and the UMLS-based features alone for the outcomes of in-hospital mortality and hospital admission. The prognostic value of symptomatology alone compared to vital signs alone, as well as the marginal addition to risk prediction models utilizing demographics and comorbidity burden were tested, and the characteristics of various models that might represent different clinical practice environments are described. For a comparison of the effect of extraction methods on model performance these models were also run using the gazetteer-based features.

4. Results

Results for documentation practices in the ED for COVID-19 patients are presented in Section 4.1. In Section 4.2, results for the NLP annotation system extraction and run-time performance analyses, as outlined in Sections 3.5.2 and 3.5.3, are presented. Lastly, in Section 4.3 results of the analyses of associations described in Section 3.6 are presented.

4.1 Prevalence of Acute CDC Symptoms

COVID-19 positive patients who were seen in the ED within the window defined in Section 3.1 and those also admitted to hospital had presenting symptoms summarized in Tables 4 and 5, respectively.

Symptoms	Negative mention	No mention	Positive mention
Cough	63 (1.26%)	2,688 (53.70%)	2,255 (45.05%)
Fever	1,372 (27.41%)	897 (17.92%)	2,737 (54.67%)
Dyspnea	1,952 (38.99%)	521 (10.41%)	2,533 (50.60%)
Fatigue	174 (3.48%)	3,111 (62.15%)	1,721 (34.38%)
Aches	455 (9.09%)	3,377 (67.46%)	1,174 (23.45%)
Headaches	823 (16.44%)	2,947 (58.87%)	1,236 (24.69%)
Loss of taste or smell	192 (3.84%)	4,446 (88.81%)	368 (7.35%)
Sore throat	787 (15.72%)	3,500 (69.92%)	719 (14.36%)
Rhinitis congestion	828(16.54%)	3,428 (68.48%)	750 (14.98%)
Diarrhea	1,500 (29.96%)	2,551 (50.96%)	955 (19.08%)
Nausea vomiting	1,095 (21.87%)	2,511 (50.16%)	1,400 (27.97%)

Table 4: Summary of extracted acute CDC symptoms for 5,006 COVID-19 positive patients seen in the ED (using Boolean ensemble for extraction of symptoms described in Section 3.3.5).

4.2 Performance Evaluation

Section 4.2.1 presents results of the performance of the symptom extraction methods using metrics described in Section 3.5.2. Each clinical annotator system used in this study is compared to the manually annotate reference standard described in Section 3.5.1. This was done primarily to understand where future tuning of these systems was required. In Section 4.2.2 we present results of a run-time analysis of these systems as described in Section 3.5.3 to gauge potential for use in near real-time extraction of symptoms from clinical notes.

4.2.1 Acute CDC Symptom Extraction Performance

Performance of extraction for acute CDC symptoms along with their corresponding negation status was assessed against the manually annotated reference (described in Section 3.5.1) for the Boolean ensemble, individual UIMA-based annotation systems and the Hybrid

Symptoms	Negative mention	No mention	Positive mention
Cough	32 (1.49%)	1,258 (58.43%)	863 (40.08%)
Fever	569 (26.43%)	471 (21.88%)	1,113 (51.70%)
Dyspnea	689 (32.00%)	253 (11.75%)	1,211 (56.25%)
Fatigue	57 (2.65%)	1,175 (54.58%)	921 (42.78%)
Aches	188 (8.73%)	1,606 (74.59%)	359 (16.67%)
Headaches	391 (18.16%)	1,444 (67.07%)	318 (14.77%)
Loss of taste or smell	67 (3.11%)	2,011 (93.40%)	75 (3.48%)
Sore throat	293 (13.61%)	1,705 (79.19%)	155 (7.20%)
Rhinitis congestion	324 (15.05%)	1,620 (75.24%)	209 (9.71%)
Diarrhea	560 (26.01%)	1,153 (53.55%)	440 (20.44%)
Nausea vomiting	424 (19.69%)	1,125 (52.25%)	604 (28.05%)

Table 5: Summary of extracted acute CDC symptoms for 2,153 COVID-19 positive patients seen in the ED and admitted to hospital (using Boolean ensemble for extraction of symptoms described in Section 3.3.5).

COVID Symptoms Gazetteer. Accuracy and weighted micro-average scores of performance metrics (described in Section 3.5.2) were calculated over all acute CDC symptoms and are given in Table 6.

NLP Annotation System	Accuracy	Weighted Micro-Average		
		Precision	Recall	f1-score
BioMedICUS	0.77	0.80	0.77	0.78
CLAMP	0.86	0.86	0.86	0.86
cTAKES	0.82	0.85	0.82	0.83
MetaMap	0.85	0.86	0.85	0.85
Boolean Ensemble	0.87	0.87	0.87	0.87
Hybrid COVID Symptoms Gazetteer	0.85	0.88	0.85	0.85

Table 6: Extraction performance of Boolean ensemble, Hybrid COVID Symptoms Gazetteer and individual UIMA-based annotation systems on the manually annotated reference over all acute CDC symptoms.

A similar analysis was also done for the Boolean ensemble, individual UIMA-based annotation systems and the Hybrid COVID Symptoms Gazetteer, for each acute CDC symptom, as shown in Figure 5.

4.2.2 NLP Annotation System Run-Time

The following run-times were recorded for each individual NLP annotation system to process the same set of 3,000 ED notes (as described in Section 3.5.3):

- Hybrid COVID Symptoms Gazetteer: 1,913.80 seconds
- BioMedICUS: 2,445.33 seconds

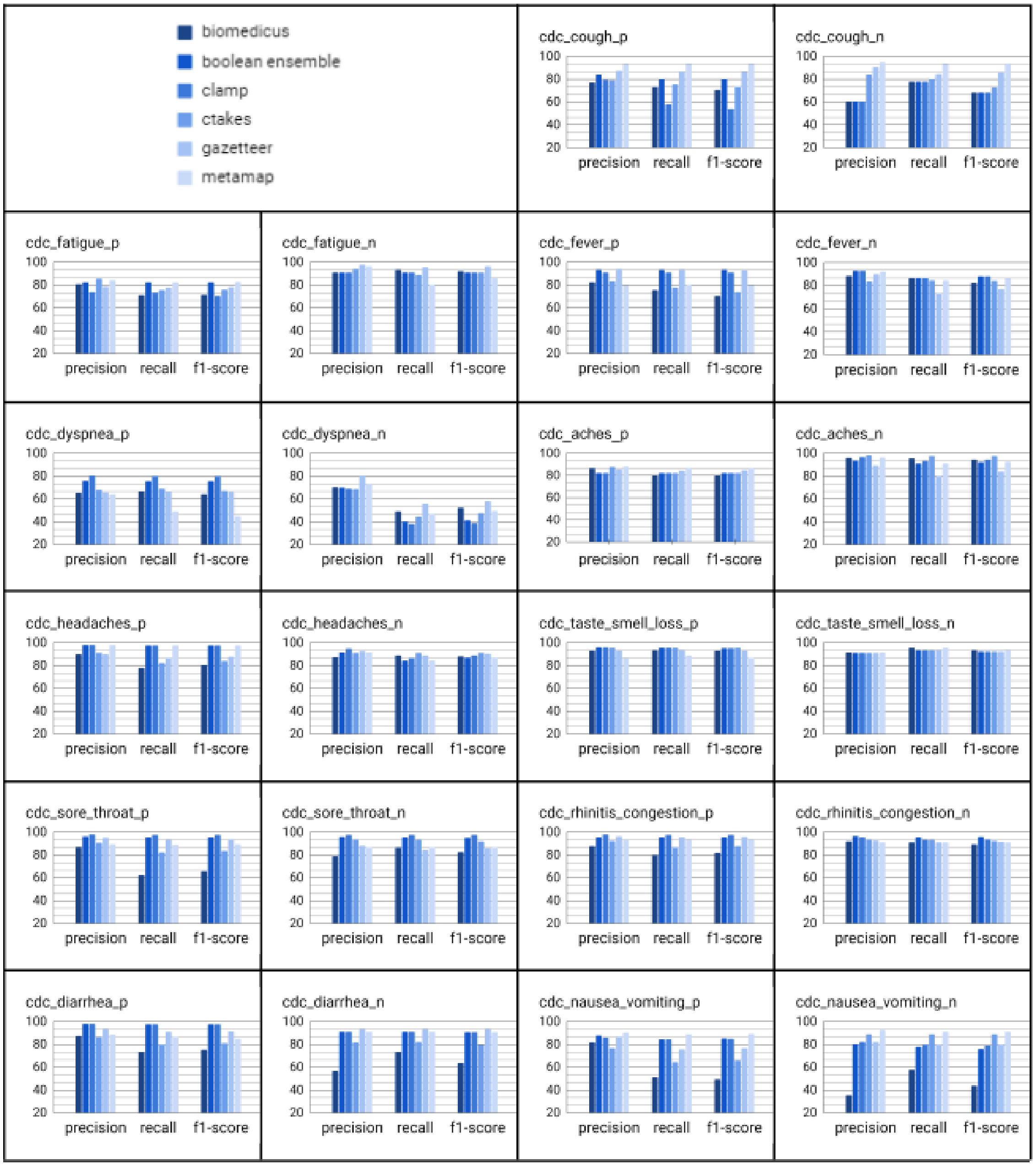


Figure 5: Acute CDC Symptom extraction performance evaluation of Boolean ensemble, Hybrid COVID Symptoms Gazetteer and individual UIMA-based annotation systems on the manually annotated reference over all symptoms. The x-axis represents the performance metric used and the y-axis represents the percentage value.

- cTAKES: 7,719.43 seconds
- CLAMP: 24,621.30 seconds
- MetaMap: 81,328.53 seconds

4.3 Statistical Analysis

Section 4.3.1 presents results of the analysis for all clinical outcomes described above in Section 3.6. Section 4.3.2 presents results of the analysis as described above with specific focus on English versus non-English speakers, as described above in 3.6. Lastly, Section 4.3.3 presents results of our analysis of model outcome differences with a comparison between symptom extraction methods, as described above in 3.6.

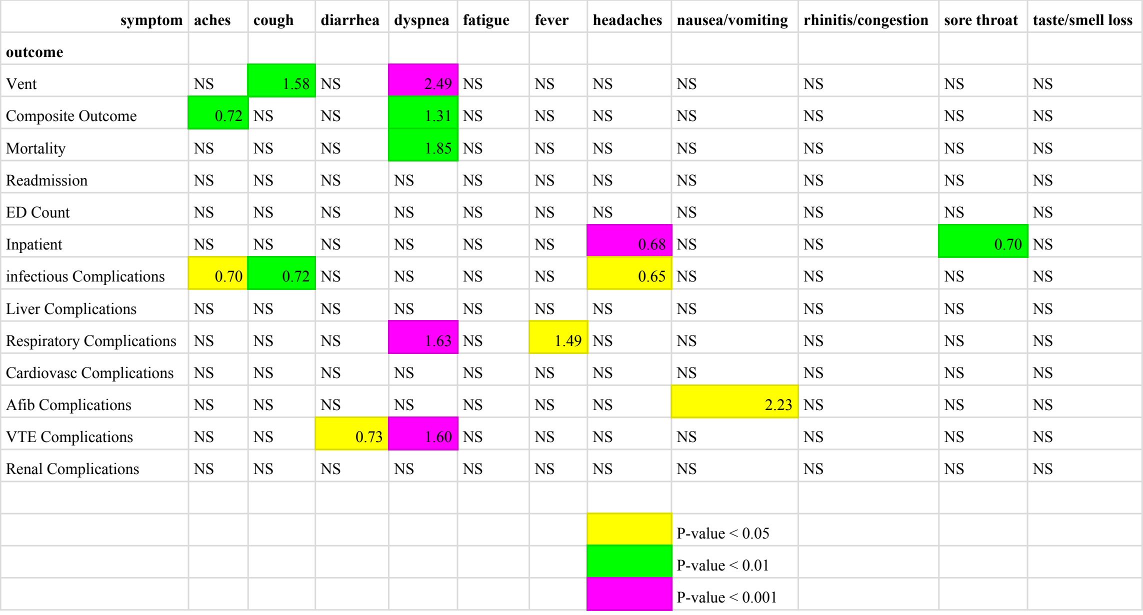
4.3.1 Associations of the Acute CDC Symptoms Across All Outcomes

An examination of the associations of the UMLS-based features across the outcomes of in-hospital mortality, hospital admission and readmission, ventilation, and the composite outcome, as described in Section 3.6, found the following acute CDC symptoms were of importance: aches was associated with reduced risk for the composite outcome (OR 0.72, p-value = 0.01); cough was associated with increased risk for ventilation (OR 1.58, p-value = 0.01); dyspnea was associated with increased risk for in-hospital mortality (OR 1.85, p-value < 0.001), ventilation (OR 2.50, p-value < 0.001), and the composite outcome (OR 1.31, p-value = 0.003); both headaches (OR 0.68, p-value < 0.001) and sore throat (OR 0.70, p-value = 0.01) were associated with reduced risk of hospital admission.

Next, associations of the UMLS-based features and various complications, including respiratory complications, liver complications, VTE, atrial fibrillation, infectious complications, and hospital readmission indicated the following acute CDC symptoms were of importance: The presence of aches (OR 0.70, p-value = 0.03), cough (OR 0.72, p-value = 0.01), or headaches (OR 0.65, p-value = 0.01) was associated with reduced risk of infectious complications; the presence of dyspnea (OR 1.63, p-value < 0.001) or fever (OR 1.49, p-value = 0.01) was associated with increased risk of respiratory complications, and nausea and vomiting was associated with increased risk of atrial fibrillation (OR 2.42 p-value = 0.04); diarrhea was associated with reduced risk of VTE (OR 0.73, p-value = 0.05) while dyspnea was associated with increased risk of VTE (OR 1.60, p-value < 0.001). Results are given in Table 7.

4.3.2 Associations For English Versus Non-English Speaking Between Acute CDC Symptoms and Outcomes

An examination of the associations of the UMLS-based features across the outcomes of in-hospital mortality and hospital admission, grouped by English and non-English speaking populations, revealed that the presence of aches was associated with a reduced risk of the composite outcome in both populations (non-English speaking [OR 0.63, p-value = 0.05]; English speaking [OR 0.76, p-value = 0.04]). Dyspnea was associated with increased risk for the composite outcome for both groups (non-English speaking [OR 1.58, p-value = 0.02]; English speaking [OR 1.24, p-value = 0.04]). Dyspnea was also associated with



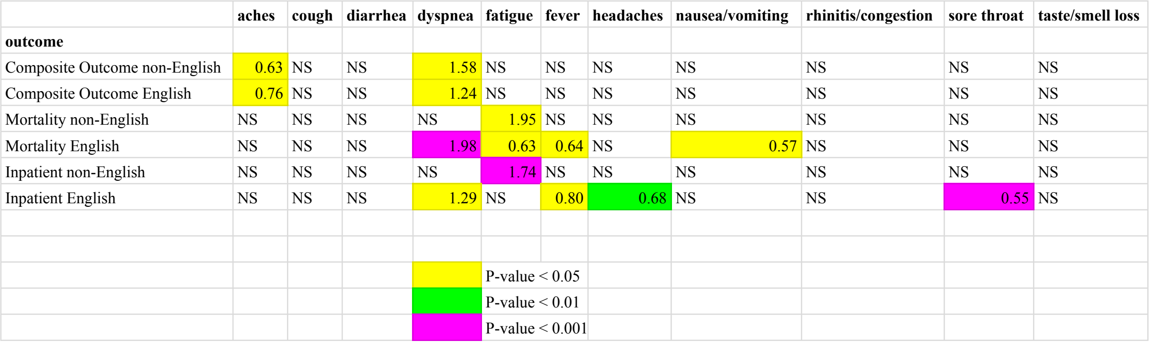
symptom	aches	cough	diarrhea	dyspnea	fatigue	fever	headaches	nausea/vomiting	rhinitis/congestion	sore throat	taste/smell loss
outcome											
Vent	NS	🟩 1.58	NS	🟪 2.49	NS	NS	NS	NS	NS	NS	NS
Composite Outcome	🟩 0.72	NS	NS	🟩 1.31	NS	NS	NS	NS	NS	NS	NS
Mortality	NS	NS	NS	🟩 1.85	NS	NS	NS	NS	NS	NS	NS
Readmission	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
ED Count	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Inpatient	NS	NS	NS	NS	NS	NS	🟪 0.68	NS	NS	🟩 0.70	NS
infectious Complications	🟨 0.70	🟩 0.72	NS	NS	NS	NS	🟨 0.65	NS	NS	NS	NS
Liver Complications	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Respiratory Complications	NS	NS	NS	🟪 1.63	NS	🟨 1.49	NS	NS	NS	NS	NS
Cardiovasc Complications	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Afib Complications	NS	NS	NS	NS	NS	NS	NS	🟨 2.23	NS	NS	NS
VTE Complications	NS	NS	🟨 0.73	🟪 1.60	NS	NS	NS	NS	NS	NS	NS
Renal Complications	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
							🟨	P-value < 0.05			
							🟩	P-value < 0.01			
							🟪	P-value < 0.001			

Table 7: Significant odds ratios for UMLS-based features versus outcomes for all patients. The composite outcome was defined as having any of the following outcomes occur: in-hospital death, admission requiring ICU or need for mechanical ventilation, or hospital length of stay $>$ 7 days.

increased risk of mortality (OR 1.98, p-value < 0.001) and hospital admission (OR 1.29, p-value $= 0.01$) in the English speaking population but was not significant for the non-English speaking population. Fatigue was associated with increased risk of both in-hospital mortality (OR 1.95, p-value $= 0.02$) and hospital admission (OR 1.74, p-value < 0.001) for non-English speakers. In contrast, for English speakers, fatigue was associated with reduced risk of in-hospital mortality (OR 0.63, p-value $= 0.01$). Fever was associated with reduced risk of in-hospital mortality (OR 0.64, p-value $= 0.03$) and hospital admission (OR 0.80, p-value $= 0.04$) in English speakers only. Similarly, headaches (OR 0.68, p-value $=$ 0.002) and sore throat (OR 0.55, p-value < 0.001) were found to be associated with reduced risk of hospital admission, while nausea and vomiting were associated with reduced risk of in-hospital mortality (OR 0.57, p-value $= 0.03$) for only the English speaking group. All other associations were not significant. Results are given in Table 8.

4.3.3 Effect of Acute CDC Symptoms

Multiple logistic regression models using only UMLS-based and gazetteer-based features derived from acute CDC symptoms (as discussed in Sections 3.3.6 and 3.4, respectively), were compared to a model using only vital signs to determine the effect on the outcome of hospital admission. There was no difference between the models using the symptoms; however, both models were significantly different when compared to the model using vital



	aches	cough	diarrhea	dyspnea	fatigue	fever	headaches	nausea/vomiting	rhinitis/congestion	sore throat	taste/smell loss
outcome											
Composite Outcome non-English	🟨 0.63	NS	NS	🟨 1.58	NS	NS	NS	NS	NS	NS	NS
Composite Outcome English	🟨 0.76	NS	NS	🟨 1.24	NS	NS	NS	NS	NS	NS	NS
Mortality non-English	NS	NS	NS	NS	🟨 1.95	NS	NS	NS	NS	NS	NS
Mortality English	NS	NS	NS	🟪 1.98	🟨 0.63	🟨 0.64	NS	🟨 0.57	NS	NS	NS
Inpatient non-English	NS	NS	NS	NS	🟪 1.74	NS	NS	NS	NS	NS	NS
Inpatient English	NS	NS	NS	🟨 1.29	NS	🟨 0.80	🟩 0.68	NS	NS	🟪 0.55	NS
				🟨	P-value < 0.05						
				🟩	P-value < 0.01						
				🟪	P-value < 0.001						

Table 8: Significant odds ratios for UMLS-based features versus outcomes for the population of English versus non-English speakers. Composite outcome was defined as having any of the following outcomes occur: in-hospital death, admission requiring ICU or need for mechanical ventilation, or hospital length of stay > 7 days.

signs. Symptoms alone did not result in any significant difference between all models on the outcome of in-hospital mortality. Results for all models are presented in Table 9. In all cases, age, sex at birth, and the Elixhauser comorbidity index have an outsized effect on outcomes and are thus used as base predictors in all models. For output from this analysis please see Online Appendix 3.

Model	Hospital Admission		In-Hospital Mortality	
	AUC ROC	95% CI	AUC ROC	95% CI
Without symptoms	0.89***	(0.88, 0.90)	0.90	(0.88, 0.91)
Without vitals (UMLS-based features)	0.84***	(0.83, 0.85)	0.87**	(0.85, 0.88)
Without vitals (Gazetteer-based features)	0.84***	(0.83, 0.85)	0.86**	(0.84, 0.88)

Table 9: AUC ROC with 95% confidence intervals. *** denotes statistically significant differences between models ”Without symptoms“ and “Without vitals;” (*NB*: For In-hospital mortality, there is no difference between the UMLS-based and Gazetteer-based models as denoted by **).

5. Discussion of Results

Key findings based on extraction of acute CDC symptoms across the population of patients defined in Section 3.1 are given below. Section 5.1 presents a qualitative assessment of symptoms extracted among those seen in the ED and those admitted to hospital, with focus on how these are documented by providers. Section 5.2 highlights the overall extraction performance of the Boolean ensemble, each individual UIMA-based annotation system and the Hybrid COVID Symptoms Gazetteer, with focus on gaps in the extraction task, and ends with a discussion of our run-time analysis. Lastly, Section 5.3 presents some surprising findings with respect to the analysis of associations described in Section 4.3.

5.1 Prevalence of Documented Acute CDC Symptoms

As seen in Tables 4 and 5 in Section 4.1, there is a large degree of variability in proportion between acute CDC symptom occurrence and symptom documentation rates. For the entire population examined in this study, the presence or absence of taste and smell loss was not documented approximately 88% of the time, while dyspnea was not documented approximately 10% of the time. While EHR use and documentation practices are known to differ (Lanham et al., 2013; Pakhomov et al., 2008), practices during the current pandemic differ significantly from everyday practice. The immediate pivot by the healthcare and scientific community was impressive and brought about new information that was disseminated very rapidly (Ingraham & Tignanelli, 2020). Despite these efforts and CDC recommendations that specific symptoms should be used to assess COVID-19, documentation of these symptoms was varied. While it is not possible to know if the documentation accurately reflects the actual clinical encounter, the use of documentation to convey information to other healthcare providers remains critical. Thus, improving documentation practices during a pandemic is an area for further investigation.

5.2 NLP Annotation Systems Extraction Performance Evaluation

Results in Table 6 in Section 4.2.1 indicate all NLP annotation systems used in this study have close values for precision and recall, with the exception of BioMedICUS (which had the lowest recall value). Table 6 shows the Hybrid COVID Symptoms Gazetteer had highest precision and thus minimized occurrence of false positives over all acute CDC symptoms and their negation status. On the other hand, the Boolean ensemble had the highest recall and was thus best at not missing ED notes classification to appropriate acute CDC symptom labels (viz., reducing false negatives). The ability to minimize false negatives is extremely important in clinical practice, since ideally an NLP annotation system should be able to detect the presence of all symptom mentions. This is the ideal for running diagnostic screening tests.

A comparison between Table 6 and Figure 5 was done to determine how each acute CDC symptom affected extraction performance. An analysis of the overall weighted micro-average f1-score for each NLP annotation system with respect to the system's f1-score for each symptom was done to give insight into how NLP annotation systems were affected by each acute CDC symptom in terms of overall weighted micro-average f1-score. A few top symptoms that reduced the overall f1-score for each given system are listed in Table 10.

Examination of Table 10 and Figure 5 indicates that cdc_dyspnea_n had a consistently low f1-score for all NLP annotation systems used in this study. A manual audit of a few notes with positive or negative mention of dyspnea in the manually annotated reference, but with no mention in the Boolean Ensemble (shown in Table 11), revealed limitations in the query expansion process with potential for loss of relevant concepts due to the query expansion rules presented in Figure 2. There was also potential for an increase in false negatives for the Hybrid COVID Symptoms Gazetteer, since the UMLS version implemented for

System	Top symptoms
BioMedICUS	cdc_cough_n, cdc_dyspnea_p, cdc_dyspnea_n, cdc_nausea_vomiting_p, cdc_nausea_vomiting_n
CLAMP	cdc_cough_p, cdc_cough_n, cdc_fatigue_p, cdc_dyspnea_n
cTAKES	cdc_dyspnea_p, cdc_dyspnea_n, cdc_nausea_vomiting_p
MetaMap	cdc_dyspnea_p, cdc_dyspnea_n
Boolean ensemble	cdc_cough_n, cdc_dyspnea_n
Hybrid COVID Symptoms Gazetteer	cdc_dyspnea_p, cdc_dyspnea_n

Table 10: List of acute CDC symptoms that reduced the overall weighted micro-average f1-score for all NLP annotation systems.

the knowledge base was not the fully licensed version (which includes several important lookup sources, including SNOMED CT).

Results presented in Figure 5 also indicate that all NLP annotation systems used in this study had f1-scores of approximately 80% or more for the following acute CDC symptoms: cdc_fatigue_n; cdc_fever_n; cdc_aches_p and cdc_aches_n; cdc_headaches_p and cdc_headaches_n; cdc_taste_smell_loss_p and cdc_taste_smell_loss_n; cdc_sore_throat_n; and cdc_rhinitis_congestion_p and cdc_rhinitis_congestion_n. Thus, these symptoms aligned well with the manually annotated reference.

It should be noted that the manually annotated reference used in this study may not be representative of the entire population of patients seen in the ED, and thus, results from this evaluation cannot be generalized. However, when used as a baseline for evaluating potential gaps in our lexicon, this allowed for rapid modification to the lexicon as needed for our production NLP pipeline. To truly gauge extraction performance of NLP annotation systems used in this study for the general task of symptom extraction, further investigation with a larger manually annotated reference is warranted.

It is noteworthy that the NLP pipeline used by Stephens et al. (2020), which utilized MetaMap Lite, successfully identified the presence of acute CDC symptoms at 0.89 precision against a set of 20 notes annotated for 200 symptoms. However, symptoms for 14 cases could not be determined. Also of note, the Boolean ensemble performed generally higher compared to all other NLP annotation systems for classifying the full set of positive and negative mentions for all acute CDC symptoms as reported in Table 6 and Figure 5. While it has yet to be determined if the differences between NLP annotation systems are significant, this is consistent with previous research which shows the benefit of Boolean ensemble combinations, especially when combined with query expansion of terms (Silverman et al., 2019; Finley et al., 2017; Kuo et al., 2016; Tignanelli et al., 2020).

Lastly, although the Hybrid COVID Symptoms Gazetteer was in experimental stages at the time of this study it had a comparable f1-score to the other NLP annotation systems and was much faster for extraction as seen in Section 4.2.2. The Hybrid COVID

Manually Annotated Reference	**Boolean Ensemble**	**COVID Symptoms Gazetteer**	**Phrases present in the notes**	Explanation
1	0	0	'Acute respiratory failure', 'Hypoxia', 'Acute respiratory failure with hypoxia', 'Pneumonia'	Phrases undetected by Boolean Ensemble and Hybrid COVID Symptoms Gazetteer
1	0	1	'Shortness of breath'	Phrases undetected by Boolean Ensemble but detected by Hybrid COVID Symptoms Gazetteer
1	1	0	'Difficulty breathing'	Phrases undetected by Hybrid COVID Symptoms Gazetteer but detected by Boolean Ensemble

Table 11: Examples of false negatives for positive or negative mentions of dyspnea as returned by Boolean Ensemble and Hybrid COVID Symptoms Gazetteer along with explanations.

Symptoms Gazetteer was 1.28 times faster than BioMedICUS; 4.03 times faster than cTAKES; 12.87 times faster than CLAMP; and 42.50 times faster than MetaMap.

5.3 Statistical Analysis

In Section 5.3.1 we discuss critical implications based on results of the analyses of associations against key clinical outcomes presented in Section 4.3.1. In Section 5.3.2 we discuss how results presented in Section 4.3.2 implicate that being a native English speaker may be associated with decreased risks from acute COVID-19 compared to non-native speakers. Lastly, in Section 5.3.3 we discuss how model differences presented in Section 5.3.3 can be exploited to provide streamlined point of care.

5.3.1 Associations of Acute CDC Symptoms Across All Outcomes

In this study, we were able to determine multiple associations between acute CDC symptoms and patient outcomes using the methods described in Section 3.6 with results presented in Section 4.3.1. As expected, cough and dyspnea were associated with a 58% and 150% increased risk of being placed on mechanical ventilation, respectively. These symptoms are likely indicative of the known pulmonary pathology and manifestations of COVID-19 (Ingraham et al., 2020a). On the other hand, aches were associated with 28% lower risk of the composite outcome; and headaches and sore throat were associated with lower risk of hospital admission. These less specific symptoms may represent a different disease phenotype versus a difference in the progression of the disease, depending on when the patient presented to the emergency room. Similar trends were seen in regard to developing complications. Dyspnea remained a prominent risk factor for respiratory complications, which is not surprising.

Less intuitive findings include how aches, cough, and headaches are associated with lower risk of developing infectious complications. Further study is need to clarify this. Lastly, there are other associations that are intuitive and may serve to denote high risk patients in certain settings. For example, nausea and vomiting were found to be associated with atrial fibrillation, likely due to their association with electrolyte abnormalities. Given the risk of atrial fibrillation in the ICU and its sequelae, knowing a patient is at high risk for developing atrial fibrillation from the ED can better prepare providers for possible complications during the patient's hospital admission (Bosch et al., 2018). It should be noted that contrary to results in this study, others, including Ramachandran et al. (2020), have reported there was no association between gastrointestinal symptoms and poor outcomes in COVID-19 patients. While these analyses were not meant to infer causality, they may still shed light on ways to further improve prognostic models and highlight areas that warrant further investigation. During a pandemic where any insight into a disease is invaluable, knowledge of which symptoms are associated with adverse outcomes could save lives in areas where resources are scarce and appropriate triage is critical.

5.3.2 Associations For English Versus Non-English Speaking Between Acute CDC Symptoms and Outcomes

Disparate outcomes among minority populations in COVID-19 are well documented, and have gained much attention throughout the COVID-19 pandemic (Eisner et al., 2011; Laster Pirtle, 2020; Harlem & Lynn, 2020; Mendy et al., 2020; Meneses-Navarro et al., 2020; Ransing et al., 2020; Turner-Musa et al., 2020; Wang & Tang, 2020). Newer data are finding that language barriers are likely contributing factors. Ingraham et al. (2021) found that non-English speaking was significantly associated with increased risk of severe disease and need for hospitalization, across patients with confirmed COVID-19 disease, despite controlling for race and neighborhood-level socioeconomic status. Since a goal of this study was to assess acute CDC symptoms and their association with patient outcomes, the effect of primary language on these associations was a logical extension of this analysis.

In this study, using the methods described in Section 3.6 with results presented in Section 4.3.2, we showed that symptoms with similar associations in both English and non-English speaking models, such as aches and dyspnea, were associated with reduced and increased risk of the composite outcome, respectively. In contrast, there were multiple significant associations found in the English speaking model that were not significant in the non-English speaking model.

Most concerning were the findings with opposing positive and negative risk for the same outcome. Fatigue was associated with a 37% lower risk of mortality in English speaking patients, while fatigue was associated with a 95% increased risk of mortality in those without English as their primary language. For non-English speakers, failure to collect symptoms is a major issue, given the dearth of significant results. These results support accounting for primary language, given the differences in models. Much of the signal may be from the lack of ability to collect symptoms and will contribute to the limitations of any model using symptoms.

Lastly, disparities in documented symptomatology between English and non-English speakers may be due to less accurate history collection with non-English speakers. This is notable especially during the COVID-19 pandemic, since interpreters are not physically present in the room: this can limit communication when compared to face-to-face encounters (Locatis et al., 2010). Leveraging a standardized and more sophisticated method of obtaining and documenting symptomatology may reduce this disparity and improve delivery.

5.3.3 Effect of Acute CDC Symptoms

Abdulaal et al. (2020) constructed a lexicon of symptoms and comorbidities for use in their predictive model for COVID-19. In contrast, the predictive models evaluated by Stephens et al. (2020) for influenza expanded on the EHR symptoms using the UMLS. Both studies illustrate the utility of symptoms for improving predictive modeling, whether through a carefully constructed lexicon or using the UMLS. Our study leveraged both lexicon-derived and UMLS-derived symptom data in the Hybrid COVID Symptoms Gazetteer to capture the strengths of both strategies.

Using the methods described in Section 3.6 with results presented in Section 4.3.3, we showed for the outcome of in-hospital mortality the symptoms-only model (both for the Boolean ensemble and the Hybrid COVID Symptoms Gazetteer) was not significantly different from the vitals-only model. The observation that a symptoms-only inventory could be used for risk prediction and prognostication strongly supports the use of telemedicine for COVID-19 triage, which could decrease the need for in-person evaluations and the resources and risks associated with these evaluations. This could provide clinicians with a powerful tool to make informed treatment decisions based on reported symptoms to help optimize both clinical resources and patient outcomes. Given that a clinical encounter is more than a symptom inventory, it would be premature to state that a patient symptom inventory would perform equivalently, though this remains an exciting area for future investigation.

It should be noted that predictions for being admitted to hospital were improved through the use of vital signs over both models using only symptoms. Further exploration across all outcomes is thus warranted.

5.4 Limitations

This study had multiple strengths, including a large number of patients, a well-characterized cohort and definitive outcomes. However, there were several limitations noted below.

5.4.1 Generalizability

The major limitation of this study is generalizability. Data collected for this study were from a single regional healthcare network and only included patients seen in the ED who tested positive for acute COVID-19. Also of note, the population of patients in this study were predominantly white, although the presented methods have potential to be used in health care networks with more diverse populations of patients.

Also, our study only focused on the symptomatology associated with the CDC's guidelines for acute COVID-19. As noted by Champika et al. (2020), there are notable symptoms not found in any of the major symptoms of acute COVID-19 catalogued by the United States/China CDC, WHO, etc. We are thus exploring ways to systematically expand the lexica for both acute COVID-19 and PASC. To help facilitate this, methods from this study are also being extended to OP clinical settings to conduct longitudinal monitoring of PASC.

Of note, due to the way the testing window was defined for inclusion into this study (as discussed in Section 3.1 and elaborated in Footnote 2), a large share of COVID-19 positive patients were missed in this study. However, since December 2020, with the introduction of rapid PCR testing, we have expanded this window to ± 14 days of COVID-19 diagnosis, thus expanding our cohort.

5.4.2 Manually Annotated Reference

Baseline performance analysis of the extraction methods used in the study was carried out on a small corpus of 46 randomly selected ED notes, as described in Section 3.5.1, meeting inclusion criteria outlined in Section 3.1. These notes were manually annotated by a single rater and served as a reference to compare against how the 4 UIMA-based annotation systems, their Boolean combination ensemble, and the Hybrid COVID Symptoms Gazetteer examined in this study, each labeled for the acute CDC symptoms in our lexicon and dictionary with negation status. While this approach would not be sufficient to generalize extraction performance over the entire corpus, it allowed for quick assessment of alignment of NLP-based methods with manual annotations, as well as rapid identification for the presence of possible disagreements and gaps in the different methods of symptom annotation and extraction. Use of a limited set of notes for baseline NLP annotation system performance evaluation has been implemented in other similar studies (Stephens et al., 2020; Tignanelli et al., 2020). To address issues of generalizability and validity we are in the process of creating a larger reference set of ED admission and general OP notes being annotated by multiple raters for acute COVID-19 and PASC symptoms.

5.4.3 NLP Methods

We did not explore the performance of other Boolean combinations for extraction of acute CDC symptoms derived from annotations generated by NLP-ADAPT-KUBE (of which there were 302 total possibilities[10]). While results for the Boolean combination used in this study were encouraging, and while this ensemble was the top performing Boolean combination in other experiments as outlined in Section 3.3.5, there is the possibility that other Boolean combinations could outperform this one. We plan to assess these ensembles against an expanded manually annotated reference to evaluate how the symptom extraction performance and model associations are affected. To this end, we are currently performing experiments using NLP-Ensemble-Explorer to examine the general behavior of Boolean ensemble combinations.

The rule-based lookup portion of the Hybrid COVID Symptoms Gazetteer also has potential to add a very large source of terms and phrases that could have been implemented into its lexicon. For example, terms that met the criteria for our query expansion, as defined

[10] Given four NLP annotation systems with two Boolean operations, this is computed by taking into account associativity, logical equivalence and the number of different ways n + 1 factors can be completely parenthesized as determined by the Catlan number. See *Enumerative Combinatorics*: Volume 2 by Richard P. Stanley.

in Section 3.2.2, but did not have a UMLS mapping are a potential source of more terms for the lexicon of COVID-19 symptoms.

While the Hybrid COVID Symptoms Gazetteer minimized the occurrence of false positives, there was a notable exception missed in our initial implementation. Since negations were determined by a preset span of text from the negating term, some negated terms were overlooked and thus mislabeled as a positive mention. As an example, the Hybrid COVID Symptoms Gazetteer detected a positive mention instead of a negative mention for ”wheezing” in the following sentence: ”Patient denies fever, myalgias, nausea, vomiting, abdominal pain, dysuria, hematuria, numbness and tingling, leg pain, difficulty walking, headache, visual disturbance, wheezing, and any other symptoms at this time.” This was because the span of the text containing ”wheezing” was not associated with the word ”denies,” which negates the mention for ”wheezing.”

We plan on addressing this issue by using sentence boundary detection. Furthermore, use of the UMLS as a knowledge base without any filtering by Semantic Types has the potential to mislabel text. To address this, we have designed a new version of the COVID Symptoms Gazetteer with more general rules for negation, and also without use of UMLS, as discussed in (Sahoo et al., 2021).

Lastly, regarding the lexicon used in this study, the *word2vec* model used for query expansion was trained on general patient notes and may not include syntactic nuances found in ED admission notes. To address this, we plan on fine tuning a transformer model like *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019)) using ED admission and OP notes. The fine-tuning process would include weak supervision of the BERT model for NER using few acute COVID-19 and PASC symptoms to extract additional symptoms.

5.4.4 Run-Time of NLP Annotation Systems

Use of the UIMA-based annotation systems in NLP-ADAPT-KUBE for large-volume real-time extraction of symptoms from notes is inefficient, both in terms of resource utilization and processing time. While Stephens et al. (2020) used MetaMap Lite for near real-time extraction of symptoms, it had several issues making it less than ideal candidate for real-time extraction of notes as discussed in Section 5.2.

MetaMap on its own performed comparably to the Boolean ensemble and the Hybrid COVID Symptoms Gazetteer for extraction of acute CDC symptom (as shown in Table 6 and Figure 5 under Section 4.2.1). However, based on the results of our run-time evaluation in Section 4.2.2, MetaMap took about 22 hours to process 3,000 notes. To consider using any of the default pipelines for any of the UIMA-based annotation systems for large-volume real-time extraction of symptoms they would need to be scaled across multiple compute nodes using a queuing system, like that used in the UIMA-AS implementation of NLP-ADAPT-KUBE (Finzel et al., 2020).

In comparison to the UIMA-based systems, the Hybrid COVID Symptoms Gazetteer is a good contender for large-volume real-time extraction of symptoms from notes, as shown in the NLP annotation system run-time results presented in Section 4.2.2. The Hybrid COVID Symptoms Gazetteer was 1.28 times faster than the next fastest UIMA-based annotation system (BioMedICUS) and 42.50 times faster than the slowest UIMA-based annotation system (MetaMap). Furthermore, scaling of the Hybrid COVID Symptoms Gazetteer across multiple nodes is much simpler to implement, since each instance can work off of its own manifest of files to process, compared to the need to implement a queuing system for the UIMA-based annotation systems (thereby increasing implementation complexity). The simple architecture of the Hybrid COVID Symptoms Gazetteer, and its extraction performance compared to UIMA-based systems, along with its speed make it a viable tool for sites needing high volume, near real-time extraction from clinical notes.

5.5 Potential Barriers

While our study indicates the use of symptomatology in CDS systems holds promise, potential barriers exist. An important aspect of this study was analysis of symptomatology-based models in the evaluation of non-English speaking patients. This observation is particularly important, since lack of English proficiency has been shown to increase the risk of mortality and developing severe COVID-19 (Ingraham et al., 2021). English proficiency likely plays a significant role in the clinician's understanding of the patient's presenting symptoms, thus influencing treatment decisions. Knowing whether an interpreter is available for an encounter could potentially lead to more conclusive results for our model. While it may be reasonable to assume most patients received some degree of interaction with a qualified interpreter, given the quick adaptation to incorporating video conferencing to most patient care rooms throughout the hospital system, the lack of data regarding the interaction remains a source of potential bias in our analyses. In tying ED admission notes to a well-defined lexicon or ontology (viz., the UMLS), the goal is to help clinicians break through some of these language barriers and make well-informed treatment decisions, likely through the assistance of interpreters. However, as noted in Section 5.3, restrictions during the COVID-19 pandemic limited use of interpreters during clinical encounters. In the future, mobile applications may be able to circumvent this barrier and assist clinicians by collecting data of various predictors, including symptoms, without being affected by language barriers, since these applications could be created in multiple languages (Stephens et al., 2020).

Finally, COVID-19 has exposed disparities in health literacy, particularly in terms of medical misinformation and politicization of response to the virus (Ingraham & Tignanelli, 2020; Ingraham et al., 2020b; Skilton, 2020). In this environment, the use of symptomatology could bypass health illiteracy by using the common language of symptoms to capture critical data for at-risk patients and simplify treatment decisions by leveraging symptoms for screening through an equitable platform across all languages.

6. Future Work

Methods presented in this study were developed quickly in response to the COVID-19 pandemic. Evaluation of the methods employed using these tools thus constituted a baseline approach to allow quick extraction of acute COVID-19 symptoms for use in this study. For the task of acute CDC symptom extraction there is currently no state-of-the-art. We describe the tools, methods and use cases on which we are currently focusing to fill this niche.

6.1 Gazetteer

Generation of specialized lexicons is a costly endeavor. For this study, development of the 164 derived COVID-19 symptoms described in Section 3.2.1 occurred over the course of four weeks and went through several iterations before full agreement about the completeness and correctness of terms in the lexicon was reached. To minimize valuable subject matter expert time involved in lexicon creation, and to increase yield of terms, we are currently exploring use of BERT models as noted in Section 5.4.3 for lexicon generation. Preliminary results from experiments using weak supervision and only 40 acute COVID symptoms to fine-tune BERT, for the purpose of NER, yielded approximately 320 unique COVID-19 symptoms from a set of 10,000 clinical notes for patients with COVID-19. While still in experimental stages, we have yet to validate the generated lexicon using BERT for extraction of COVID-19 symptoms on an expanded reference set of manually annotated notes.

To reduce the processing time even further, we have implemented a version of the COVID-19 Symptoms Gazetteer that utilizes all available cores without use of the UMLS (to reduce potential for false negatives as noted in Section 5.4.3). Experimental results for run-time, resource utilization and symptom extraction performance for acute COVID-19 are provided by Sahoo et al. (2021). Lastly, we are exploring use of a messaging queue to replace the need for a manifest list of files to process, which will allow for very high-throughput processing of documents at scale across multiple compute nodes. The potential of a custom lexical gazetteer both in terms of performance and processing time also motivates us to experiment on the benefits of using a gazetteer for diseases other than COVID-19. To this end, we are looking at ways to use domain-specific gazetteer-enhanced modules for improving NER models across disparate corpora as outlined by Liu et al. (2019).

Finally, we are examining the behaviour of combining the 4 UIMA-based annotation systems in NLP-ADAPT-KUBE and the Hybrid COVID Symptoms Gazetteer as a Neural Network (NN) ensemble. Preliminary results for the NN ensemble produced a weighted micro-average f1-score of 0.93 over all the acute CDC symptoms on the manually annotated reference, which was 0.06 higher than the f1-score for the Boolean ensemble used in this study.

6.2 Manually Annotated Reference

At the time of this study, the only reference of clinical notes manually annotated for acute CDC symptoms available to us was the set of notes described in Section 3.5.1. As noted above in Section 5.4.2, we are creating a larger and more comprehensive manually annotated reference. A major issue is that the creation of a manually annotated reference is a costly process in terms of time and resources. Use of weakly-supervised models trained using labeling functions created by subject matter experts have been shown to speed up this process to create annotated reference sets that are comparable to hand-labeled ones (Ratner et al., 2020). We are thus exploring the use of weak-supervision using Snorkel to assist in creating an expanded annotated reference set of annotated notes (Yang et al., 2019).

6.3 Symptomatology Use Cases

This study has shown the benefit of using extracted symptoms for acute COVID-19 from unstructured clinical encounter narratives for CDS. Several avenues of inquiry we are currently investigating are listed below.

Treatment of COVID-19 patients requires urgency to mitigate potential harm, including (a) initiation of anticoagulation to prevent thrombo-embolic disease, and (b) ensuring appropriate isolation procedures are established as soon as possible. Establishing a relationship between timing of care and symptomatology opens up several avenues of investigation, including care quality.

Symptomatology will substantially aid in identifying an accurate representation of an index encounter leading to a hospital stay. This includes collecting data that are poorly documented, such as confusion or altered mentation, along with other important diagnostic information at the point of care. Comparing a clinician's decision at the index encounter with the final diagnosis has the potential to open up multiple avenues of investigation into diagnostic decision making, atypical presentations, and error.

Finally, diagnostic uncertainty is a highly understudied area in clinical care delivery. For a subset of the COVID-19 population, use of symptomatology for clinical prediction will help identify patients with an uncertain diagnosis or prognosis. Comparing quality of care delivery and healthcare utilization between patients where there is demonstrable uncertainty and where there is not may help illustrate an important component of healthcare costs and patient safety. Flagging patients with an uncertain diagnosis may reduce potential for errors in diagnostic reasoning and trigger additional data collection to improve prognostication in real-time.

Furthermore, the investigation and understanding of the PASC remains in its infancy (Al-Aly et al., 2021). Given PASC is largely a symptoms-based diagnosis, it serves as an ideal platform to further leverage the methods presented in this study.

7. Conclusion

This study demonstrates that NLP methods for extraction of acute COVID-19 symptoms using the UMLS and a rule-based gazetteer offer potential to enhance clinical prognosis for various outcomes related to COVID-19. It also shows that clusters of symptoms for COVID-19 carry significant predictive power. Our results indicate that symptomatology-based prognostic models offer clinicians a potentially powerful tool that could facilitate telemedicine encounters, replacing the need for other in-person options that might jeopardize the well-being of health care workers and patients. The methods presented in this study have potential for diseases beyond acute COVID-19, including PASC, and other acute and chronic conditions. While this study has identified potential barriers (i.e., primary language) for how clinicians currently collect symptoms, it also offers solutions to circumvent these barriers. Most importantly, we have shown that evidence-based management of symptoms has potential to improve outcomes.

8. Acknowledgements

Serguei Pakhomov and Christopher Tiganelli both contributed equally as senior authors for this study. Also, we would like to extend our gratitude to Jenzi Silverman, Pa Xiong Vang, Tanya Melnick, Matthew Larson, Reed McEwan and Elizabeth Lindemann for their help in proofreading this manuscript.

9. Conflict of Interests and Funding Sources

This research was supported by Fairview Health Services, the National Institutes of Health's National Center for Advancing Translational Sciences grant U01TR002062, the National Institutes of Health's National Heart, Lung, and Blood Institute's grant T32HL07741 (NEI), the Agency for Healthcare Research and Quality (AHRQ) R01HS026743 (MGU) and Patient-Centered Outcomes Research Institute (PCORI), grant K12HS026379 (SS, CJT). Additional support for MN-LHS scholars is offered by the University of Minnesota Office of Academic Clinical Affairs and the Division of Health Policy and Management, University of Minnesota School of Public Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of Fairview Health Services, the National Institutes of Health, AHRQ, PCORI, or Minnesota Learning Health System Mentored Career Development Program (MN-LHS). This grant was also supported by the University of Minnesota CTSA grant UL1TR000114.

Appendix A. Configuration for NLP Annotation System Run-Time Analysis:

- Architecture: x86 64
- CPUs: 8 cores
- Thread(s) per core: 1
- Core(s) per socket: 1
- Socket(s): 8
- Model name: Intel(R) Xeon(R) Gold 6152 CPU @ 2.10GHz
- Operating System (OS): Ubuntu 18.04.4 LTS (Bionic Beaver)
- RAM: 64 GB
- Platform: Azure VMWare

Appendix B. Example Query Expansion

The following example illustrates the steps used in the query expansion:

1. Start with the term "cough" from the list of 164 derived COVID-19 symptoms.

2. Mapping “cough” to the UMLS produces the CUI: C1961131 and the preferred term: ”cough.”

3. Using the preferred term “cough” generate the top 100 semantically similar terms (in Table B.1 below we display only the top 11)

4. Mapping the semantically similar terms from step 3 to the UMLS we get the results below in Table B.2.

Semantically similar terms	**Cosine distance**
dry cough	0.9200
congestion	0.8729
coughing	0.8582
nonproductive	0.8484
nonproductive cough	0.8398
uri symptoms	0.8363
productive cough	0.8327
wheezing	0.8317
sore throat	0.8244
non productive	0.8211
non productive cough	0.8156

Table B.1: Top 11 semantically similar terms to the parent term “cough”.

Mapped Cui	**Mapped preferred term**
C0850149	dry cough
C0700148	congestion
C0010200	coughing
No UMLS mapping	nonproductive
C0850149	dry cough
No UMLS mapping	uri symptoms
C0239134	productive cough
C0043144	wheezing
C0031350	pharyngitis
No UMLS mapping	non productive

Table B.2: Top 11 semantically similar terms mapped to UMLS.

Appendix C. Example Binary Tree Expansion of Boolean Combination

As an example of how Boolean ensembles merge annotation output, the tree representation of arbitrary sets A, B, and C with the merge operation ((A ∩ B) ∪ C) is given in Figure 3. Starting from the bottom, we traverse the leaves of the tree. If the leaf is a set (in the case of A or B) we look to the leaf’s parent for an operation. A’s parent is an intersection operation node, which consumes both a left and a right leaf (A and B) as operands and combines all common elements between both sets A and B as an intersection. Included in Figure C.1 is the sub-expression (A ∩ B), which is the result of the intersection operation and which will be a leaf for future operations. In this simple case, there is only one parent node remaining, a union operation node. This union consumes two leaves — (A ∩ B) and the remaining leaf (C), to produce the final expression ((A ∩ B) ∪ C), which combines all elements between both sets (A ∩ B) and C as a union. The merged set for the given Boolean expression is thus complete.

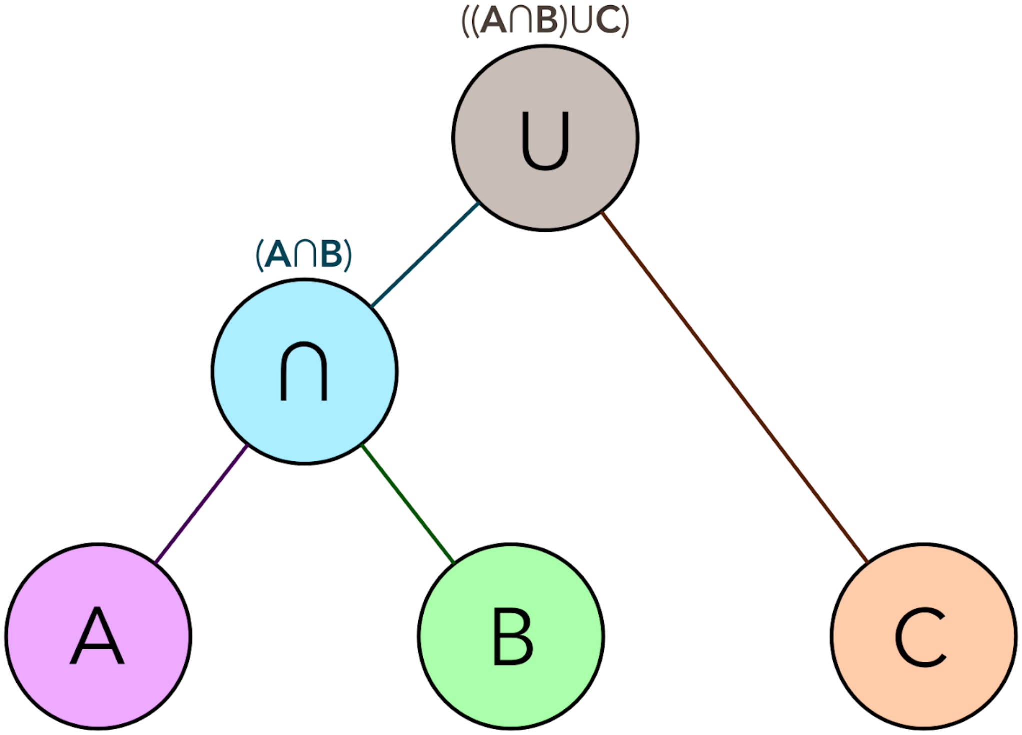


Figure C.1: Tree representation of Boolean expression.

Appendix D. Prevalence of Data used in Associations Analysis

Variable	Missing	Total	Percent Missing
Inpatient	0	5,006	0.0000
Readmission	0	5,006	0.0000
age_raw	0	5,006	0.0000
elixsum	0	5,006	0.0000
race	0	5,006	0.0000
Non_Englis~g	0	5,006	0.0000
male	0	5,006	0.0000
SBP_min24h	130	5,006	2.6000
Temp_max24h	145	5,006	2.9000
RR_max24h	192	5,006	3.8400
HR_max24h	121	5,006	2.4200
cdc_aches3	0	5,006	0.0000
cdc_fatigue3	0	5,006	0.0000
cdc_fever3	0	5,006	0.0000
cdc_cough3	0	5,006	0.0000
cdc_dyspnea3	0	5,006	0.0000
cdc_headac~3	0	5,006	0.0000
cdc_taste_~3	0	5,006	0.0000
cdc_sore_t~3	0	5,006	0.0000
cdc_rhinit~3	0	5,006	0.0000
cdc_nausea~3	0	5,006	0.0000
cdc_diarrh~3	0	5,006	0.0000

Table D.1: Data prevalence.[11]

[11] Variable definitions:

- Inpatient: Binary variable determining if admitted to hospital
- Readmission: Binary variable determining if readmitted to hospital
- age_raw: Age as float at time of encounter
- elixsum: Elixhauser comorbidity index (described in Section 3.6)
- race: Coded race/ethnicity
- Non-English: Binary variable determining if Non-English speaker
- SBP_min24h: Systolic Blood Pressure 24h Min (mmHg)
- Temp_max24h: Temperature 24h Max (F°)
- RR_max24h: Respiratory Rate 24h Max (breaths/min)
- HR_max24h: Heart Rate 24h Max (beats/min)
- cdc_*: Acute COVID-19 symptoms

References

Abdulaal, A., Patel, A., Charani, E., Denny, S., Mughal, N., & Moore, L. (2020). Prognostic Modeling of COVID-19 Using Artificial Intelligence in the United Kingdom: Model Development and Validation. *Journal of Medical Internet Research*, *22*(8), e20259. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Aklouche, B., Bounhas, I., & Slimani, Y. (2018). Query Expansion Based on NLP and Word Embeddings. In *Proceedings of the Twenty-Seventh Text Retrieval Conference (TREC 2018)*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST). Backup Publisher: National Institute of Standards and Technology (NIST).

Al-Aly, Z., Xie, Y., & Bowe, B. (2021). High-dimensional characterization of post-acute sequalae of COVID-19. *Nature*.

Albright, D., Lanfranchi, A., Fredriksen, A., Styler, William F, I., Warner, C., Hwang, J. D., Choi, J. D., Dligach, D., Nielsen, R. D., Martin, J., Ward, W., Palmer, M., & Savova, G. K. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, *20*(5), 922–930.

Andrews, E., Berghofer, K., Long, J., Prescott, A., & Caboral-Stevens, M. (2020). Satisfaction with the use of telehealth during COVID-19: An integrative review. *International Journal of Nursing Studies Advances*, *2*, 100008.

Aronson, A. (2001). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, *2001*, 17–21.

Aronson, A., & Lang, F.-M. (2010). An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association : JAMIA*, *17*, 229–36.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proc. AMIA Symp*, pp. 17–21.

Bodenreider, O. (2020). The UMLS Semantic Network. `https://semanticnetwork.nlm.nih.gov`.

Boettiger, C. (2015). An introduction to Docker for reproducible research. `https://doi.org/10.1145/2723872.2723882`.

Bompelli, A., Silverman, G. M., Finzel, R. L., Vasilakes, J., Knoll, B., Pakhomov, S., & Zhang, R. (2020). Comparing NLP Systems to Extract Entities of Eligibility Criteria in Dietary Supplements Clinical Trials Using NLP-ADAPT. In Michalowski, M., & Moskovitch, R. (Eds.), *Artificial Intelligence in Medicine - 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25-28, 2020, Proceedings*, Vol. 12299 of *Lecture Notes in Computer Science*, pp. 67–77. Springer.

Bosch, N. A., Cimini, J., & Walkey, A. J. (2018). Atrial Fibrillation in the ICU. *Chest*, *154*(6), 1424–1434.

Bursi, F., Weston, S. A., Redfield, M. M., Jacobsen, S. J., Pakhomov, S., Nkomo, V. T., Meverden, R. A., & Roger, V. L. (2006). Systolic and diastolic heart failure in the community.. *JAMA*, *296*(18), 2209–2216. Place: United States.

CDC (2020). Symptoms of Coronavirus. `https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html`.

Champika, S. K. G., Dineshani, H., Dileepa, E., & Saroj, J. (2020). Symptomatology of Coronavirus Disease 2019 (COVID-19) - Lessons from a meta-analysis across 13 countries [Preprint]. `https://doi.org/10.21203/rs.3.rs-39412/v2`.

Couto, F., Campos, L., & Lamurias, A. (2017). MER: a Minimal Named-Entity Recognition Tagger and Annotation Server..

Croft, P., Altman, D. G., Deeks, J. J., Dunn, K. M., Hay, A. D., Hemingway, H., LeResche, L., Peat, G., Perel, P., Petersen, S. E., Riley, R. D., Roberts, I., Sharpe, M., Stevens, R. J., Van Der Windt, D. A., Von Korff, M., & Timmis, A. (2015). The science of clinical practice: disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice.. *BMC medicine*, *13*, 20.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Dikaleh, S., Sheikh, O., & Felix, C. (2017). Introduction to kubernetes. In *Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering*, CASCON '17, p. 310, USA. IBM Corp.

Eisner, M. D., Blanc, P. D., Omachi, T. A., Yelin, E. H., Sidney, S., Katz, P. P., Ackerson, L. M., Sanchez, G., Tolstykh, I., & Iribarren, C. (2011). Socioeconomic status, race and COPD health outcomes. *Journal of epidemiology and community health*, *65*(1), 26–34. Edition: 2009/10/23.

Elkin, P. L., Froehling, D., Wahner-Roedler, D., Trusko, B., Welsh, G., Ma, H., Asatryan, A. X., Tokars, J. I., Rosenbloom, S. T., & Brown, S. H. (2008). NLP-based identification of pneumonia cases from free-text radiological reports.. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2008*, 172–176.

Fan, Y., Pakhomov, S., McEwan, R., Zhao, W., Lindemann, E., & Zhang, R. (2019). Using word embeddings to expand terminology of dietary supplements on clinical notes.. *JAMIA open*, *2*(2), 246–253.

Ferrucci, D., & Lally, A. (2004). Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*, 327 – 348.

Finley, G. P., Knoll, B. C., Liu, H., Xu, H., Melton, G. B., & Pakhomov, S. V. S. (2017). Using ensembles of NLP engines without a common type system to improve abbreviation disambiguation. *AMIA Joint Summits Proceedings*, *2017*, 2.

Finzel, R., & Silverman, G. (2019). nlp-adapt-kube. `https://github.com/nlpie/nlp-adapt-kube`.

Finzel, R. L., Silverman, G. M., Datar, S., & Liu, S. (2020). AIME 2020 Tutorial 2/AMIA 2019 W22: Large Scale Ensembled NLP Systems with Docker and Kubernetes - nlpie/Workshop_large_scale_nlp..

Gillißen, P. (2020). phrase_matcher.py. `https://github.com/explosion/spaCy/blob/master/examples/information_extraction/phrase_matcher.py`.

Harlem, G., & Lynn, M. (2020). Descriptive analysis of social determinant factors in urban communities affected by COVID-19.. *Journal of public health (Oxford, England)*, *42*(3), 466–469.

He, Z., Perl, Y., Elhanan, G., Chen, Y., Geller, J., & Bian, J. (2017). Auditing the assignments of top-level semantic types in the UMLS semantic network to UMLS concepts. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1262–1269.

Hincapié, M. A., Gallego, J. C., Gempeler, A., Piñeros, J. A., Nasner, D., & Escobar, M. F. (2020). Implementation and Usefulness of Telemedicine During the COVID-19 Pandemic: A Scoping Review. *Journal of Primary Care & Community Health*, *11*, 2150132720980612. Publisher: SAGE Publications Inc.

Hoogendoorn, M., Szolovits, P., Moons, L., & Numans, M. (2016). Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. *Artificial intelligence in medicine*, *69*, 53–61.

Ingraham, N. E., Barakat, A. G., Reilkoff, R., Bezdicek, T., Schacker, T., Chipman, J. G., Tignanelli, C. J., & Puskarich, M. A. (2020a). Understanding the renin-angiotensin-aldosterone-SARS-CoV axis: a comprehensive review.. *The European respiratory journal*, *56*(1).

Ingraham, N. E., Boulware, D., Sparks, M. A., Schacker, T., Benson, B., Sparks, J. A., Murray, T., Connett, J., Chipman, J. G., Charles, A., & Tignanelli, C. J. (2020b). Shining a light on the evidence for hydroxychloroquine in SARS-CoV-2.. *Critical care (London, England)*, *24*(1), 182.

Ingraham, N. E., Purcell, L. N., Karam, B. S., Dudley, R. A., Usher, M. G., Warlick, C. A., Allen, M. L., Melton, G. B., Charles, A., & Tignanelli, C. J. (2021). Racial and Ethnic Disparities in Hospital Admissions from COVID-19: Determining the Impact of Neighborhood Deprivation and Primary Language. *Journal of General Internal Medicine*.

Ingraham, N. E., & Tignanelli, C. J. (2020). Fact Versus Science Fiction: Fighting Coronavirus Disease 2019 Requires the Wisdom to Know the Difference.. *Critical care explorations*, *2*(4), e0108.

Klie, J.-C., & Castilho, R. E. d. (2020). DKPro Cassis - Reading and Writing UIMA CAS Files in Python. `https://doi.org/10.5281/zenodo.3994108`.

Knoll, B. (2019). The biomedical information collection and understanding system (biomedicus). `https://github.com/nlpie/biomedicus`.

Kuo, T.-T., Rao, P., Maehara, C., Doan, S., Chaparro, J. D., Day, M. E., Farcas, C., Ohno-Machado, L., & Hsu, C.-N. (2016). Ensembles of NLP Tools for Data Element Extraction from Clinical Notes. *AMIA Annual Symposium Proceedings*, *2016*, 10.

Lang, F.-M. (2016). MetaMap2016 Usage Notes. `https://metamap.nlm.nih.gov/Docs/MM_2016_Usage.pdf`.

Lanham, H. J., Sittig, D. F., Leykum, L. K., Parchman, M. L., Pugh, J. A., & McDaniel, R. R. (2013). Understanding differences in electronic health record (EHR) use: linking individual physicians' perceptions of uncertainty and EHR use patterns in ambulatory care. *Journal of the American Medical Informatics Association*, *21*(1), 73–81. _eprint: https://academic.oup.com/jamia/article-pdf/21/1/73/17375951/21-1-73.pdf.

Laster Pirtle, W. N. (2020). Racial Capitalism: A Fundamental Cause of Novel Coronavirus (COVID-19) Pandemic Inequities in the United States.. *Health education & behavior : the official publication of the Society for Public Health Education*, *47*(4), 504–508.

Li, J., Chen, Z., Nie, Y., Ma, Y., Guo, Q., & Dai, X. (2020). Identification of Symptoms Prognostic of COVID-19 Severity: Multivariate Data Analysis of a Case Series in Henan Province. *Journal of Medical Internet Research*, *22*(6), e19636. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Liang, W., Yao, J., Chen, A., Lv, Q., Zanin, M., Liu, J., Wong, S., Li, Y., Lu, J., Liang, H., Chen, G., Guo, H., Guo, J., Zhou, R., Ou, L., Zhou, N., Chen, H., Yang, F., Han, X., Huan, W., Tang, W., Guan, W., Chen, Z., Zhao, Y., Sang, L., Xu, Y., Wang, W., Li, S., Lu, L., Zhang, N., Zhong, N., Huang, J., & He, J. (2020). Early triage of critically ill COVID-19 patients using deep learning. *Nature Communications*, *11*.

Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S. R., Ravikumar, K. E., Wu, S. T., Kullo, I. J., & Chute, C. G. (2013). An information extraction framework for cohort identification using electronic health records.. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, *2013*, 149–153.

Liu, T., Yao, J.-G., & Lin, C.-Y. (2019). Towards Improving Neural Named Entity Recognition with Gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5301–5307, Florence, Italy. Association for Computational Linguistics.

Ljubesic, N., Boras, D., Bakaric, N., & Njavro, J. (2008). Comparing Measures of Semantic Similarity. In *ITI 2008 - 30th International Conference on Information Technology Interfaces*, pp. 675–682.

Locatis, C., Williamson, D., Gould-Kabler, C., Zone-Smith, L., Detzler, I., Roberson, J., Maisiak, R., & Ackerman, M. (2010). Comparing in-person, video, and telephonic medical interpretation.. *Journal of general internal medicine*, *25*(4), 345–350.

Mendy, A., Apewokin, S., Wells, A. A., & Morrow, A. L. (2020). Factors Associated with Hospitalization and Disease Severity in a Racially and Ethnically Diverse Population of COVID-19 Patients. *medRxiv*, 2020.06.25.20137323.

Meneses-Navarro, S., Freyermuth-Enciso, M. G., Pelcastre-Villafuerte, B. E., Campos-Navarro, R., Meléndez-Navarro, D. M., & Gómez-Flores-Ramos, L. (2020). The challenges facing indigenous communities in Latin America as they confront the COVID-19 pandemic. *International Journal for Equity in Health*, *19*(1), 63.

Meystre, S., & Haug, P. J. (2005). Automation of a problem list using natural language processing.. *BMC medical informatics and decision making*, *5*, 30.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *CoRR*, *abs/1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., & Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 26, pp. 3111–3119. Curran Associates, Inc.

Miller, B. N., & Ranum, D. L. (2013). Parse Tree. In *Problem Solving with Algorithms and Data Structures using Python*, p. Section 7.6.

Moore, B. J., White, S., Washington, R., Coenen, N., & Elixhauser, A. (2017). Identifying Increased Risk of Readmission and In-hospital Mortality Using Hospital Administrative Data: The AHRQ Elixhauser Comorbidity Index. *Medical care*, *55*(7), 698–705.

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*, 319–327.

Nguyen, D. H. M., & Patrick, J. D. (2014). Supervised machine learning and active learning in classification of radiology reports. *Journal of the American Medical Informatics Association*, *21*(5), 893–901. _eprint: https://academic.oup.com/jamia/article-pdf/21/5/893/5970216/21-5-893.pdf.

NLM (2009). UMLS® Reference Manual [Internet]. `https://www.ncbi.nlm.nih.gov/books/NBK9676/`.

NLM (2018). Semantic Types and Groups. `https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml`.

NLM (2020). Word-Sense Disambiguation. `https://wsd.nlm.nih.gov/`.

Pakhomov, S., Finley, G. P., McEwan, R., Wang, Y., & Melton, G. (2016). Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, *32 23*, 3635–3644.

Pakhomov, S. V., Jacobsen, S. J., Chute, C. G., & Roger, V. L. (2008). Agreement between patient-reported symptoms and their documentation in the medical record.. *The American journal of managed care*, *14*(8), 530–539.

Pizarro, J. (2021). Jenojp/negspacy. `https://github.com/jenojp/negspacy`.

Plisko, V. (2014a). Conjunction. `https://www.encyclopediaofmath.org/index.php/Conjunction`.

Plisko, V. (2014b). Disjunction. `https://www.encyclopediaofmath.org/index.php/Disjunction`.

Ramachandran, P., Onukogu, I., Ghanta, S., Gajendran, M., Perisetti, A., Goyal, H., & Aggarwal, A. (2020). Gastrointestinal Symptoms and Outcomes in Hospitalized Coronavirus Disease 2019 Patients. *Digestive Diseases*, *38*(5), 373–379.

Ransing, R., Ramalho, R., Filippis, R. d., Ojeahere, M. I., Karaliuniene, R., Orsolini, L., Costa, M. P. d., Ullah, I., Grandinetti, P., Bytyçi, D. G., Grigo, O., Mhamunkar, A., Hayek, S. E., Essam, L., Larnaout, A., Shalbafan, M., Nofal, M., Soler-Vidal, J., Pereira-Sanchez, V., & Adiukwu, F. (2020). Infectious disease outbreak related stigma and discrimination during the COVID-19 pandemic: Drivers, facilitators, manifestations, and outcomes across the world. *Brain, Behavior, and Immunity*, *89*, 555–558.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, *29*(2), 709–730.

Reátegui, R., & Ratté, S. (2018). Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making volume*, *18*(18), 74.

Rosendal, M., Carlsen, A. H., Rask, M. T., & Moth, G. (2015). Symptoms as the main problem in primary care: A cross-sectional study of frequency and characteristics. *Scandinavian journal of primary health care*, *33*(2), 91–99. Edition: 2015/05/11 Publisher: Taylor & Francis.

Sahoo, H., & Silverman, G. (2020). COVID Symptoms Gazetteer. `https://github.com/nlpie/covid_symptom_gazetteer/tree/hybrid`.

Sahoo, H., Silverman, G. M., Ingraham, N. E., Lupei, M., Puskarich, M. A., Finzel, R. L., Sartori, J., Zhang, R., Knoll, B. C., Liu, S., Liu, H., Melton, G. B., Tignanelli, C. J., & Pakhomov, S. V. S. (2021). A rule-based system for COVID-19 symptom identification and classification. *JAMIA Open*, *in-press*(in-press).

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Schuler, K. K., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, *17 5*, 507–13.

Silverman, G. (2020a). nlp-ensemble-explorer/extract_ensemble.py. `https://github.com/nlpie/nlp-ensemble-explorer/blob/polarity/ensemble_explorer/extract_ensemble.py`.

Silverman, G. (2020b). parse_uima.py. `https://github.com/nlpie/nlp-ensemble-explorer/blob/polarity/ensemble_explorer/parse_uima.py`.

Silverman, G. M., Lindemann, E. A., Rajamani, G., Finzel, R. L., McEwan, R., Knoll, B. C., Pakhomov, S. V. S., Melton, G. B., & Tignanelli, C. J. (2019). Named Entity Recognition in Prehospital Trauma Care. *Studies in Health Technology and Informatics*, *264*, 1586–7.

Silverman, G. M., Solinksy, J. C., Finzel, R. L., Usher, M. G., Lupei, M., Ingraham, N. E., Lusczek, E., Knoll, B. C., McEwan, R., Xu, H., Jiang, X., Melton, G. B., Pakhomov, S. V. S., & Tignanelli, C. J. (2020). Using Ensembles of NLP Systems for Extraction of UMLS Concepts from Unstructured Data for use in Covid-19 Clinical Data Analytics. `https://canvas.umn.edu/courses/180230/discussion_topics/808177`.

Skilton, N. (2020). Health literacy and disparities in COVID-19–related knowledge, attitudes, beliefs and behaviours in Australia..

Skube, S. J., Hu, Z., Simon, G. J., Wick, E. C., Arsoniadis, E. G., Ko, C. Y., & Melton, G. B. (2020). Accelerating Surgical Site Infection Abstraction With a Semi-automated Machine-learning Approach. *Annals of Surgery*.

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2018). CLAMP a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association : JAMIA*, *25*, 331 – 336.

Stagg, V. (2015). *ELIXHAUSER: Stata module to calculate Elixhauser index of comorbidity*. Published: Statistical Software Components, Boston College Department of Economics.

Stephens, K. A., Au, M. A., Yetisgen, M., Lutz, B., Suchsland, M. Z., Ebell, M. H., & Thompson, M. (2020). Leveraging UMLS-driven NLP to enhance identification of influenza predictors derived from electronic medical record data. `http://biorxiv.org/lookup/doi/10.1101/2020.04.24.058982`.

Tignanelli, C., Silverman, G., Lindemann, E., Trembley, A., Gipson, J., Beilman, G., Lyng, J., Finzel, R., McEwan, R., Knoll, B., Pakhomov, S., & Melton, G. (2020). Natural Language Processing of Prehospital Emergency Medical Services Trauma Records Allows for Automated Characterization of Treatment Appropriateness. *Journal of Trauma and Acute Care Surgery*, 607–614.

Turner-Musa, J., Ajayi, O., & Kemp, L. (2020). Examining Social Determinants of Health, Stigma, and COVID-19 Disparities.. *Healthcare (Basel, Switzerland)*, *8*(2).

Uzuner, O., South, B. R., & Duvall, S. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556.

Wang, Z., & Tang, K. (2020). Combating COVID-19: health equity matters.. *Nature medicine*, *26*(4), 458. Place: United States.

Wen, A., Fu, S., Moon, S., El Wazir, M., Rosenbaum, A., Kaggal, V. C., Liu, S., Sohn, S., Liu, H., & Fan, J. (2019). Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ digital medicine*, *2*, 130.

Wollenstein-Betech, S., Cassandras, C. G., & Paschalidis, I. C. (2020). Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: Hospitalizations, mortality, and the need for an ICU or ventilator. *International Journal of Medical Informatics*, *142*, 104258.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Dahly, D. L., Damen, J. A. A., Debray, T. P. A., de Jong, V. M. T., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Martin, G. P., McLernon, D. J., Andaur, C. L., Reitsma, J. B., Sergeant, J. C., Shi, C., Skoetz, N., Smits, L. J. M., Snell, K. I. E., Sperrin, M., Spijker, R., Steyerberg, E. W., Takada, T., Tzoulaki, I., van Kuijk, S. M. J., van Bussel, B., van Royen, F. S., Verbakel, J. Y., Wallisch, C., Wilkinson, J., Wolff, R., Hooft, L., Moons, K. G. M., & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal.. *BMJ (Clinical research ed.)*, *369*, m1328.

Yang, X., Bian, J., Fang, R., Bjarnadottir, R. I., Hogan, W. R., & Wu, Y. (2019). Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *Journal of the American Medical Informatics Association*, *27*(1), 65–72.

Řehůřek, R. (2016). gensim.models.Word2Vec.most_similar. `https://tedboy.github.io/nlps/generated/generated/gensim.models.Word2Vec.most_similar.html`.