# COHORT IDENTIFICATION FROM FREE-TEXT CLINICAL NOTES USING SNOMED CT'S SEMANTIC RELATIONS

Eunsuk Chang

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Carolina Health Informatics Program (CHIP).

Chapel Hill
2022

Approved by:

Javed Mostafa

Fei Yu

Emily R. Pfaff

David D. Potenziani

Kimberly Robasky

# ABSTRACT

Eunsuk Chang: Cohort Identification from Free-Text Clinical Notes Using SNOMED CT's
Semantic Relations
(Under the direction of Javed Mostafa)

In this paper, a new cohort identification framework that exploits the semantic hierarchy of SNOMED CT is proposed to overcome the limitations of supervised machine learning-based approaches. Eligibility criteria descriptions and free-text clinical notes from the 2018 National NLP Clinical Challenge (n2c2) were processed to map to relevant SNOMED CT concepts and to measure semantic similarity between the eligibility criteria and patients. The eligibility of a patient was determined if the patient had a similarity score higher than a threshold cut-off value, which was established where the best F1 score could be achieved. The performance of the proposed system was evaluated for three eligibility criteria. The current framework's macro-average F1 score across three eligibility criteria was higher than the previously reported results of the 2018 n2c2 (0.933 vs. 0.889). This study demonstrated that SNOMED CT alone can be leveraged for cohort identification tasks without referring to external textual sources for training.

I dedicate my work to my beloved wife, Boreum, who has been a constant source of support and encouragement during the challenge of graduate school and life. I am truly thankful for having you in my life. I also dedicate this dissertation to my son, Minjay, who is the greatest gift I have ever been blessed with.

## ACKNOWLEDGEMENTS

I cannot express enough thanks to my committee for their continued support and encouragement: Dr. Javed Mostafa, my committee chair; Drs. Fei Yu, Emily Pfaff, David Potenziani, and Kimberly Robasky. I offer my sincere appreciation for the learning opportunities provided by my committee. Special thanks to Javed, my advisor, whose understanding of my situation has made my navigation in academia possible and has made me a better researcher. Without his assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. I would like to thank you very much for your support and understanding over these past three years.

Getting through my dissertation required more than academic support, and I have many, many people to thank for listening to and, at times, having to tolerate me. My completion of this project could not have been accomplished without the support of my colleagues in LAIR and my child, Minjay—thank you for allowing me time away from you to research and write.

Finally, I must express my very profound gratitude to my parents and to my wife for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this dissertation. This accomplishment would not have been possible without them. Thank you.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AIDS | Acquired immunodeficiency syndrome |
| API | Application programming interface |
| AUI | Atom Unique Identifier |
| CAD | Cardiovascular disease |
| CDSS | Clinical decision support system |
| CNN | Convolutional neural network |
| CUI | Concept Unique Identifier |
| EHR | Electronic health record |
| FDA | Food & Drug Administration |
| FFF | Fully connected feedforward |
| FHIR | Fast Health Interoperability Resource |
| FN | False negative |
| FP | False positive |
| FSN | Fully specified name |
| HL7 | Health Level 7 |
| i2b2 | Informatics for Integrating Biology to the Bedside |
| ICD | International Classification of Disease |
| IC | Information content |
| LOINC | Logical Observation Identifiers, Names, and Codes |
| LSTM | Long Short-Term Memory |
| LUI | Lexical Unique Identifier |

| | |
|---|---|
| NLP | Natural language processing |
| MedDRA | Medical Dictionary for Regulatory Activities |
| MeSH | Medical Subject Heading |
| n2c2 | National NLP Clinical Challenges |
| PML | Progressive multifocal leukoencephalopathy |
| RNN | Recurrent neural network |
| SUI | String Unique Identifier |
| SVM | Support vector machines |
| TN | True negative |
| TP | True positive |
| TREC | Text Retrieval Conference |
| TURP | Transurethral resection of the prostate |
| UIMA | Unstructured Information Management Architecture |
| UMLS | Unified Medical Language System |

# I. INTRODUCTION

## 1.1 Problem Statement

Selecting the appropriate subjects and sample population is a critical step to a successful clinical trial. Clinical trial designers need to ensure that the recruited patients satisfy the inclusion and exclusion criteria of the trial to eliminate confounding factors and to avoid the study being underpowered. However, one of the major challenges to the timely conduct of research is patient recruitment for clinical trials (Fletcher et al. 2012; Mitchell et al. 2014; Treweek et al. 2013). Because of the insufficient number of participating patients, recruitment difficulties often end up with many abandoned or underpowered clinical trials (Schroen et al. 2010). Revoked or delayed clinical trials even claim patients' lives; on one calculation, delays in recruitment to clinical trials of streptokinase kinase in myocardial infarction led to as many as 10,000 unnecessary death in the U.S (Collins 1992). A patient with a poor prognosis and limited treatment options would have been able to, if eligible, access safe and efficient treatment in a late-stage trial if the patient could be identified and recruited.

To fill the gap between strict inclusion/exclusion criteria and difficulties in the recruitment of patients, efforts have been made to identify eligible patients from electronic health records (EHRs) (Hernandez et al. 2015; Jensen et al. 2012; Mc Cord & Hemkens 2019). EHRs have attracted the attention of researchers and clinicians as a potential tool to accelerate recruitment by examining a large number of clinical records (Thadani et al. 2009; Miotto and

Weng 2015), generally searching for discrete data points such as age or laboratory results. A key challenge, however, is that more detailed information about medical conditions is often embedded in the extensive occurrence of clinical narratives in EHRs in the form of unstructured text (Small et al. 2017; Afzal et al. 2018; Ford et al. 2016; Kossovsky et al. 1999), and some kinds of patient information are accessible exclusively through the unstructured parts of EHRs. For example, social history, family history, temporal context (e.g., a procedure has been *undergone* or *planned*), and treatment modality (e.g., aspirin to *prevent* stroke or to *treat* acute myocardial infarction) can be found in free-text clinical notes only and cannot be accessed via coded or structured data. Therefore, recruiting staff must carefully read patient records in order not to miss relevant information and potential subjects for the trial.

The manual workload burden for manual screening of patient records is one of the major obstacles to successful research. Manual screening of clinical data is not only time-consuming but also expensive: according to one estimate, the cost of manual screening can cost up to $336.48 per subject for cancer clinical trials when including patient consenting and access to external, related medical records (Penberthy et al. 2012). Clinical trial recruitment is additionally hampered by the fact that most clinical practices lack the staffing necessary for manual patient screening (Ni et al. 2015).

To reduce or avoid the workload and cost of labor-intensive manual screening, many automated tools to identify patient cohorts from free-text clinical records have been developed (Beauharnais et al. 2012; Ni et al. 2015). One problem with using unstructured patient records is that EHRs are not designed with the purpose of patient recruitment in mind (Bache et al. 2013), and information is inconspicuously embedded in free text. Clinical notes are often written in an ungrammatical way, and document formats and styles vary by institutions and physicians, which makes the extraction of important patient data arduous.

Various natural language processing (NLP) techniques have been proposed to process the unstructured texts and improve the accuracy of patient identification from a large collection of clinical notes in EHRs in a fast and effective way. To date, supervised machine learning NLP models—defined here as models trained by learning a mapping between input variables and the output variables (i.e., labels) which were annotated by human experts—which learn relationships between words from large corpora of documents have been rigorously studied (Köpcke et al. 2013; Ni et al. 2019; Chakrabarti et al. 2017; Xu et al. 2011). Supervised learning denotes various algorithms that generate a function that maps inputs to desired outcomes by looking at several input-output examples of the function. Supervised learning approaches are successfully used when labels of each data point are available. However, some important issues exist with the supervised learning NLP when it comes to cohort selection tasks.

First, the large amount of time and labor required to annotate and train a machine learning model is an ongoing concern for the NLP community (Spasic & Nenadic 2020). Traditional supervised learning requires an extensive annotation and labeling of training data. The large amount of human labor and cost for annotation has limited the size of training data available, which significantly hampers the validity of supervised machine learning models (Spasic and Nenadic 2020). The use of structured data (e.g., International Classification of Disease (ICD) and Logical Observation Identifiers, Names, and Codes (LOINC) codes) as labels may not be helpful for annotation efforts, as those structured data are primarily used for reimbursement and cannot be harmonized with clinical notes. When it comes to biomedical corpora, the recurring expense for hiring highly specialized biomedical professionals to annotate data further impedes building large validated datasets.

Second, supervised learning NLP techniques offer no hard evidence about the generalizability of the machine learning models. As the provenance of data available for training is confined to only a small number of contributing institutions, the trained model is often not readily generalizable to other institutions.  It has repeatedly been reported that there has often been a significant drop in performance when a system is trained in one institution and tested in another (Napolitano et al. 2016; Ye et al. 2017). In addition, available labels can serve only one research question at a time and may not be applied to another. For example, a dataset at hand has labeled binary classification to data (e.g., a patient is currently smoking or not) but a researcher may want a multinomial analysis of outcomes (e.g., current smoker, former smoker, never smoker, or unknown). In this case, the dataset needs to be re-annotated from the very beginning, nullifying previous efforts to label such extensive data. Although computational phenotyping and harmonization efforts across heterogenous clinical records have been an important research topic for the last decade (Pacheco et al. 2018), there has been a limited accomplishment in harmonizing information from the unstructured free-text part of EHRs.

Lastly, many researchers have recently drawn attention to the errors and biases embedded in data and models (Geiger et al. 2021). As the contemporary effort to build large data often involves automatic labeling or crowdsourcing, data are increasingly susceptible to labeling errors (Sambasivan et al. 2021). In an analysis of 10 test sets from datasets that include ImageNet, an image database used to train countless computer vision algorithms, the authors found an average of 3.4% label errors across all of the datasets (Northcutt et al. 2021). In addition to labeling errors, datasets are not emancipated from bias encoded within the corpus when it comes to textual data (Bender et al. 2021; Kurita et al. 2019). For example, a hospital may have a disproportionally high prevalence of specific diagnoses, or a doctor may present unique text features because of the hospital's practices or disposition to use specific jargon or parlance (Sohn et al. 2018). Labeling errors and biased distribution of cases could lead researchers to draw incorrect conclusions about which models perform best in the real world, potentially undermining the framework by which the researchers benchmark machine learning systems.

Other than the above-mentioned problems that arise from supervised learning NLP with annotated datasets, another problem worth noting is that many previous NLP systems were intended to solve a single NLP subtask. Cohort identification NLP tools generally involve multiple subtasks, such as named entity recognition (Chen et al. 2015; Wu et al. 2017), negation identification (Peng et al. 2018; Sohn et al. 2012), and assertion detection (Minard et al. 2011; de Bruijn et al. 2011). When these subcomponents are integrated into one cohort identification system, their performance may be suboptimal because each subcomponent has proved its efficacy only on its own training data, which often comes from general language domains such as Wikipedia. Furthermore, those individual subtask NLP components are not designed to coordinate with others. As knowledge elicited from clinical notes goes through multiple, heterogeneous NLP pipelines in a cohort identification system, interactions among NLP subtasks could undermine the performance of the system. For example, when using a negation detection algorithm that looks for a preposition "without" in order to eliminate negated entities in a sentence (e.g., to eliminate *orthopnea* in "dyspnea without orthopnea"), the algorithm will fail to retrieve the concept "MRI without contrast" by eliminating the concept's description-level word "contrast." The negation and entity extraction subtasks need to be integrated to prevent distorting or losing the original meaning of the original text.

## 1.2 Significance

There are two distinguished approaches to artificial intelligence: knowledge-based and machine learning-based approaches. Researchers that solve reasoning problems using knowledge-based tools are sometimes at odds with researchers that instead use machine learning tools. The machine learning approach extracts information directly from historical data and extrapolates it to make predictions. It automatically builds a classifier by learning the characteristics of each category from a set of labeled examples. It is highly dependent on collective intelligence that is housed at the present time in the general public who are rarely experts in the field. In contrast, the knowledge-based approach, which is represented by ontologies, uses a team of experts to try to encode and construct a large number of the properties of the world. Although each approach has pros and cons, the following points favor the ontology-based approach.

### 1.2.1 Less Effort to Create and Maintain Knowledge

The semantic structures of ontologies are constructed by human experts. While *data* is supervised by humans (e.g., by the annotation and labeling of data) in supervised machine learning, it is the *internal structure* of knowledge that is supervised by humans in ontology-based approaches. In this regard, the knowledge structure (e.g., identified relations among entities) is consistent and explicit in ontologies, while it is variable and subject to change (depending on training data) in supervised learning systems.

Although an ontology, too, needs extensive human efforts to construct such a large knowledge resource, it is generalizable and can be repeatedly leveraged to answer discrete research questions in a variety of settings without the cost of retraining on a large dataset. Once meticulously constructed, ontology systems save clinical researchers from the labor-intensive and error-prone annotation process required for creating training data specific to new projects. The more those ontologies are used, the more likely the benefit from the generalizability feature of ontologies exceeds the cost of constructing and maintaining ontology-based systems.

When a new kind of knowledge emerges, an ontology can accommodate it instantly by creating a new concept for it. Ontologies, in conjunction with other artificial intelligence-driven data analytics tools, can find new patterns and trends in data. By categorizing identified direct relations to a causality relation ontology, ontologies can help with early hypothesis testing in pharmaceuticals. Emerging infectious diseases such as COVID-19 can be reflected in the knowledge structure of ontologies as soon as it concerns medical professionals. By a machine learning approach, in contrast, COVID-19 cannot be predicted by the systems until a sufficient number of cases are accumulated. To make COVID-19 predictable by the machine learning-based system, a large number of cases and non-cases should be collected before they are labeled as such. Under these circumstances, rapidly changing conditions and sparse data, ontologies can be more readily used than machine learning

Cohort identification tasks are sometimes required to identify patients with rare or newly emerging cases. Acute discovery of pathophysiology, natural history, and epidemiology of newly emerging diseases is crucial for cohort identification systems to prove their utility in that the essence of clinical research is observing a fact in clinical settings by way of safe experimentation and data.

### 1.2.2 Generalizability

Beyond terminology standardization, ontologies can contribute to the generalizability of knowledge obtained from one institution to another by providing reliable and consistent semantic relations among biomedical entities (Wang et al. 2018). Ontologies made this possible by providing a formal, explicit specification of a shared conceptualization of biomedical entities (Gruber 1993).

Biomedical ontologies provide layout for knowledge-sharing among different institutions (Pundt and Bishr 2002). Standard biomedical terminologies and ontologies are advocated by the U.S. government to be used in favor of semantic interoperability of patient information among healthcare organizations and institutions (D'Amore et al. 2014). In addition to interoperability at the inter-institutional level, ontologies can help health professionals constantly deal with ever-changing knowledge in health care at the individual level (Wongthongtham & Zadjabbari 2012).

Besides coded data, biomedical entities embedded in unstructured parts of EHR can be sharable if they are mapped to standard ontologies. Many standard ontologies offer attribute relations (e.g., *has_indirect_procedure_site*) as well as hierarchical relations (e.g., *is_a* relation) among biomedical entities. If the context and meaning of free-text clinical narratives are consistently expressed by semantic relations of ontology concepts, the knowledge extracted from one institution is readily generalizable to another, or a third party can easily merge knowledge from multiple institutions to conduct analysis. For purpose of illustration, assume that a researcher wanted to identify patients who had undergone any procedure on the pancreas from Hospitals A and B. He can identify Patient 1 who underwent the Whipple procedure at Hospital A and Patient 2 who underwent total pancreatectomy at Hospital B because the two procedures share the common attribute value of *procedure site – pancreatic structure*, even though the two procedure names are not lexically related.

### 1.2.3 Less Bias in NLP Data Models

An ontology-based approach exploits the formal semantic structure of the ontology to represent patients (e.g., such as building vectors to represent patient records) instead of learning patterns from training data. Cohort identification tasks can become more reliable when bias in data is avoided. If patients with acquired immunodeficiency syndrome (AIDS) do not present with progressive multifocal leukoencephalopathy (PML) in Hospital A, there is no way for supervised machine models to learn that PML is a complication of AIDS. Thus, the model cannot infer that a patient with PML in Hospital B would have already been diagnosed with AIDS. If we exploited an ontology, on the other hand, the model would be aware that rare complications like PML were associated with AIDS because it is explicitly described in the ontology that PML *is a* complication of AIDS, even though there were no such cases in the hospital.

### 1.2.4 Quantifying Eligibility Using Similarity Measure

In addition to its ability to represent relations between entities embedded in EHRs, ontologies can be used to measure the similarity between an eligibility criterion and a clinical note. Supervised machine learning-based algorithms are limited in this task because the eligibility criteria in general are short in length and cannot provide enough features that characterize patients.

The asymmetry in length between eligibility criteria and clinical notes can be overcome by the knowledge-based approach. In ontology, each concept is logically described through the relationships (e.g., *is_a*) to other concepts. Ontology's ability to add relations between concepts and synonyms, hypernyms, and hyponyms has proven effective at mapping vector-represented terms and is an ideal source to measure the semantic distance (and similarity) between medical concepts. The semantic similarity measure provides more advantages to cohort identification tasks by providing a quantitative degree of eligibility of each patient (Alonso and Contreras 2016; Mabotuwana et al. 2013). Beyond a binary classification of eligibility (e.g., predicting a patient is "eligible" or "not eligible"), the quantitative estimation of the degree of eligibility can help clinical trial recruiting staff determine how extensively they will include potential subjects.

## 1.3 Purpose of the Study

To address the aforementioned challenges of supervised learning NLP models, my research focuses on a knowledge-driven, ontology-based information retrieval framework for identifying a cohort of patients from unstructured clinical notes in EHRs.

The primary purpose of this study is to propose a cohort identification framework that does not demand annotated training data for supervised learning. No subcomponent of the proposed framework requires datasets for training or learning. Instead, the proposed framework makes use of a semantic structure of ontologies. The current study will demonstrate the semantic structure of SNOMED CT alone can be exploited for cohort identification tasks.

SNOMED CT is one of the most comprehensive biomedical ontology systems (Millar 2016). It is made up of a set of concepts, descriptions, and relations that serve as a common reference layout for comparing and aggregating data from multiple individuals, systems, or institutions concerning the entire health care process. SNOMED CT is primarily used in EHRs to capture, store, and access clinical data of patients, and it has recently been adopted as a tool to code and classify unstructured medical narratives, working alongside various NLP and machine learning techniques in broad fields of biomedicine owing to its broad coverage of biomedical entities (Chang and Mostafa 2021).

The current study will present that the performance of a cohort identification framework based on the exclusive use of SNOMED CT's hierarchical semantic structure is on a par with, if not superior to, the currently best available cohort identification systems, most of which use supervised machine learning classifiers. To the best of my knowledge, this is the first attempt to depend exclusively on SNOMED CT's hierarchies for a cohort identification task.

A cohort identification task includes multiple subtasks, including but not limited to concept extraction, assertion detection, time-related information extraction, laboratory test result extraction, inference, and meta-information extraction. Of note, the current study focuses on the extraction of concepts for relevant symptoms, disorders, procedures, anatomical sites, or medications from free-text clinical notes for cohort selection. Assertions and time-related information were extracted from free texts using lexical- and syntactic-level rules. Cohort identification tasks examined in this study do not involve the extraction of laboratory test results (as in "HbA1c greater than 7%") or inference (as in "patient speaks English"), which requires task-specific algorithms trained from datasets.

The rest of the paper is structured such that a brief background and related work are presented in Chapter II, followed by the description of the design of the proposed cohort identification framework in Chapter III and the presentation of experimental results of the proposed framework in Chapter IV.

# II. BACKGROUND AND RELATED WORK

## 2.1 Introduction

### 2.1.1 Ontologies and Standard Terminologies

Terminology systems provide a standardized meaning of technical language (Schulz et al. 2019). Terminology systems such as Medical Subject Heading (MeSH) and ICD-10 define semantic relations among pre-existing terms and intend to coerce entities to be classified into non-overlapping categories to enable statistical analysis of disease epidemiology. On the other hand, ontologies aim at the categorization of objects and description of their relationships by logic-based axioms and multi-hierarchical structure. They provide formal reasoning support for knowledge representation by identifying terms and enabling concepts to be related to each other (Schulz et al. 2019). Ontology-controlled vocabularies have the potential to promote multiple levels of integration in medical research and practice, enabling data integration, knowledge integration, information integration, and interdisciplinary integration (Liaw et al. 2014).

### 2.1.2 SNOMED CT

SNOMED CT is a reference terminology which provides context-free, well-defined representations of entities of a domain. Providing a common reference point, SNOMED CT plays an essential role in achieving interoperability among heterogeneous healthcare systems and institutions by offering an accepted collection of coordinated reference ontologies/terminologies and harmonized development processes (Blobel 2010). What differentiates it from aggregation terminologies such as ICD-10 is that SNOMED CT describes the classes of entities within a multi-hierarchical semantic structure, whereas aggregation terminologies enforce the principles of single hierarchies and disjoint classes (Schulz et al. 2017).

SNOMED CT has gained momentum to become a prevalent clinical terminology system when the U.S. federal government required in 2013 that SNOMED CT be included in EHR systems in order for them to be certified for Stage 2 Meaningful Use (National Library of Medicine 2016). SNOMED CT-coded patient data helped efficiently identify significant facts in the oceans of data, enabling effective meaning-based retrieval and linking the EHRs to authoritative clinical knowledge. SNOMED CT can be used as a code system to store clinical information; an interface terminology to capture and display clinical information; an index system to retrieve clinical information; a common terminology to communicate in a meaningful way; a dictionary to query, analyze and report, and link to knowledge resources; and extensible foundation to represent new types of clinical data (SNOMED CT International 2021).

### *2.1.2.1 SNOMED CT content components*

SNOMED CT is made up of a set of concepts, descriptions, and relations that serve as a common reference layout for comparing and aggregating data from multiple individuals, systems, or institutions concerning the entire health care process. The central component in SNOMED clinical terms is the concept, which is a clinical idea associated with a unique identifier. The meaning of that idea is specified by an association with a term that is known as the fully specified name (FSN) and the link between that identifiers; the meaning of that clinical idea is permanent and unchangeable. Descriptions provide terms that are human-readable and allow the meaning of the concept to be seen by users. Each concept is associated with several descriptions, and each description has an associated human-readable term. There are several types of description, each with a particular value. All concepts have at least one FSN and at least one synonym, which are the main description types of SNOMED CT. In Figure 1, the SNOMED CT concept 23069007 has the FSN of "Cerebrovascular accident (disorder)," and synonyms of "CVA – Cerebrovascular accident" and "Stroke."

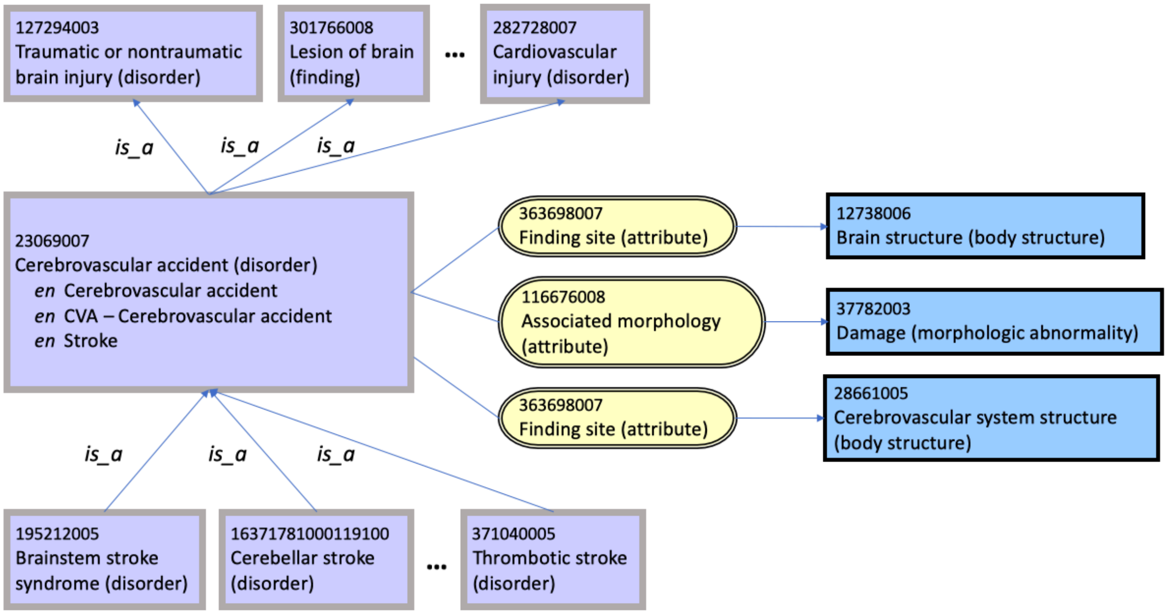

127294003
Traumatic or nontraumatic brain injury (disorder)
301766008
Lesion of brain (finding)
...
282728007
Cardiovascular injury (disorder)
is_a
is_a
is_a
23069007
Cerebrovascular accident (disorder)
en Cerebrovascular accident
en CVA – Cerebrovascular accident
en Stroke
363698007
Finding site (attribute)
12738006
Brain structure (body structure)
116676008
Associated morphology (attribute)
37782003
Damage (morphologic abnormality)
363698007
Finding site (attribute)
28661005
Cerebrovascular system structure (body structure)
is_a
is_a
is_a
195212005
Brainstem stroke syndrome (disorder)
16371781000119100
Cerebellar stroke (disorder)
...
371040005
Thrombotic stroke (disorder)

**Figure 1.** Diagram of hierarchical and attribute-range relations of the target concept **230690007 | Cerebrovascular accident (disorder) |** of SNOMED CT. Purple squares delineate *is_a* axial semantic relations. Double-circled yellow ovals denote attributes. Blue square represents their corresponding value concepts.

Each SNOMED CT concept is associated with other concepts by a set of relations that expresses the characteristics of a concept. Those relations are crucial to the computer processability of SNOMED CT data. A concept might be a subtype of another concept or may have a particular attribute that has a value that is provided by another concept. Subtype relationships are known as *is_a* relationships in SNOMED CT and create a hierarchy linking each concept to more general concepts. The taxonomic hierarchical *is_a* relations enable retrieval of specific concepts in response to more general queries. On the other hand, attribute relations provide additional defining information about a concept and why that concept is different from its supertypes. In the Figure 1 example, *Cerebrovascular accident* is a subtype of *Traumatic or nontraumatic brain injury* and has a finding site (attribute) which is *brain structure*.

SNOMED CT also supports post-coordination to combine established concepts to represent more complex concepts. When it comes to mapping biomedical entities in free text into SNOMED CT codes, post-coordination makes it possible to construct logical expressions for complicated entities that are not defined by pre-coordinated expressions. Using logical axiom, post-coordination makes the expression of biomedical entities expandable with fewer pre-existing concepts and thus requires a lower terminology maintenance burden.

## 2.2 Literature Review of SNOMED CT Use

In this section, the recent use of SNOMED CT will be thoroughly reviewed, focusing on previous literature on the use of SNOMED CT for classifying or coding clinical documents and previous cohort identification systems to select patients from free-text clinical narratives. To present the up-to-date trend in the topic, only those published after 2013 will be discussed. Parts of this section have been published elsewhere (Chang and Mostafa 2021)

### 2.2.1 SNOMED CT's Content Coverage

The extensive coverage for biomedical entities by SNOMED CT makes SNOMED CT-based systems such as clinical decision support systems (CDSSs) and cohort identification systems comprehensively applicable to a broad spectrum of diseases and specialties. Since SNOMED CT is such a large terminology system, the granularity of a SNOMED CT's concept coverage is highly dependent on the source of text and domain. The coverage of SNOMED ranged from 22.5% for video portal tags (Konstantinidis et al. 2013) to 99.9% for clinical charts (Oluoch et al. 2015). The content coverage of SNOMED CT usually examined along with other standardized terminologies. Though SNOMED CT was generally considered the most comprehensive terminology, this terminology was outperformed by MeSH for herbal and dietary supplement terms (Manohar et al. 2015), by RadLex for radiology text (Kahn 2014), by NCI Thesaurus for cancer description (Schulz et al. 2019), and by Omaha System for social determinants of health (Monsen et al. 2018). The low coverage for some areas of biomedicine may be caused by the enormous size of the semantic structure of SNOMED CT; if the size of the terminology is large, mapping biomedical entities to standardized vocabularies becomes less accurate because coders have more choices to map (Schulz et al. 2019).

### 2.2.2 SNOMED CT for Classifying or Coding Clinical Documents

The semantic capabilities of SNOMED CT to normalize data components can integrate data from heterogeneous sources. The semantic interoperability features of SNOMED CT-coded entities—rather than words—can be represented as features of classification tasks.

Biomedical concepts in free text can be extracted and coded by SNOMED CT concepts. In such cases, SNOMED CT codes and semantic relations were coupled with machine learning text mining algorithms like support vector machines (SVM) (Butt et al. 2013; Koopman et al. 2015; Koopman et al. 2018; Zolnoori et al. 2019; Greenbaum et al. 2019), neural networks (Arbabi et al. 2019; Arguello-Casteleiro et al. 2019; Banerjee et al. 2019; Lerner et al. 2020; Peterson and Liu 2020), decision trees/random forest (Ma and Weng 2016; Liaqat et al. 2020; Petrova et al. 2015), or ensemble/boosting algorithms (Mujtaba et al. 2018; Lin et al. 2017) to map biomedical entities embedded in free texts into SNOMED CT concepts. In other cases, open-source or commercial NLP applications such as MetaMap (Plaza 2014; Barros et al. 2018; Shen et al. 2018) and cTAKES (Sanchez Bocanegra et al. 2017; Meystre et al. 2014; Valtchinov et al. 2020) were used to extract or map clinical phenotypes to SNOMED CT. Manual extraction and mapping of the SNOMED CT concepts were conducted in (Chan et al. 2017; Lingren et al. 2014; Karimi et al. 2015; Lindman et al. 2019). SNOMED CT concepts extracted from those free texts were then fed into algorithms that classified clinical data elements based on organ system (Hier and Pearson 2019), cause of death (Roldán-García et al. 2016; Ternois et al. 2019), or additional need attributed to patients (Hitchins & Hogan 2018).

Other terminology systems that were leveraged in combination with SNOMED CT for annotating or mapping biomedical phenotypes include ICDs (e.g., ICD-9, ICD-10, or ICD-10-CM) (Koopman et al. 2015; Ternois et al. 2019), MeSH (Martínez García et al. 2015; Ruch et al. 2008), LOINC (Zvára et al. 2017; Hostetter et al. 2015), RxNorm (Sohn and Liu 2014), and Medical Dictionary for Regulatory Activities (MedDRA) (Karimi et al. 2015). The Unified Medical Language System (UMLS) was adopted to produce mapping among multiple ontologies

(Pradhan et al. 2015; Soysal et al. 2018; Becker and Böckmann 2017).

F1 scores, often accompanied by recall and precision scores, were used most frequently to report the performance of machine learning tools (Grundel et al. 2021; Tahmasebi et al. 2019; Siefridt et al. 2020; Abidi et al. 2016) because they best represent a balance between precision and recall. The free text was annotated with SNOMED CT codes to create a corpus or data library (Pearce et al. 2019). Recent attempts were made to analyze Twitter mentions of disease concerns (Barros et al. 2018) and to encode COVID-19-related clinical phenotypes (Jani et al. 2020). In one study, instead of encoding free text, SNOMED CT synonyms were tokenized and back-incorporated into a corpus to expand word features in common English (Hagen et al. 2013). Another study represented lexical expressions such as ambiguity and negation cues with SNOMED CT concepts (Velupillai et al. 2014).

### 2.2.3 Implementation of SNOMED CT

SNOMED CT codes were employed by data models to provide the means of terminology standardization, which were then integrated into data registries/repositories (Asgari et al. 2013; Asklund et al. 2015; Guien et al. 2018; Lomonaco et al. 2014; Morash et al. 2018; Norton et al. 2016; Park et al. 2016; Pandey et al. 2019; Banda et al. 2016), query/terminology services (Elkin et al. 2018; Lamy et al. 2015; Metke-Jimenez et al. 2018; Silva Layes et al. 2019; Song et al. 2014; Sun et al. 2015), and knowledge-based CDSSs (Danahey et al. 2017; Greibe 2013; Müller et al. 2019; Maheronnaghsh et al. 2013; Hendriks et al. 2019). Applications were built to query structured data such as demographics, medications, and laboratory test results, as well as unstructured data in clinical notes using SNOMED CT (Elkin et al. 2018). Other use cases involved identifying cases of a specific disease (i.e., congestive heart failure, bladder cancer, etc.) based on SNOMED CT codes, either from free-text clinical notes (Wang et al. 2015) or structured data registries (Rasmussen et al. 2018). The medical domains of registries included cancer (Asgari et al. 2013; Pedrera et al. 2020; Tomic et al. 2015), adverse drug effects (Banda et al. 2016; Knowledge Base Workgroup of the Observational Health Data Sciences and Informatics (OHDSI) collaborative 2017), cardiovascular diseases (Mohd Sulaiman et al. 2017; Pandey et al. 2019), and gene information (Lomonaco et al. 2014; Park et al. 2016). Terminology services and application environment adopted the Health Level 7 (HL7) standards such as the Fast Health Interoperability Resource (FHIR) to access the terminology and facilitate the exchange of semantic messages (Abhyankar et al. 2015; Ali et al. 2017).

### 2.2.4 Retrieval of Patient Data Using SNOMED CT

Researchers conducted retrospective observational studies to collect patient data using SNOMED CT codes. They identified and retrieved retrospective study subjects according to the inclusion criteria for each study.

To obtain more patient information than codified SNOMED CT concepts can provide, attempts had been made to map biomedical entities in free-text clinical narratives into structured SNOMED CT expressions. Since new clinical ideas can be expressed using existing SNOMED CT concepts by identifying their relations, free-text phrases or sentences can be formulated by the SNOMED CT relations structure. Various automated approaches were developed to convert free-text clinical notes (Li 2018; Patrick et al. 2007), problem descriptions (Peterson and Liu 2020), and phases (Kate 2013) into SNOMED CT codes. By extracting patient information embedded in unstructured parts of EHRs, automated mapping of free text inputs into SNOMED CT provides means to provide more granular information about the patient.

### 2.2.5 Conclusion

Owing to its broad coverage for biomedical entities, SNOMED CT is widely used for applications in various settings. The semantic capabilities of SNOMED CT can integrate data from heterogeneous sources. Semantic interoperability provided by SNOMED CT has facilitated the exchange of biomedical data.

## 2.3 Review of Semantic Similarity Metrics

### 2.3.1 Semantic Similarity Using Taxonomic Structure of Biomedical Ontology

Controlled terminologies, combined with the UMLS or not, and the proliferation of textual resources (e.g., free-text clinical notes in EHR) in health care offered a rich resource for creating automated methods to measure semantic similarity between concepts (Pakhomov et al. 2010). The taxonomic structure of biomedical ontology/vocabularies has been used to quantify similarity because it does not depend on the availability of large corpora as opposed to classical approaches that assess document similarity based on term appearance probabilities in corpora.

Previous works in the general language domain suggest that corpus-based distributional measures of similarity suffer from limitations that stem from the corpora's imbalance, sparseness, and textual ambiguity (Budanitsky & Hirst 2006; Lin 1998). More recent studies in the biomedical domain (Sánchez & Batet 2011; Zare et al. 2015) showed that ontology-based measures such as intrinsic information content (IC) outperformed the corpus-based approaches.

Patient similarity investigates distances between a variety of components of patient data, including semantic similarity measures in free-text clinical notes, and determines methods of clustering patients based on short distances between some of their characteristics. Similarity metrics can be utilized for systematic identification of a subset of patients to improve prediction of treatment outcomes, identify cohorts in clinical trials, and help clinicians manage particular patients by referring to similar previous cases (Lee & Das 2010; Lee et al. 2015).

### 2.3.2 Measuring Semantic Similarity Between Two Biomedical Concepts

The measures of semantic similarity are critical in cohort identification tasks in which rigorous completeness in information retrieval is required for the classification of textual data (Pedersen et al. 2007; Sánchez & Batet 2011). The estimation of similarity between individual concepts is an essential component of measuring similarity between documents.

Taxonomic concepts in knowledge sources and ontologies are usually organized hierarchically. Two main hierarchies in SNOMED CT are parent-child, which is *is_a* hierarchy, and broader-narrower relations (McInnes & Pedersen 2015; Zare et al. 2015). In such a hierarchical structure, concepts under the direct parent will be more similar than those which share a more distant parent.

Hadj Taieb and his colleagues (2014) suggested that when developing similarity metrics between two concepts, three important features need to be addressed: depth, hyponyms, and leaves. The depth levels of concepts play an essential role in defining the semantic similarity because a child concept specifies the meaning of its parent concept while going down from the root concept towards a target concept. Hyponyms are significant in determining the specificity of a concept because a concept with a greater number of hyponyms (i.e., subtypes that are direct and indirect descendants) is less specific. This implies that a concept is premature to be fully defined at its current level if many hyponyms are subsumed by the concept deep in the hierarchy.

As the computation burden increases proportional to the number of hyponyms, which again increases exponentially down to the hierarchy, the number of leaves of a concept is a reasonable proxy to the number of hyponyms.

The semantic similarity between two taxonomic concepts can be estimated by three broad approaches: probabilistic, path-based, and information-theoretic approaches (Jia et al. 2019). Probabilistic approaches consider the frequency distribution of the concept in a set and are generally used for categorical data. Path-based approaches account for information about the co-location of the terms in taxonomy and measure the distance between the links relating to them. Information-theoretic approaches measure the semantic similarity according to the information the two concepts contain. Originally, the IC in computational linguistics quantifies the amount of information embedded in a concept appearing in a discourse (Sánchez et al. 2011). In this section, the last two approaches for semantic similarity measurement will be discussed, for the current work is to exploit the taxonomic structure of SNOMED CT.

#### *2.3.2.1 Path-based measures*

Caviedes and Cimino (2004) developed similarity metrics based on the shortest path (*spath*) between two concepts in the semantic structure of the UMLS. They evaluated their metrics with parent-child relations clusters drawn from SNOMED CT, MeSH, and ICD-9-CM, among which were mapped by a subset of the UMLS. SNOMED CT's correlation with expert scores was 0.60 and improved to be 0.79 when coupled with MeSH.

Pedersen et al. (2007) suggested a modification of the above method, called *Path Measure* or *Path Length (Path)*, which is calculated as the reciprocal of the length of the shortest path in *is_a* relationship in SNOMED CT.

$$sim_{Path}(c_1, c_2) = \frac{1}{spath(c_1,c_2)} \tag{1}$$

Leacock and Chodorow (1998) (*lch*) counted in the number of nodes between both concepts and the total depth of the taxonomy (*D*) in a non-linear fashion by scaling them logarithmically. In this way, the semantic similarity between two concepts can be larger if they are located deeper in the taxonomy.

$$sim_{lch}(c_1, c_2) = 1 - \frac{\log{(spath)}}{log2D} \tag{2}$$

Wu and Palmer (1994) (*wup*) incorporated the idea of least common subsumer (LCS), the most specific taxonomical ancestor that two concepts share, into their proposed metric. When the relative depth of both concepts in the taxonomy is taken into account, two concepts that are located lower in the taxonomy will have higher semantic proximity.

$$sim_{wup}(c_1, c_2) = \frac{2 \times depth(LCS(c_1,c_2)}{spath(c_1,c_2) + 2 \times depth(LCS(c_1,c_2))} \tag{3}$$

Nguyen and Al-Mubaid (2006) (*nam*) proposed a measure with common specificity and local granularity features by factoring in both the depth of taxonomy and LCS of the two concepts. The shortest path measure was modified to scale the depth of LCS with respect to the total depth of the taxonomy (*D*).

$$sim_{nam}(c_1, c_2) = \log\left(2 + (spath(c_1, c_2) - 1) \times (D - depth(LCS(c_1, c_2))\right) \qquad (4)$$

Shobhana and Radhakrishnan (2016) proposed a new measure of similarity that recognizes multiple inheritances in ontologies. It combined super concepts with their common specificity features. Fuzzy rules were created to calculate similarity scores and then tailored to the optimal rule base by a modified genetic algorithm. The F1 measure of the proposed method was higher than the existing similarity measures.

Instead of measuring the LCS, McInnes et al. (2014) have introduced the *u-path* principle derived from the dense multi-hierarchical taxonomy of SNOMED CT. The *u-path* measure quantified the strength of similarity between two concepts based on their proximity within a hierarchy using undirected path information. The overall result showed a higher correlation than other path-based measurements on two standard benchmark datasets.

#### *2.3.2.2 Information Content (IC)-based measures*

The IC of a concept is estimated by the amount of semantic content of the concept. In the semantic hierarchy of an ontology, the IC of a concept is correlated with the probability of a concept occurring on the ontology hierarchy. The IC of a concept decreases as the layer of the concept goes deeper, hence the IC of the concept increases. This enables the semantic similarity between two concepts to be measured by assessing their IC. IC may be estimated from the distribution statistics of concepts within a text corpus (corpus IC) or from the semantic structure of ontology (intrinsic IC). In this section, we will discuss the intrinsic IC that typically determines the semantic similarity between two concepts based on the IC of their LCS.

Resnik (1995) (*res*) proposed that the amount of information two terms share can be measured by the IC of the LCS of both terms. Two terms whose LCS is located lower in the hierarchy are more similar than those that share an LCS located higher in the hierarchy.

$$sim_{res}(c_1, c_2) = IC(LCS(c_1, c_2)) \tag{5}$$

Lin (1998) (*lin*) proposed an alternative approach to address Resnik's problem that any pair of concepts under the same LCS will have the same similarity value. To consider the unique features of each concept, Lin proposed to use a ratio of the common information content of two terms (i.e., the IC of LCS) to the content of information separately associated with each of them (i.e., the aggregate of the IC of each concept).

$$sim_{lin}(c_1, c_2) = \frac{2\times IC(LCS(c_1,c_2))}{IC(c_1)+IC(c_2)} \tag{6}$$

Jiang and Conrath (1997) (*j&c*) combined the edge-based approach with the node-based approach, expressing the edge counting scheme in terms of IC. Considering the depths of and density around nodes, it measures the difference between the IC of two concepts and that of LCS (i.e., $IC(LCS(c_1,c_2))$).

$$sim_{j\&c} = \frac{1}{IC(c_1)+IC(c_2)-2\times IC(LCS(c_1,c_2))} \tag{7}$$

### 2.3.3 Measuring Semantic Similarity Between Two Sets of Concepts

Using the set theory, one can evaluate the similarity between sets according to their overlapping and differential elements (Hubálek 1982). Researchers employed the set theory to study the similarity paradigm for measuring the representations of sets (i.e., features of terms). In a set-based approach, set operations are applied to the predicted semantic similarities between sets, and terms are represented as a collection of features. Features that are distinct from each other set and common features of both sets are considered to define similarity coefficients. Classical methods such as Dice, Jaccard, and Cosine distance measures can be used to measure set-level similarity.

A variety of methods for calculating a set-level similarity has been proposed, measuring the resemblance of two taxonomic concept sets. Girardi et al. (2016) referred to the path distance between concepts in a hierarchy to detect similarities in patient records using categorical values such as ICD-10. According to Girardi et al. (2016), The distance between two sets X and Y is

$$d_H(X,Y) = \frac{1}{|X \cup Y|}(\sum_{x \in X \setminus Y} \frac{1}{|Y|} \sum_{y \in Y} d(x,y) + \sum_{y \in Y \setminus X} \frac{1}{|X|} \sum_{x \in X} d(y,x) \tag{8}$$

where $d(x,y)$ is denoted as

$$d(x,y) = \frac{p_{min}(x,y)}{l(x)+l(y)} \tag{9}$$

where $p_{min}(x,y)$ is the minimum number of edges between concepts $x$ and $y$, and $l(x)$ is the depth of a concept in the hierarchy. The distance metric between individual concepts takes into account the minimum path between concepts as well as their depths (level) in the hierarchy. At a set level, both the denominator and numerator of the distance $d_H(X,Y)$ measure how dissimilar the two sets are: the fewer concepts the two sets share, the greater the numerator gets while the denominator decreases. The proposed metric was evaluated on 800 patient records and showed marked improvement in clustering patient records, in terms of the average distance between patients with the same diseases, over the Jaccard distance, which does not take account of concept depth in the hierarchy, and Haase-Li concept distance (Haase et al. 2004).

To the best of my knowledge, the work by Alonso and Contreras (2016) is the only study that tried to measure similarity between query and patient records using the UMLS as a mapping layer. They leveraged both path- and intrinsic IC-based similarity metrics between concepts that were extracted from free-text queries and patient records using MetaMap, a biomedical NLP tool. They first calculated the Concept Unique Identifier (CUI)-level similarity by obtaining the maximum similarity between CUIs in query subphrase and patient records. For each CUI in the query subphrase, the maximum subphrase-level similarity in terms of the CUI was defined by the maximum CUI-level similarity values in the query subphrases. The average of those maximum subphrase-level similarity values was obtained for every query phrase to yield phrase-level similarity, and the final similarity between the query and patient records is the sum of those phrase-level similarity values divided by the number of phrases in the patient record. The test dataset from the 2011 Text Retrieval Conference (TREC) was used for evaluation; the result showed the performance of the path-based similarity (F1 score = 0.430) was similar to that of intrinsic IC-based one (F1 score = 0.427).

### 2.3.4 Similarity Measure within Multiple Ontologies

As the granularity and degree of coverage of an ontology vary by discipline, as discussed in section 2.2.1, attempts were made to use multiple ontologies for measuring semantic similarity between two concepts. Al-Mubaid and Nguyen (2009) proposed a new ontology-structure-based approach to map the concepts across multiple ontologies in the UMLS, which is based on the cross-modified path length and common specificity between two concepts as well as local granularity in each ontology cluster. Going further from Al-Mubaid and Nguyen's approach, Batet et al. (2013) considered the difference in size between ontologies and provided relative values that were normalized based on the ontological structure. Those multi-ontology-based approaches commonly made use of WordNet, an ontology for general terms, to improve the correlation with human similarity ratings.

### 2.3.5 Similarity Metrics in Practice

When semantic similarity exists between biomedical terms, the thesauri of synonymous terms can be used for information retrieval and natural language processing. The taxonomic structure of SNOMED CT has been exploited to classify or cluster entities like clinical models (e.g., templates or archetypes) (Gøeg et al. 2015), radiology reports (Mabotuwana et al. 2013), research articles for systematic review (Ji et al. 2017), and clinical trials (Wei & Fu 2017).

Other than classifying clinical documents, McInnes and Pedersen (2015) developed a framework for quantifying "relatedness," which represents how closely two concepts are associated although they are not semantically linked (i.e., *headache* and *aspirin* are related though they are not closely connected by *is-a* relationship). Significant differences were reported in the correlation of semantic similarity and relatedness with human judgment. They also found that the correlation could be substantially enhanced when a definition-based relatedness was coupled with an IC similarity measure.

Chandar et al. (2015) proposed a similarity-based approach to recommending candidate n-grams to new SNOMED CT concepts. They clustered n-gram SNOMED CT concept sets based on their distance between the feature vectors using the K-means algorithm and aligned them with the SNOMED CT semantic structures. The ranked list of n-grams made it simpler and easier to define new concepts for SNOMED CT.

Martínez et al. (2013) and Sánchez et al. (2014) proposed a general framework to mask textually sensitive health data by exploiting semantic similarity based on the taxonomic structure of SNOMED CT. Harispe et al. (2014) proposed a method for unifying semantic similarity measures based on ontology, decomposing the underlying core elements of various semantic similarity measures. Hsieh et al. (2013) used web-indexed pages as training corpus to retrieve page counts of a given term for estimating semantic similarity.

### 2.3.6 Conclusion

Semantic similarity between two concepts can be measured by probabilistic, path-based, or information-theoretic approaches. Path-based approaches account for information about the co-location of the terms in taxonomy and measure the distance between the links relating to them. Information-theoretic approaches measure the semantic similarity based on the information the two concepts share. To conform to the human sense of similarity, relatedness measures quantify shared information contents or use context vectors of terms, not depending on subsumption relations of an ontology.

## 2.4 Review of Cohort Identification Systems

### 2.4.1 Cohort Identification Systems Submitted to 2018 National NLP Clinical Challenges

The Track 1 Shared Task of the 2018 National NLP Clinical Challenges (n2c2) was to identify patients who meet eligibility criteria from narrative medical records (Stubbs et al. 2019).

The task required participating systems to provide decisions on each patient's eligibility for 13 clinical trials. Each eligibility criterion was derived from real clinical trials listed on ClinicalTrials.gov and had multiple subcriteria, which required intra-criterion rules or logic to organize and integrate information gathered from subcriteria into the final conclusion. Given the complexity of the logic in the criteria, various NLP subtasks, such as concept extraction, assertion detection, time-related information extraction, laboratory results extraction, and meta-information extraction, were needed to be coordinated. The definition and applicable NLP components of 13 criteria are provided in Table 1.

The dataset consisted of records for 288 deidentified patients. The training set consisted of 202 records while the remaining 86 were set aside as the test set during the challenge. Only the training dataset with gold-standard labels was released for the participants, and the test dataset was not disclosed at the time of the challenge. The organizers evaluated the submitted systems based on the test dataset that was held out. The gold-standard labels of the test dataset have been opened to the public at the time of the development of the current framework. Only the textual content of patient records was available. Each of the 288 patient records aggregated longitudinal records of 2-5 visits in chronological order, and patient-level eligibility for each of the 13 criteria was presented as "met" or "not met" at the end of the patient record. Each visit record contains patient demographics, chief complaint, present illness, medical history, family history, social history (e.g., smoking history and alcohol consumption), signs and symptoms, physical examination findings, laboratory data, radiology and/or pathology reports, prescribe medications, problem list, or referral letter.

**Table 1.** Definition and basic information of the 13 eligibility criteria as used in 2018 n2c2 Shared Task 1. The columns of "NLP components" were adopted from Chen et al. (2019b)

| Criterion Name | Criteria | NLP components | | | | | Number of records | |
|---|---|---|---|---|---|---|---|---|
| | | CE | AD | TM | LR | MI | Met | Not Met |
| DRUG-ABUSE | Drug abuse, current or past | Y | Y | | | | 15 | 273 |
| ALCOHOL-ABUSE | Current alcohol use over weekly recommended limits | Y | Y | Y | | Y | 10 | 278 |
| ENGLISH | Patient must speak English | Y | Y | | | | 265 | 23 |
| MAKES-DECISIONS | Patient must make their own medical decisions | Y | Y | | | | 277 | 11 |
| ABDOMINAL | History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction | Y | Y | | | | 107 | 181 |
| MAJOR-DIABETES | Major diabetes-related complication, defined as any of the following that are a result of (or strongly correlated with) uncontrolled diabetes:<br>• Amputation<br>• Kidney damage<br>• Skin conditions<br>• Retinopathy<br>• Nephropathy<br>• Neuropathy | Y | Y | | | | 156 | 132 |
| ADVANCED-CAD | Advanced cardiovascular disease, defined as having 2 or more of the following:<br>• Taking 2 or more medications to treat cardiovascular disease (CAD)<br>• History of myocardial infarction<br>• Currently experiencing angina<br>• Ischemia, past or present | Y | Y | Y | | Y | 170 | 118 |
| MI-6MOS | Myocardial infarction in the past 6 months | Y | Y | Y | | Y | 26 | 262 |
| KETO-1YR | Diagnosis of ketoacidosis in the past year | Y | Y | Y | | Y | 1 | 287 |
| DIETSUPP-2MOS | Taken a dietary supplement (excluding Vitamin D) in the past 2 months | Y | Y | Y | | Y | 149 | 139 |
| ASP-FOR-MI | Use of aspirin to prevent myocardial infarction | Y | Y | | | | 230 | 58 |
| HBA1C | Any HbA1c value between 6.5 and 9.5% | Y | Y | | Y | | 102 | 186 |
| CREATININE | Serum creatinine > upper limit of normal | Y | Y | | Y | Y | 106 | 182 |

*CE*, concept extraction; *AD*, assertion detection; *TM*, time-related information extraction; *LR*, laboratory results extraction; *MI*, meta-information extraction.

To create the gold standard labels, seven annotators, all of whom had medical training (five registered nurses, one medical doctor, and one medical assistant), independently annotated all records. Each record was assigned to three different annotators. For each criterion, the annotators examined patient records for evidence that the criteria were met or not met, and they annotated the relevant phrases accordingly. The parts of the text that provided evidence for the annotators' assertions were not disclosed to the public. The gold standard labels were determined by majority vote. The overall agreement between them before adjudication was 84.9%. The macro-averaged precision, recall, and F1 score for annotators compared to the gold standard were 0.957, 0.959, and 0.958, respectively (Stubbs & Uzuner 2015).

The goal of 2018 n2c2 Shared Task Track 1 was to identify patients who meet each of the 13 eligibility criteria from the narrative medical records, as well as to scrutinize whether NLP systems can make use of clinical narratives to identify patients eligible for clinical trials. The participants of the 2018 n2c2 exploited various NLP methods from empirical rules (Chen et al. 2019; Spasic et al. 2019) to deep learning techniques (Chen et al. 2019; Segura-Bedmar and Raez 2019) to hybrid methods combining the advantages of two or more approaches (Tannier et al. 2019; Vydiswaran et al. 2019) (Table 2). The mean micro F1 score for all submissions was 0.799 and the maximum score was 0.91. Among the top 10 systems, there were 4 rule-based systems and 6 hybrid or machine learning systems (Stubbs et al. 2019).

**Table 2**. Micro average F1 scores achieved by selected cohort identification systems on the 2018 n2c2 dataset.

| Affiliations (Authors) | Methods | Micro F1 |
|---|---|---|
| Medical University of Graz (Oleynik et al. 2019) | Rule-based, deep learning | 0.9100 |
| University of Michigan (Vydiswaran et al. 2019) | Hybrid | 0.9075 |
| Sorbonne Université (Tannier et al. 2019) | Hybrid | 0.9069 |
| Meta Data Quest (Chen et al. 2019b) | Rule-based | 0.9028 |
| Cardiff University (Spasic et al. 2019) | Rule-based | 0.8814 |
| Universidad Carlos III de Madrid (Segura-Bedmar & Raez 2019) | Deep learning | 0.7721 |

A team from the Medical University of Graz (Oleynik et al. 2019) evaluated shallow and deep learning classifiers. To address the possible overfitting of complex models trained on small datasets, they used BioWordVec-pretrained embeddings (which were trained on PubMed and MIMIC-III) to reuse unsupervised input representation schemes trained on a large dataset, followed by a fine-tuning of schemes using a small annotated dataset. The rule-based approach used regular expressions and textual markers to detect negation and context and to extract laboratory findings. Support vector machines were trained on a bag-of-words representation of the input document using *tf-idf*. The logistic regressions model was trained with 100 epochs and a window size of 5. The Long Short-Term Memory (LSTM), which is based on a recurrent neural network that is typically used for time-series events, was optimized with a learning rate of 0.02 and 25 epochs of training. Overall, rule-based (F1 score 0.91) and shallow models such as support vector machine (F1 score 0.80) showed higher micro F1 scores than deep learning strategies (F1 score 0.74). The pretrained models outperformed the self-trained models throughout classifiers, improving the F1 score of logistic regression from 0.8068 to 0.8113 and that of LSTM from 0.7504 to 0.7522. The overall micro F1 score for all thirteen criteria (0.910) was at the top of the 2018 n2c2 participants.

Another team, from Med Data Quest (Chen et al. 2019b), designed a challenge-oriented, rule-based NLP system, whose performance was compared with a hybrid general NLP system the authors previously built for general medical information extraction. Their system was plugged in with lexical-, syntactic, and meta-level evidence to indicate the presence of the target concepts, to validate the relations between the core concepts and their modification attributes, and to use section-level, note-level, document-level, and patient-level information such as patient's gender and date of the note. The hybrid general NLP system was used for text processing (tokenization, sentence division, and section detection), grammar analysis (POS correction, concept correction, and normalization), entity and relation extraction (NER, relation detection, modifiers detection, and disambiguation), and knowledge reasoning (relation reasoning, concept reasoning, and concept arrangement). Built upon the UMLS and Unstructured Information Management Architecture (UIMA), the hybrid system combined bidirectional LSTM models and rules to assign the relationship between entities, such as treatment relations between diseases and drugs. The F1 score of the challenge-oriented, rule-based system was 0.9028 and that of the hybrid general system was 0.8145.

Spasic et al. (2019) investigated the above-mentioned 2018 n2c2 tasks taking advantage of rule-based pattern matching, mostly regular expressions, to extract context-sensitive features from longitudinal free-text patient records. Information not directly relevant to a patient or eligibility criterion—negated entities, family history, allergies, and irrelevant time window—was filtered out. The bag-of-words representations were then passed to supervised classifiers. To prevent the loss of the context of individual words, context tags were attached to an individual token that is lexically distinguishable from other tokens. Supervised machine learning was performed when a sufficient number of training data was available in each criterion. A rule-based approach focusing on a small set of relevant features was chosen for the remaining criteria. A total of four machine learning algorithms—support vector machine, logistic regression, naïve Bayesian classifier, and gradient tree boosting—was performed on unseen test data. The system achieved an overall F1 score of 88.14% and outperformed three baseline systems, which were the rule-based, the hybrid, and the hierarchical neural network.

Segura-Bedmar and Raez (2019) exploited several deep learning architectures such as a simple convolutional neural network (CNN), deep CNN, recurrent neural network (RNN), and CNN-RNN hybrid architecture for 2018 n2c2 Shared Task on cohort selection for clinical trials. Unlike other existing deep learning systems, the authors used a fully connected feedforward (FFF) layer before the classification layer for all four deep learning models. The feature vector was fed into the FFF layer, which mapped it into a higher-order feature space that is easier to separate into distinct labels. Random initialization generated statistically significant better results than pre-trained word embeddings for various criteria when deep learning models were expanded with the FFF layer. With micro-F1 around 77% and macro F1 around 49%, the RNN and hybrid designs delivered the best overall results. The use of the FFF layer contributed to the improved results for all architectures except the hybrid architecture. Because of the small amount of the dataset, however, the performance was inferior to that of typical machine learning classifiers proposed by other studies.

### 2.4.2 Cohort Identification Systems Outside 2018 National NLP Clinical Challenges

Outside the 2018 n2c2, Liu et al. (2013) reported an information extraction framework for extracting named entities and their corresponding contextual information (e.g., negation, temporality, and experiencer) under cTAKES. The framework was developed under UIMA that has been adopted to create and implement a pipeline of modular tools for processing unstructured content. The knowledge-driven information extraction engine is comprised of regular expressions, normalization, and match rules. This is a viable approach when a task involves a specific subdomain or a limited number of named entities. However, they evaluated the performance of the implementation on only two phenotypes (peripheral arterial disease and heart failure) and did not prove its transferability to other domains.

### 2.4.3 Conclusion

The teams participating in the 2018 n2c2 shared task achieved average F1 scores of around 0.9 using various classification methods including rule-based approaches and deep learning. Most of the best-performing cohort identification systems submitted to the 2018 n2c2 used rule-based and hybrid approaches.

# III. METHODOLOGY

This study focused on the cohort identification from unstructured free-text narratives of EHRs given a free-text query. Retrieval of information that could readily be extracted from structured parts of EHRs will not be explored in this study. Rather, this study paid particular attention to processing data that could be obtained from free-text clinical notes, such as symptoms and signs, drugs administered, diseases diagnosed, modes of disease/injury, and radiology findings.

## 3.1 Research Question

In this study, SNOMED CT concepts extracted from eligibility criteria descriptions (query) and patient records were used to answer the following research question.

Research Question: Can SNOMED CT's semantic relations be exclusively used for querying disease names, procedures, medications, and other patient contexts from free-text clinical narratives for cohort identification?

To answer the research question, the current study examined whether a cohort identification framework that uses SNOMED CT's semantic hierarchy that represents relations between biomedical entities can achieve performance on a par with or better than currently optimal systems, most of which use supervised learning classifiers (Oleynik et al. 2019; Tannier et al. 2019; Mikolov et al. 2013), without referring to external resources such as large-scale biomedical corpora.

The proposed design of the framework is based on the following premises:

1. SNOMED CT's comprehensive coverage of biomedical entities makes it applicable to the broad spectrum of biomedical fields and reusable for diverse eligibility descriptions.
2. Terms in an eligibility criterion are semantically more general than those in patient records and should be used to retrieve those in patient records that contain their descendants. If a query term is "intra-abdominal surgery," for example, patient records that include *cholecystectomy* or *small bowel resection* should be retrieved since those are the specifics of *intra-abdominal surgery*.
3. The more a biomedical entity is mentioned in a patient record, the more closely the patient record is related to that biomedical entity.

Premise 1 ensures that the proposed framework applies to a wide range of diseases and specialties. Unlike previous cohort identification systems dedicated to detecting specific diseases such as colorectal cancer (Xu et al. 2011), cerebral aneurysm (Castro et al. 2017), and pediatric cancers (Ni et al. 2015), the current framework can serve diverse eligibility criteria owing to SNOMED CT's comprehensive coverage for biomedical entities.

According to Premise 2, concepts extracted from an eligibility criterion (hereafter referred to as "query concepts") semantically subsume some, if not all, concepts in the eligible patient records (hereafter referred to as "patient record concepts"). When similarity is measured, the proposed algorithm examines whether any descendants of the query concepts exist in patient records, and those that are not descendants of a query concept are ignored.

Premise 3 assumes that the more frequently a concept is mentioned in a patient record, the more likely the patient is related to that concept. This is a plausible premise, considering that it is a common practice for clinicians to repeatedly record the same problems in the patient record for every visit until the problem is completely resolved. For example, the word *diabetes* or *hyperglycemia* will appear in the problem lists of every visit record of a diabetic patient, unless he is not fully returned to normoglycemic status. Since each patient record in the 2018 n2c2 dataset notes multiple historical visits by the patient, it is highly likely that longstanding clinical problems will have been mentioned multiple times in the patient record.

## 3.2 Framework Design

### 3.2.1 Overview

In this study, concepts and concept relations encoded in SNOMED CT within the UMLS were leveraged to prompt concept features of eligibility criteria and each patient record. Figure 2 outlines the high-level steps of the proposed framework. In the first step, free-text clinical notes were preprocessed to eliminate redundant information that is not relevant to the patient, e.g., diseases or procedures that were negated or referred to family members. Many parts of the preprocessing were adopted from the work by Spasic et al. (2019) who preprocessed narrative medical records using regular expressions and a rule-based text mining approach. The second step, entity extraction, used MetaMap, an off-the-shelf biomedical NLP tool that automatically encodes clinical free texts into SNOMED CT concepts (Aronson 2001). In the next step, contextual features such as negation (e.g., “patient *denied* a history of sexually transmitted disease”) and temporality (e.g., “*currently* experiencing angina”) were identified by examining the syntactic structure of a sentence. Biomedical entities that were associated with negation or context discordant with the eligibility criteria were removed if necessary. This yielded a *Patient Record Concept Feature Set* for each patient record, which contained biomedical entities relevant only to the patient. Finally, the semantic similarity between the Patient Record Concept Feature Sets and *Query Concept Feature Set*, which represents eligibility criteria with SNOMED CT concepts, was measured to estimate how closely the patient record was related to the query. The more similar a patient record is to the query, the more likely the patient is eligible for the corresponding clinical trial.

A Python application programming interface (API) of MetaMap was used for entity extraction. With those extracted entities, semantic relationships among them were identified using ontology relational tables stored in a local machine and were exploited to yield similarity between query and patient records. The biomedical entity extraction tool, ontology relational tables, and other NLP subtasks were integrated on Python 3.8.5. The complete Python codes for the current framework are available at https://github.com/eunsuk-c/CohortIdentificationSNOMED/.

### 3.2.2 Data Set and Sample

A publicly available dataset was used in this study. Free-text clinical notes from the 2018 National NLP Clinical Challenge (n2c2) Shared Task and Workshop on Cohort Selection for Clinical Trials (Stubbs et al. 2019) were used for the construction of the framework and evaluation of performance. The original corpus was from the 2014 i2b2/UTHealth shared tasks.

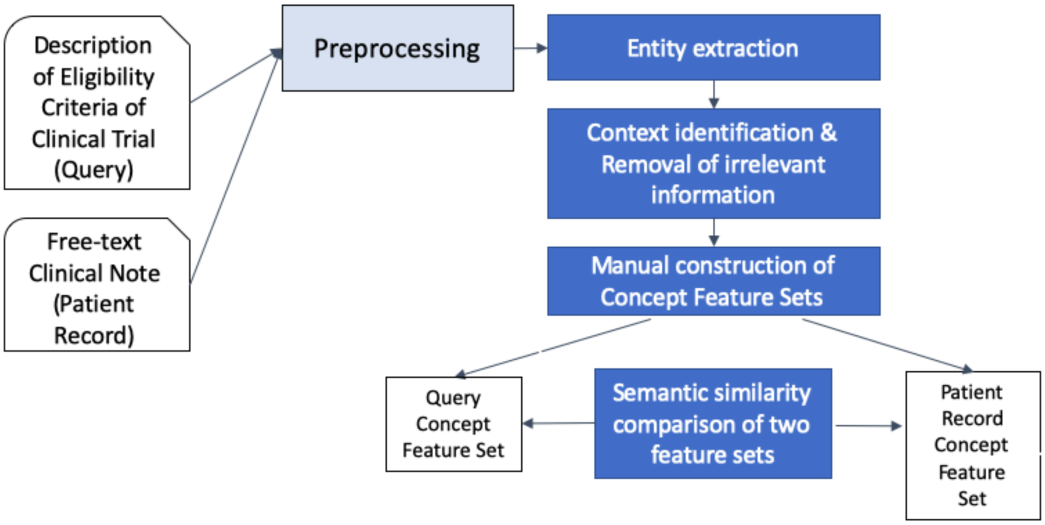

Description of Eligibility Criteria of Clinical Trial (Query)
Free-text Clinical Note (Patient Record)
Preprocessing
Entity extraction
Context identification & Removal of irrelevant information
Manual construction of Concept Feature Sets
Query Concept Feature Set
Semantic similarity comparison of two feature sets
Patient Record Concept Feature Set

**Figure 2.** Architecture of the proposed NLP framework.

Each visit note in the de-identified patient record starts with "Record date: *YYYY-MM-DD*," a changed, imaginary date. For each set of patient records, it is assumed that the most recent record is *now*, regardless of the year, day, or month recorded. While some discharge notes follow a predefined format with subheadings such as "Present Illness," "Allergy," and "Social History," most of them are written at a clinician's discretion without document sections.

Since the purpose of the current study is to exploit SNOMED CT's semantic structure for knowledge acquisition from free-text clinical notes, only three highly medical and knowledge-intensive criteria—"ABDOMINAL," "MAJOR-DIABETES," and "ADVANCED-CAD"—were used in this study. Those three criteria were ideal testbeds where cohort identification could be attained by querying symptoms, diseases, medications, and procedures, which are the most commonly used EHR data elements. In these criteria, the presence of specialized medical concepts in clinical narratives is crucial to determining patients' eligibility. Not only do those three criteria have a relatively balanced distribution of "met" and "not met" labels, but they also allow me to test broad coverage of NLP components of a cohort identification task: concept extraction, assertion detection, time-related information extraction, and meta-information extraction. To answer the research question, only those subcomponents of NLP that are required to query disease names, procedures, medications, and other patient contexts from free-text clinical narratives were tested in this study. Other NLP components of cohort identification tasks such as the extraction of laboratory test values (which is for "CREATININE" and "HBA1C" criteria) or inference from indirect evidence (which is for "ENGLISH" and "MAKES-DECISION" criteria) were not explored in this study.

### 3.2.3 Preprocessing

Most NLP systems generally require several distinct steps to preprocess unstructured, free text into processible data. Key steps of preprocessing in this cohort identification framework include removing special punctuations, detecting sentence boundaries, resolving abbreviations, parsing documents, and removing redundant information.

Most of the preprocessing was done using regular expressions. First, all characters were converted into lowercase characters. Periods within acronyms were removed (e.g., *c.c.* becomes *cc*) so that they could be used only to mark the end of sentences and split sentences. Contractions were expanded (e.g., *couldn't* becomes *could not*) so that the negative "not" could be exposed for a later stage of the removal of negated expressions (Table 3a).

**Table 3.** Examples of preprocessed sentences and phrases.

| Original text | Preprocessed text |
|---|---|
| a. Expanding contractions | |
| The patient couldn't tolerate the prescribed medications. | `the patient could not tolerate the prescribed medications.` |
| b. Converting brand names to generic names | |
| Cortril 5mg q.d. | `hydrocortisone 5 mg qd` |
| c. Section titles | |
| Allergy: the patient has no known allergy. | `allergy:`<br>`the patient has no known allergy.` |
| Family history<br>Mother was diagnosed with breast cancer at age 57. | `family history:`<br>`family member was diagnosed with breast cancer at age 57.` |
| d. Replacement with family members | |
| His mother was diagnosed with breast ca. but hasn't undergone treatment. | `his family member was diagnosed with breast cancer but has not undergone treatment.` |

A separate file of the dictionary of medical abbreviations (based on (Spasic et al. 2019), modified by author) was used for expanding abbreviations. The expansion of abbreviations was done not only for approved ones (e.g., *A. fib* for *atrial fibrillation*) but for informal ones that were frequently used in the medical community (e.g., *c.c* for *chief complaint*), using regular expressions shown in Table 4. The trade names of drugs were converted into their generic names (e.g., *Lipitor* becomes *atorvastatin*) using Drugs@FDA Data Files (U.S. Food & Drug Administration 2017) (Table 3b).

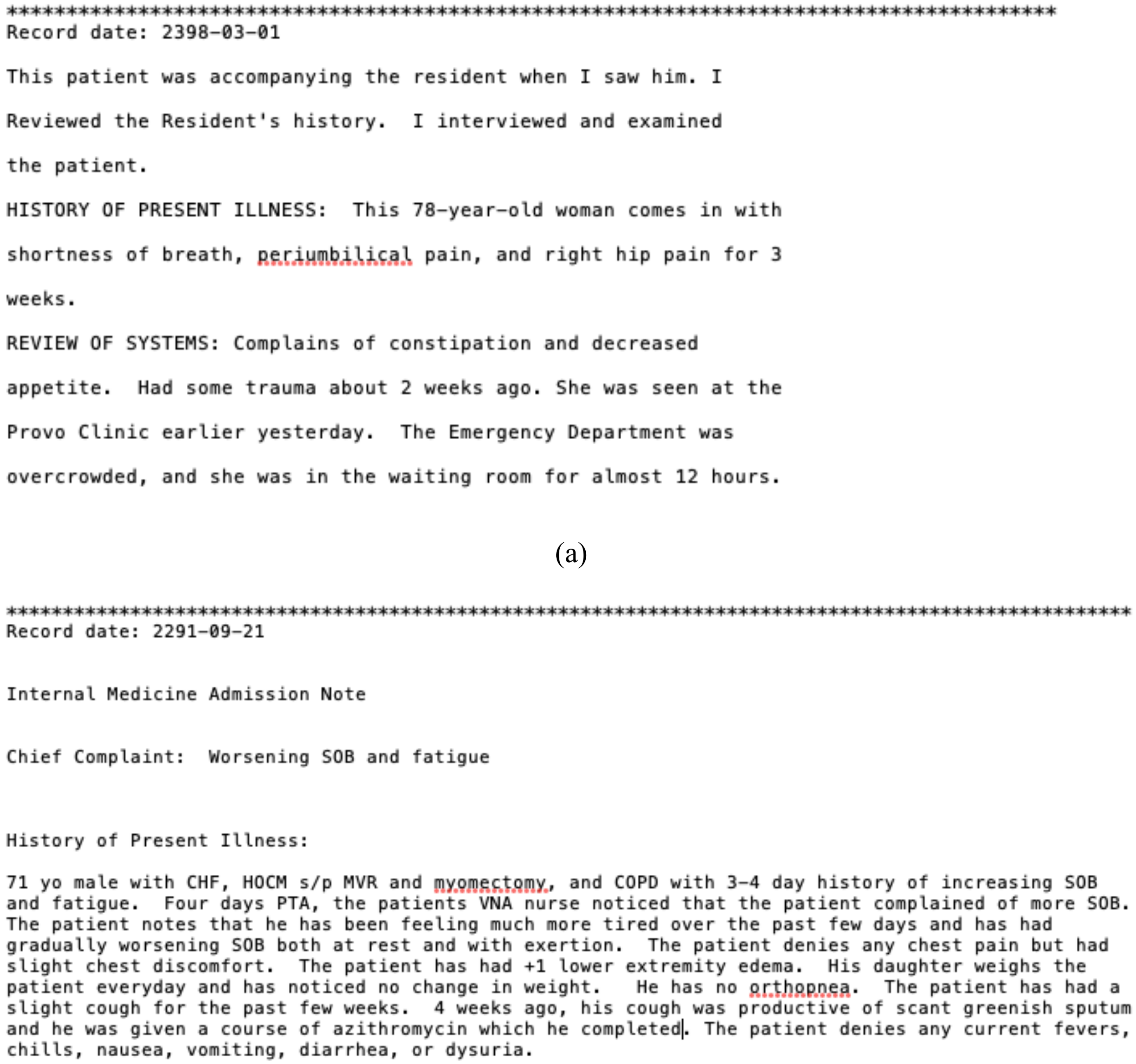
**Table 4**. Regular expressions to expand abbreviations and convert them into fully-expanded terms.

| Regular expressions | Example abbreviations | Converted to |
|---|---|---|
| \ss\/p\s | s/p | status post |
| \sc\s*\/*\s*c\s | cc | chief complaint |
| \sc\s*\/\s*o\s | c/o | complaining of |
| \sc\s*\/\s*b\s | c/ b | complicated by |
| \sc\s*\/\s*w\s | c/w | continue with |
| \sf\s*\/\s*h\s | f/h | family history |
| \sn\s*\/\s*v\s | n/v | nausea vomiting |
| \so\s*\/\s*e\s | o/e | on examination |
| \se\s*\/\s*o\s | e / o | evidence of |
| \snc\s*\/\s*at\s | nc/at | normocephalic atraumatic |
| \sd\s*\/\s*c\s | d/c | discontinued |
| \sd\s*\/\s*o\s | d/o | disorder |
| \sf\s*\/\s*up*\s | f/u | follow up |
| \sh\s*\/\s*o\s | h/o | history of |
| \sy\s*\/*\s*o\s | yo | year old |
| \sr\s*\/\s*o\s | r/o | rule out |
| \sw\s*\/\s*u\s | w/u | work up |
| \sw\s*\/\s*o[ut]*\s | w/out | without |
| \sw\s*\/*\s | w/ | with |

One challenge that remains is that the lines in the patient record documents in the dataset are arbitrarily separated. While lines in some notes are broken by paragraphs, those in other notes are done at the physician's discretion. Some line breaks take place after an arbitrary number of characters, even though the sentence is not finished (Figure 3(a)). The line breakers in broken sentences needed to be eliminated; instead, the whole sentence (or paragraph) should be pasted end to end in order to make a syntactic analysis of the sentence possible. In such cases, the line breakers that resided between lines shorter than 50 characters were removed. The hypothesis behind this is that any lines longer than 50 characters are highly likely to be syntactically connected with the next line. In this way, short lines that simply listed lab results (i.e., “`Na 135 \n K 5.3 \n`”) or problems (i.e., `Problem List: \n history of CAD \n current DM with medication \n`) can avoid being pasted with the lines below, which prevents the merging of unrelated information.

In the 2018 n2c2 dataset, the sections of patient records (“family history,” “lab results,” “radiology reports,” etc.) were arbitrarily divided by providers; one may place a colon at the end of the section headings and start a new line, while another may prefer writing a section heading and its contents on one line without line breaks in between. To separate the unstructured record by section, header terms that were embedded at the first position on a line were identified and forced to be separated from that line if they contained word parts like *history*, *exam*, *lab*, *med*, *allerg,* or *plan* (Table 3c). Section headings on a separate line were then tagged by a colon at the end so that they can be easily identified at a later stage for section division. This enabled the stratification of information by sections (see 3.2.4).

Documents were parsed and tokenized using spaCy, an open-source Python toolkit for advanced natural language processing. The document was then segmented into sentences by the occurrence of a period, and sentences were a unit of text processing. After document parsing, any entities that referred to family members (e.g., mother, daughter, etc.) were converted into the word phase *family member* (Table 3d). Those sentences that included the word phase *family member* were deleted at a later stage.



```
*****************************************************************************************
Record date: 2398-03-01

This patient was accompanying the resident when I saw him. I

Reviewed the Resident's history.  I interviewed and examined

the patient.

HISTORY OF PRESENT ILLNESS:  This 78-year-old woman comes in with

shortness of breath, periumbilical pain, and right hip pain for 3

weeks.

REVIEW OF SYSTEMS: Complains of constipation and decreased

appetite.  Had some trauma about 2 weeks ago. She was seen at the

Provo Clinic earlier yesterday.  The Emergency Department was

overcrowded, and she was in the waiting room for almost 12 hours.
```

(a)

```
***************************************************************************************************
Record date: 2291-09-21


Internal Medicine Admission Note


Chief Complaint:  Worsening SOB and fatigue



History of Present Illness:

71 yo male with CHF, HOCM s/p MVR and myomectomy, and COPD with 3-4 day history of increasing SOB
and fatigue.  Four days PTA, the patients VNA nurse noticed that the patient complained of more SOB.
The patient notes that he has been feeling much more tired over the past few days and has had
gradually worsening SOB both at rest and with exertion.  The patient denies any chest pain but had
slight chest discomfort.  The patient has had +1 lower extremity edema.  His daughter weighs the
patient everyday and has noticed no change in weight.   He has no orthopnea.  The patient has had a
slight cough for the past few weeks.  4 weeks ago, his cough was productive of scant greenish sputum
and he was given a course of azithromycin which he completed. The patient denies any current fevers,
chills, nausea, vomiting, diarrhea, or dysuria.
```

(b)

**Figure 3**. In patient record (a), line breaks occur approximately every 55-60 characters even though sentences are not finished. Each line is also separated by an empty line. In patient record (b), on the other hand, a line break exists only at the end of paragraphs, and there are no empty spaces between lines. The patient records shown here are imaginary examples reproduced by the author, adhering to the styles of the original 2018 n2c2 shared task track 1 dataset.

### 3.2.4 Note-level and Section-level Information Identification

Subcriteria such as “currently experiencing angina” contain time constraint that requires the processing of meta-information of patient records. Each patient record of the 2018 n2c2 dataset contains 2-5 physician encounters that were separated by the date heading “Record date: *YYYY-MM-DD*,” sorted by visit dates. For visit-level information identification, the date of visit was identified and separated using regular expressions from the date headings. The most recent visit note was regarded as “current.”

Section identification is useful to filter out some false or noisy information. For instance, the mention of “calcium” in the laboratory exam section is highly likely to indicate it was calcium content in the serum; that mentioned in the medication section is likely to belong to dietary supplements. If a clinical trial seeks a patient who takes calcium supplements, it will filter out the calcium-related concepts mentioned in the laboratory exam section of the notes. This will help to reduce preprocessing mistakes and mitigate the potential concept mapping errors.

Rules and regular expressions were used to identify section headings tagged at section 3.2.3 and to normalize section header terms. Identified heading texts were assigned to one of the following normalized section headers: “chief complaint,” “history of present illness,” “past medical history (includes past surgical history),” “family history,” “social history,” “review of systems,” “physical examination,” “allergies,” “medications,” “laboratory examinations,” “radiological examinations,” “problem list,” “assessment,” and “plans.”

Each patient record was stored as values of Python’s “dictionary of dictionary”, with visit dates as keys and section headers as sub-keys. In this way, users can adapt sections needed for a specific cohort identification task. The *allergies* section, for example, may not be used in most cohort identification tasks, while the biomedical entities contained in the *problem list* section provide the most decisive clues to the patient's eligibility in most cases. In addition to the sections of interest, one can define the time frame of interest she wants to retrieve; if patients who have currently experienced angina are retrieved, the fraemwork can search for the records of the most recent date.

### 3.2.5 Extraction of SNOMED CT Concepts

The texts preprocessed at the previous step were passed to the next step: extraction of SNOMED CT concepts from free text using MetaMap. MetaMap is a publicly available program maintained by the National Library of Medicine, providing access to multiple biomedical ontologies/vocabularies brought by the UMLS and mapping biomedical text to the UMLS Metathesaurus (Aronson 2001).

The UMLS organizes concepts from various source terminologies in the Metathesaurus following a unique identifier structure in the concept-, term-, string-, and atom-level hierarchies. CUI represents a biomedical concept that encompasses all its synonym terms. Lexical Unique Identifier (LUI) groups the descriptions of a concept by lexical variations. String Unique Identifier (SUI) identifies the uniqueness of any variation in character set, upper-lower case, or punctuation in the human-readable description of a concept. Atom Unique Identifier (AUI) is assigned to each occurrence of a string in a given source terminology (Table 5).

**Table 5.** Representation structure of UMLS concept C0019163 from the 2018AB version (National Library of Medicine 2021).

| CUI | LUI | SUI | AUI | Source | String |
|---|---|---|---|---|---|
| C0019163 | L0042725 | S0515550 | A2984272 | SNOMED | Type B viral hepatitis |
| C0019163 | L2871361 | S14189714 | A23469643 | SNOMED | Viral hepatitis type B (disorder) |
| C0019163 | L0019163 | S0047904 | A2882355 | SNOMED | Hepatitis B |
| C0019163 | L0019163 | S0047904 | A0067497 | MeSH | Hepatitis B |
| C0019163 | L14168949 | S17230772 | A28394375 | MeSH | Hepatitis B Virus Infection |

MetaMap was originally designed for retrieving MEDLINE/PubMed citations from the biomedical scientific literature. As ongoing research efforts have shown that MetaMap was also effective at retrieving UMLS concepts from free-text clinical narratives (Kimia et al. 2015; Przybyła et al. 2019) as well as other tasks such as knowledge discovery (Weeber et al. 2000), it has now become one of the most frequently cited medical concept retrieval tools to map various sources of clinical narratives into the UMLS Metathesaurus.

MetaMap processes input text through multiple lexical/syntactic analyses as the following (Aronson & Lang 2010):

- Tokenization, sentence boundary determination, and acronym/abbreviation identification;
- Part-of-speech tagging;
- Lexical lookup of input words; and
- Final syntactic analysis consisting of a shallow parse in which phrases and their lexical heads.

MetaMap divides an input text into noun phrases and then generates variants for each noun phrase, with a variant essentially consisting of one or more noun phrase terms, as well as all spelling variants, abbreviations, acronyms, and synonyms. Using an evaluation function, it maps a collection of candidate CUIs containing one of the variants and computes a score for each candidate CUI. Then it brings together candidates that are participating in disjoint parts and recalculates the score using the merged candidates. The CUIs with the highest scores are chosen as the best match for the input text. A recent analysis of the performance of MetaMap showed its recall, precision, and F1 score were 0.88, 0.89, and 0.88, respectively, for extracting UMLS concepts from the Informatics for Integrating Biology to the Bedside (i2b2) Obesity Challenge data (Reátegui & Ratté 2018).

To use MetaMap on the Python interface which is most widely used for natural language processing, pymetamap, a Python wrapper developed by Rios (2020) was employed. Pymetamap imports a list of sentences and extracts concepts using MetaMap before it returns concepts in the form of a list of Concept objects. Pymetamap returns the following outputs from the sentence "John had a huge heart attack."

```
Concept(index='2', mm='MM', score='13.22', preferred_name='Myocardial
Infarction', cui='C0027051', semtypes='[dsyn]', trigger='["Heart
attack"-tx-1-"heart attack"]', location='TX', pos_info='17:12',
tree_codes='C14.280.647.500;C14.907.585.500')
```

Metamap extracts CUI (`Concept.cui`) "C0027051," whose preferred name (`Concept.preferred_name`) is "Myocardial Infarction," from the trigger term (`Concept.trigger`) "Heart attack." It also provides information about semantic types (`Concept.semtypes`) and the probability score (`Concept.score`) of the extracted concept given the sentence it belonged to, which is in this example 13.22. For this study, the `cui`, `preferred_name`, and `trigger` instances of Concept objects were used.

In this study, MetaMap was configured to match SNOMED CT terms only. Every occurrence of a SNOMED CT concept was assigned a CUI of the UMLS and then converted to the corresponding AUIs since the UMLS defines semantic relations between concepts by AUIs only. If a trigger term had an assigned CUI but no SNOMED CT AUI, its SNOMED CT fully specified name (FSN) was searched on the SNOMED CT Browser[1] using Snowstorm API, and the AUI of the corresponding FSN was retrieved. If the term was not searchable on the SNOMED CT Browser, the CUI was abandoned and never used. Those UMLS concepts that did not encompass at least one SNOMED CT AUI were withdrawn.

The use of MetaMap has two lingering problems: (i) MetaMap accepts some generic words and verbs (e.g., best, normal, take, reduce) as medical entities when segmenting and extracting medical entities, and (ii) MetaMap may suggest multiple concepts for the same term, as well as several semantic types for the same concept when categorizing medical items. For example, the concept of *removal* in the noun phrase "removal of bone chips" will be suggested as either **A2979561 | Removal (procedure) |** or **A3241327 | Removal – action (qualifier value) |**. To solve (i), concepts extracted at and uni-, bi-, and tri-gram levels were also considered in addition to noun-phrase-level extraction. For example, from the sentence "The patient underwent an MRA of the head," MetaMap by default extracts the concept **A9409632| MRI angiography of head (procedure) |**. The current fraemwork was configured to extract additional concepts from unigram words as well, namely **A15145221 | Magnetic resonance imaging (MRI) of vessels (procedure) |** from the single word "MRA" and **A3140404 | Head structure (body**

[1] browser.ihtsdotools.org

**structure) |** from the single word “head.”

In addition, the semantic types of extracted concepts were classified to allocate their importance. Those concepts which belonged to “therapeutic procedure,” “clinical history/examination observable,” “drug or medicament,” or “evaluation finding,” to name a few, were given higher importance whereas those under “administrative statuses” or “relative times” were assigned lower importance. Important concepts that carry more specific meanings in clinical settings, such as symptoms, signs, anatomical sites, diseases, procedures, and medications, were grouped and assigned “quantified importance,” with 1 being the most important. The quantified importance of a semantic type was assigned manually by the author such that the current framework can best identify the most commonly used data elements such as diagnoses, procedures, and medications. Even though a concept was lexically meaningful, it received less importance if it was subsumed by no biomedically-important semantic group. In the current framework, concepts whose quantified importance was less than 0.5 were eliminated. This prevented MetaMap from mapping clinically trivial concepts, section headings (e.g., *chief complaint*, *medications*), and mistakenly extracted concepts (e.g., mapping **A2873197 | Proton |** from a letter “h”) to SNOMED CT concepts. A list of upper-level concepts whose subtypes are eliminated from mapping is shown in Appendix A. The examples of the description of concepts, their corresponding semantic group, and quantified importance are shown below:

**Concept** – quantified importance << supertype (SEMANTIC GROUP)

**Therapeutic radiology procedure** – 0.8 << Therapeutic procedure (PROC)

**Penicillin** – 0.8 << Chemical (MED)

**Initially** – 0.0 << Event orders (GEN)

**Eye examination interpretation** – 0.2 << Interpretation of findings (DESC)

For problem (ii), all extracted concepts, irrespective of their semantic tags, were considered for document expansion at a later stage. For example, MetaMap extracts the concepts **A2878436 | Amputation (procedure) |, A3241080 | Amputation – action (qualifier value) |,** and **A2991260| Amputated structure (morphologic abnormality) |** from the text string "amputation due to diabetic foot." In this case, all three concepts are added to a patient document for consideration. This does not affect the similarity measure between the text and query, as shown in section 3.2.11, since the similarity metric considers the subtypes of a query concept only.

### 3.2.6 Concept Annotation and Document Expansion with Concept Features

In this step, all biomedical terms and words in the patient records were annotated by UMLS AUIs using MetaMap, and the extracted AUIs replaced the corresponding trigger terms in the sentence. In this way, the syntactic structure of the patient record can be preserved and the relations among biomedical entities extracted from the free text can be machine-processable. Cues for negation (i.e., without), certainty (i.e., highly likely), and temporality (i.e., past history of) were also preserved in sentences so that they could be exploited later to assess contextual information surrounding a biomedical entity.

Table 6 demonstrates an example of a sentence in which trigger term strings were substituted by UMLS AUIs. As mentioned in section 3.2.5, multiple AUIs extracted by MetaMap can be used in place of a trigger term, irrespective of its sort of semantic tags, for document expansion.

**Table 6.** Example of original text and text processed to replace biomedical entities with UMLS AUIs. Note that two AUIs (A29922622 and A2878587) were mapped from a single trigger term, "anxiety." **A3501627 | Hypertensive disorder, systemic arterial (disorder) |**, **A2928669 | Diabetes mellitus (disorder) |**, **A2992622 | Anxiety disorder (disorder) |**, **A2878587 | Anxiety (finding) |**, **A2890018 | Panic attack (finding) |**, **A2984272 | Type B viral hepatitis (disorder) |**.

| Original text | Text replaced by AUIs |
|---|---|
| She has a history of hypertension and diabetes, anxiety and panic attacks, hepatitis B as well. | `she has a history of A3501627 and A2928669, A2992622 A2878587, and A2890018, A2984272 as well.` |

### 3.2.7 Eliminating Irrelevant Concept Features

At this stage, biomedical entities (i.e., AUIs) that were irrelevant to the current situation of a patient were detected using regular expressions. AUIs that were located in the same sentence with negation cues, which are listed below, were identified by empirical rules.

No, without, rule out, deny, cannot see, unlikely, negative for, is negative, hold, neither…nor, not appear, not known, not appreciate, not complain, not demonstrate, not exhibit, not feel, not reviewed, not have, does not, free of

AUIs that were not related to the rationale for current treatment were also detected if they were associated with the *prevention* or *prophylaxis* of diseases. Planned procedures were detected if the corresponding AUIs followed or were followed by the word *plan* or *schedule*.

Those irrelevant AUIs are stored in a separate "negation set" so that the AUIs in this vector can be removed at a later stage (Table 7).

**Table 7**. Example of negation set. The word *denied* triggered the negation of the sentence; the AUIs in the negated sentence were included in the negation set.

| Original text | Text replaced with AUIs | Negation set |
|---|---|---|
| The patient denied past history of sexually transmitted disease. | `The patient denied past history of A3070494.` | `{A307494}` |

### 3.2.8 Building Patient Concept Feature Set

In this step, AUIs were extracted from the AUI-replaced text prepared above and were stored in the Patient Concept Feature Set. Then the AUIs that were also included in the negation set were discarded from the Patient Concept Feature Set. In the end, the Patient Concept Feature Set contained AUIs that were relevant to the current situation of the patient. The final product of the Patient Concept Feature Set is a Python dictionary whose keys are patient-relevant AUIs and values are the number of occurrences of the corresponding AUI in the patient record. One example of a Patient Concept Feature Set is shown below:

Patient Concept Feature Set of Patient 217 = `{‘A23027694’:7, ‘A2887859’:2, ‘A10885208’:3…}`[2]

### 3.2.9 Building and Expanding Query Concept Feature Set

A Query Concept Feature Set was manually curated to represent each eligibility criterion with AUIs. Biomedical entities that are relevant to each criterion were examined across

[2] **A23027694 | Type 2 diabetes mellitus |, A2887859 | Urticaria |, A10885208 | Degenerative disorder of macula |**

training dataset, and those relevant entities in terms of SNOMED CT concepts were added to the Query Concept Feature Set if they aligned with medical expert knowledge. Sets of concepts extracted by MetaMap from each free-text eligibility criterion were refined to best represent the eligibility criterion using AUIs, which consisted of a Query Concept Feature Set. The Query Concept Feature Sets for the ADVANCED-CAD and MAJOR-DIABETES eligibility criteria were constructed as shown in Table 8 and Table 9, respectively. In the later step of measuring similarity between query and patient records, only those patient record concepts that are the direct descendants of query concepts will contribute to similarity measurement.

**Table 8.** Free-text criteria and their corresponding sets of AUIs for the ADVANCED-CAD eligibility criterion.

| Criteria in natural language | Corresponding AUIs | Note |
|---|---|---|
| Taking 2 or more medications to treat CAD | **A2872933 \| Nitrate salt \|**<br>**A2884643 \| Nitroglycerin \|**<br>**A3651057 \| Antiplatelet agent \|**<br>**A2882344 \| Heparin \|**<br>**A3483678 \| HMG-CoA reductase inhibitor \|**<br>**A3609517 \| Angiotensin-converting enzyme inhibitor agent \|**<br>**A3249419 \| Angiotensin II receptor antagonist \|**<br>**A3802645 \| beta-blocking agent \|**<br>**A3335705 \| Calcium channel blocker \|** | Satisfied only if ≥2 medications |
| History of myocardial infarction | **A7873496 \| Coronary arteriosclerosis \|**<br>**A2938836 \| Myocardial infarction \|**<br>**A24246836 \| Coronary artery stent \|**<br>**A26575074 \| EKG: myocardial infarction \|** | |
| Currently experiencing angina | **A23468083 \| Angina pectoris \|**<br>**A3313085 \| Ischemic chest pain \|** | Only the most recent visit record was examined |
| Ischemia, past or present | **A7873407 \| Ischemic heart disease \|** | **A7873407 \| Ischemic heart disease \|** is a supertype of **A23468083 \| Angina pectoris \|** |

**Table 9.** Free-text criteria and their corresponding sets of AUIs for the MAJOR-DIABETES eligibility criterion.

| Criteria in natural language | Corresponding AUIs | Note |
|---|---|---|
| Major diabetes-related complication, defined as any of the following that are a result of (or strongly correlated with) uncontrolled diabetes: | **A2928669\| Diabetes mellitus \|**<br>**A2882662 \| Hyperglycemia \|**<br>**A3032080 \| Hyperglycemic disorder \|**<br>**A3322678 \| Diabetic complication \|** | At least 1 of these AUIs must exist in the Patient Concept Feature Set |
| Amputation | **A3241080 \| Amputation - action \|**<br>**A2878436 \| Amputation \|**<br>**A2991260 \| Amputated structure \|** | |
| Kidney damage (nephropathy) | **A24807926 \| Chronic kidney disease \|**<br>**A23451959 \| Chronic renal failure \|**<br>**A2890220 \| End stage renal disease \|**<br>**A3318512 \| Diabetic renal disease \|**<br>**A3063986 \| Renal impairment \|** | |
| Skin conditions | **A3471880 \| Gangrenous disorder \|**<br>**A2881852 \| Gangrene \|**<br>**A3769261 \| Ulcer of lower extremity \|** | |
| Retinopathy | **A3064437 \| Retinal disorder \|** | |
| Neuropathy | **A2876139 \| Neuropathy \|**<br>**A3282650 \| Paresthesia of foot \|**<br>**A3119481 \| Diabetic hand syndrome\|**<br>**A3400640 \| Disorder of the peripheral nervous system \|**<br>**A3162750 \| Neuropathic pain \|** | |

### *3.2.9.1 Task-specific configuration of Query Concept Feature Set: ABDOMINAL*

In the case of the ABDOMINAL criterion, the definition of *intra-abdominal surgery* is broad and ambiguous, as it does not explicitly dictate whether it includes, for example, intra-pelvic surgeries such as prostatectomy. To figure out the annotation intention of the 2018 n2c2 dataset, I examined the AUIs extracted from the training dataset and compared those extracted from the *met* patient records with those from the *not met*. From this, it was found that procedures such as prostatectomy, dilation and curettage, and inguinal hernia repair were not considered intra-abdominal surgery by the annotators of the 2018 n2c2. In compliance with these findings, the final set of query concept features was constructed as the pseudocode below:

```
ABDOMINAL:IntraAbdominalSurgery
equivalentTo sct:OperationOnAbdominalRegion
and not (sct:OperativeProcedureOnMaleGenitourinaryTract
       or sct:EndometrialScraping
       or sct:AbdominalWallProcedure)
```

The definition of *intraabdominal surgery* for the ABDOMINAL study is equivalent to *operation on abdominal region* in SNOMED CT but not *operative procedure on male genitourinary tract* in SNOMED CT, nor *endometrial scaping* in SNOMED CT, nor *abdominal wall procedure* in SNOMED CT. In this way, '`sct:Prostatectomy`' was excluded from the query because it was a subtype concept of '`sct:OperativeProcedureOnMale-GenitourinaryTract`,' even though it was an operation on the abdominal region. This definition of *intra-abdominal surgery* used for the ABDOMINAL criterion can be graphically presented in Figure 4.

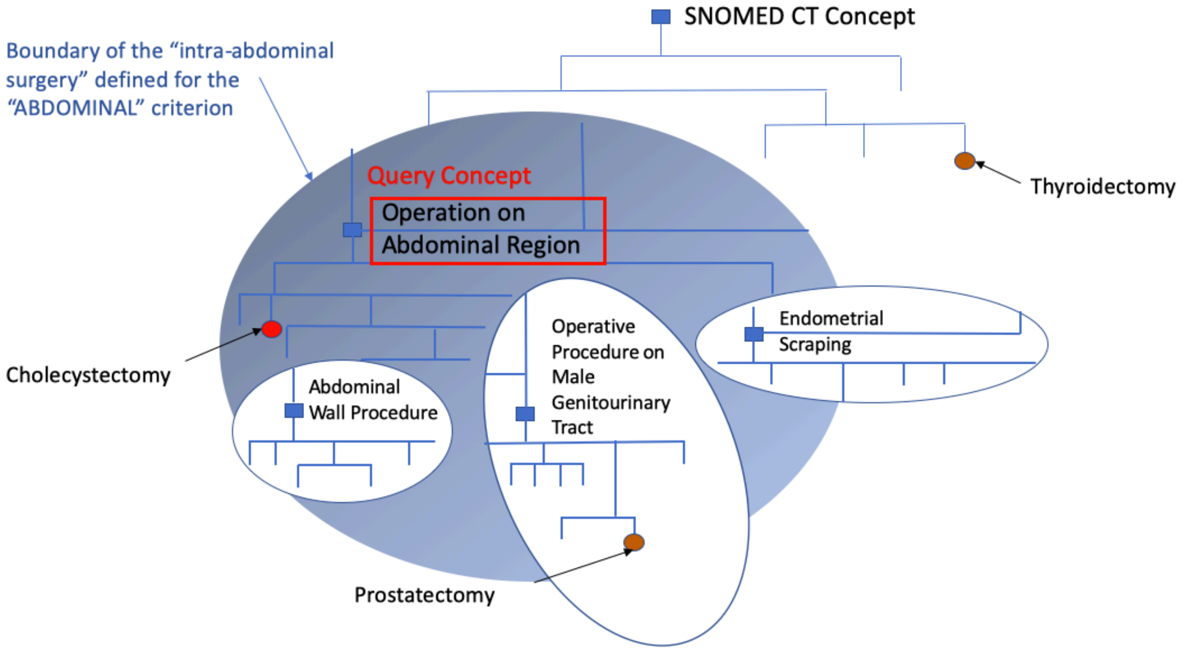

SNOMED CT Concept
Boundary of the "intra-abdominal surgery" defined for the "ABDOMINAL" criterion
Query Concept
Operation on Abdominal Region
Thyroidectomy
Cholecystectomy
Abdominal Wall Procedure
Operative Procedure on Male Genitourinary Tract
Endometrial Scraping
Prostatectomy

**Figure 4.** Graphical description of the definition of "intra-abdominal surgery" in the ABDOMINAL criterion. Only blue-shaded concepts are considered to define the boundary of intra-abdominal surgery for the ABDOMINAL criterion. The subtypes of "Abdominal Wall Procedure," "Operative Procedure on Male Genitourinary Tract," and "Endometrial Scraping" are excluded from the definition of intra-abdominal surgery for the ABDOMINAL study. Accordingly, patient records that contain the concept of *cholecystectomy* are eligible for the ABDOMINAL although those that contain the concepts of *prostatectomy*, which is a subtype of "Operative Procedure on Male Genitourinary Tract," and *thyroidectomy*, which is outside the "Operation on Abdominal Region" hierarchy, are not.

### *3.2.9.2 Task-specific configuration of expanded Query Concept Feature Set: ADVANCED-CAD*

Further processing of the Patient and expanded Query Concept Feature Sets needed to take into account the context of eligibility criteria. To satisfy the modifying adverb *currently* in the eligibility criterion "*currently* experiencing angina," direct descendants of **A23468083 | Angina pectoris |** and **A3313085 | Ischemic chest pain |** were extracted from the most recent visit note. This was enabled by referring to the "dictionary of dictionary" prepared at the stage described in section 3.2.4, where each patient's visit date was recorded as a key of the dictionary and the corresponding patient record contents as a value.

#### *3.2.9.3 Task-specific configuration of expanded Query Concept Feature Set: MAJOR-DIABETES*

The MAJOR-DIABETES criterion implies that a patient has to have *both* diabetes and diabetic complications to be eligible. For the purpose of illustration, assume that the framework identifies a subtype of "nephropathy," such as *acute nephropathy* from a patient record. If the patient record does not mention any concept of "diabetes," however, the patient is not eligible, because we cannot infer that acute nephropathy was caused by diabetes in that patient. To achieve this, the proposed framework proceeded to check whether the patient had diabetic complications after confirming that the patient had diabetes. The patient was then determined to be eligible if he had any diabetic complications.

Further inspection of training data revealed that the entities on the left column of Table 10 did not receive credits to be included in the Query Concept Feature Set for MAJOR-DIABETES. Therefore, the AUIs presented on the left column of Table 10 and their subtypes were not included in the expanded Query Concept Feature Set, even though they were the subtypes of the concepts presented in the right column of Table 10. Graphically, the definition of *nephropathy* is represented by blue area in Figure 5.

**Table 10.** AUIs excluded from the Expanded Query Concept Feature Set for MAJOR-DIABETES criterion (left column) and their supertype concepts (right column). All subtypes of each AUIs on the left column were also automatically excluded from the expanded Query Concept Feature Set.

| Excluded **AUI \| Description \|** | Subsumed by |
|---|---|
| **A3092358 \| Acute renal impairment \|** | **<< A3063986 \| Renal impairment \|** |
| **A18073061 \| Cardiorenal syndrome \|** | |
| **A3665747 \| Renal failure associated with renal vascular disease \|** | |
| **A2953455 \| Acute renal failure syndrome \|** | |
| **A2972962 \| Milk alkali syndrome \|** | |
| **A26587941 \| Renal impairment caused by Polyomavirus \|** | |
| **A24246786 \| Prerenal renal failure \|** | |
| **A15154924 \| Color blindness \|** | **<< A3064437 \| Retinal disorder \|** |
| **A3367488 \| Congenital anomaly of retina \|** | |
| **A3307688 \| Age related macular degeneration \|** | |
| **A2892313 \| Exudative retinopathy \|** | |
| **A8387117 \| Hamartoma of retina \|** | |
| **A2891853 \| Hemangioma of retina \|** | |
| **A3030153 \| Hereditary vitreoretinopathy \|** | |
| **A2890283 \| Hypertensive retinopathy \|** | |
| **A3034481 \| Injury of retina \|** | |
| **A3578402 \| Neoplasm of retina \|** | |
| **A2939246 \| Night blindness \|** | |
| **A3904017 \| Paraneoplastic retinopathy \|** | |
| **A3238102 \| Retinal abnormality - non-diabetes \|** | |
| **A2892298 \| Retinal defect \|** | |
| **A7873926 \| Retinopathy of prematurity \|** | |
| **A3070798 \| Sickle cell retinopathy \|** | |
| **A11734429 \| Toxic retinopathy \|** | |
| **A3318307 \| Axonal neuropathy \|** | **<< A2876139 \| Neuropathy \|** |
| **A3322744 \| Cranial nerve disorder \|** | |
| **A3121005 \| Disorder of nerve repair \|** | |
| **A3034272 \| Inflammatory neuropathy \|** | |
| **A3318879 \| Neuromyopathy \|** | |
| **A3162763 \| Neuropathy due to infection \|** | |
| **A2976161 \| Paraneoplastic neuropathy \|** | |
| **A2874833 \| Postinfectious neuralgia \|** | |

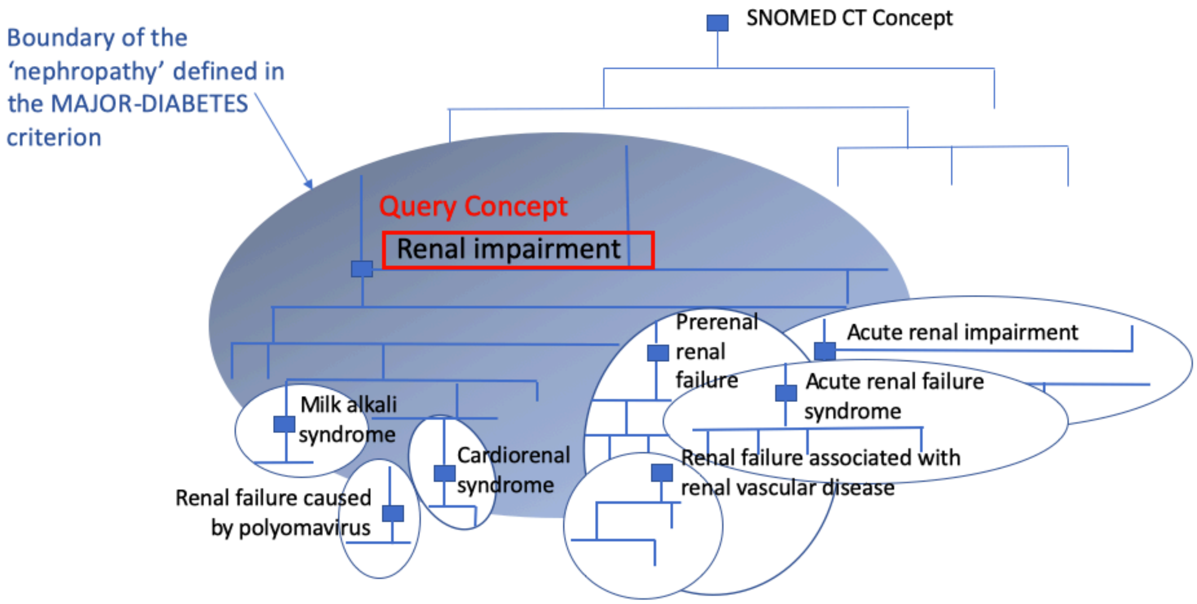
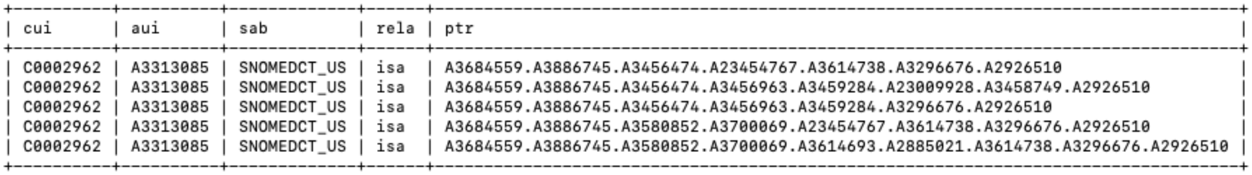

SNOMED CT Concept
Boundary of the 'nephropathy' defined in the MAJOR-DIABETES criterion
Query Concept
Renal impairment
Prerenal renal failure
Acute renal impairment
Acute renal failure syndrome
Milk alkali syndrome
Cardiorenal syndrome
Renal failure associated with renal vascular disease
Renal failure caused by polyomavirus

**Figure 5.** Graphical description of the definition of *nephropathy* in the MAJOR-DIABETES criterion.

### 3.2.10 Defining *is_a* Relations Among Concepts

The UMLS integrates over 2 million names for over a million concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts, to enable interoperability between computer systems.

To exploit semantic relations among UMLS concepts, a full version of 2016AB release files was imported using MetamorphoSys, a UMLS installation wizard, and Metathesaurus customization tool to a local machine. MySQL 8.0.22 was used to load MRCONSO and MRHIER relational tables up in the MySQL database. The MRCONSO table contains information about a concept's CUI, AUI, LUI, SUI, and descriptions (i.e., fully specified name, synonym, and preferred term). The MRHIER table defines a concept's *is_a* hierarchical relation back to the root to define all of its supertype concepts (Figure 6). A concept may have more than one *is_a* route to the root. The supertypes of patient record concepts can be identified using the MRHIER table and will be employed in the next step of similarity measurement.



```
+----------+----------+-------------+------+----------------------------------------------------------------------------------+
| cui      | aui      | sab         | rela | ptr                                                                              |
+----------+----------+-------------+------+----------------------------------------------------------------------------------+
| C0002962 | A3313085 | SNOMEDCT_US | isa  | A3684559.A3886745.A3456474.A23454767.A3614738.A3296676.A2926510                  |
| C0002962 | A3313085 | SNOMEDCT_US | isa  | A3684559.A3886745.A3456474.A3456963.A3459284.A23009928.A3458749.A2926510         |
| C0002962 | A3313085 | SNOMEDCT_US | isa  | A3684559.A3886745.A3456474.A3456963.A3459284.A3296676.A2926510                   |
| C0002962 | A3313085 | SNOMEDCT_US | isa  | A3684559.A3886745.A3580852.A3700069.A23454767.A3614738.A3296676.A2926510         |
| C0002962 | A3313085 | SNOMEDCT_US | isa  | A3684559.A3886745.A3580852.A3700069.A3614693.A2885021.A3614738.A3296676.A2926510 |
+----------+----------+-------------+------+----------------------------------------------------------------------------------+
```

**Figure 6**. A screenshot of the MRHIER table searched for *is_a* route to the root from the concept **A2926510 | Chest pain |** using a MySQL query "`select cui, aui, sab, rela, ptr from MRHIER where aui = A2926510`." **A2926510 | Chest pain |** has five routes to the root concept **A3684559 | SNOMED CT Concept |**. The ancestors of the concept **A2926510 | Chest pain |** include **A3684559 | SNOMED CT Concept |**, **A3886745 | Clinical finding |, A3456474 | Finding by site |**, **A3580852 | Neurological finding |**, **A23454767 | Finding of sensation by site |**, **A3456963 | Finding of body region |**, **A3700069 | Sensory nervous system finding |**, **A3614738 | Pain finding at anatomical site |**, **A3459284 | Finding of trunk structure |**, **A3614693 | Pain / sensation finding |**, **A23009928 | Finding of upper trunk |**, **A2885021 | Pain |**, **A3458749 | Finding of region of thorax |**, and **A3296676 | Pain of truncal structure |**, in the order of increasing depth in the *is_a* hierarchy.

If a supertype of a patient record concept is in the Query Concept Feature Set, the semantic similarity between the query concept and the patient record concept is measured as described in section 3.2.11. If none of the supertypes of a patient concept is included in the Query Concept Feature Set, the patient concept is abandoned and semantic similarity is not calculated. In the Figure 6 example, if the concept **A3614693 | Pain / sensation finding |** is in the Query Concept Feature Set and the concept **A2926510 | Chest pain |** is in the Patient

Concept Feature Set, the semantic similarity between **A3614693 | Pain / sensation finding |** and **A2926510 | Chest pain |** can be calculated since **A3614693 | Pain / sensation finding |** is a supertype of **A2926510 | Chest pain |**.

### 3.2.11 Similarity Measurement

In this section, a new similarity metric to measure the similarity between the Query and Patient Record Concept Feature Sets is proposed. As shown in section 2.3, existing metrics had measured the similarity between individual concepts or sets. Although concept-level or set-level similarity measures work well for comparing or clustering sets that are equivalent in terms of length and information content, they are limited in measuring similarities between query and patient records because patient records hold a disproportionally more specific and greater amount of information than queries do. Moreover, cosine similarity and other set-level similarity metrics can overestimate the similarity scores in favor of longer but less pertinent patient records when MetaMap extracts a substantial amount of unintended and irrelevant concepts from free texts.

To address these problems, I propose a new similarity measure as Equations (10) and (11). The proposed similarity measure first makes use of a new weight metric for a concept that appeared in a patient record with respect to an individual query concept as follows:

$$wt_q(p) = \begin{cases} \log(\frac{depth(q)}{depth(p)} + 1) \cdot \frac{1}{\sqrt{|subtypes(p)|+1}} \quad if\ p\ is\ subsumed\ by\ or\ identical\ to\ q \\ 0\ otherwise \end{cases} \quad (10)$$

*wt*$_q$(*p*) is the weight of a SNOMED CT concept *p* in a patient record (i.e., Patient Concept Feature Set) with respect to a SNOMED CT concept *q* in a query (i.e., Query Concept Feature Set). *depth(x)* is the minimum number of nodes in the path from a concept *x*, to the root of the SNOMED CT taxonomy. The weight is larger than 0 only if the concept from the patient record is a subtype of or identical to the concept from the query. By considering the subtypes of a query concept only, it eliminates from consideration irrelevant concept features (i.e., concepts located outside the query concept's hierarchy). If the patient concept *p* is not subsumed by the query concept *q*, the similarity is 0. *depth*(*q*)/*depth*(*p*) estimates how much specific *p* is in relation to *q*, and is less than or equal to 1 in most cases because query terms are usually more general (i.e., located at a lower depth in the SNOMED CT hierarchy) than the terms in patient records. If there were multiple *is_a* paths from *q* to *p*, the shortest path was selected for similarity computation.

The deeper the patient concept *p* is located in the taxonomy, i.e., has a more specialized meaning, the greater the weight gets. The weight also depends dynamically on query concept *q*; the more specific the query term is, (i.e., located deeper in the hierarchy), the greater the weight gets. For example, given the concept "acute peptic ulcer with hemorrhage" in a patient record, the query "ulcer" yields the weight score of 0.209, and the more concrete query "peptic ulcer" yields 0.283.

|*subtypes*(*p*)| is the number of all subtypes of patient record concept *p*. It is assumed that a concept with many subtype concepts is less specific than that with a smaller number of subtype concepts because general concepts need to be further defined at a lower level of the hierarchy (Hadj Taieb et al. 2014). The number of subtype concepts was counted by identifying the direct descendant of the concept of interest and then iteratively counting its direct descendants (descendants of descendants) down to the *is_a* hierarchy until there are no more descendants. Even though the concepts "hospital department" and "choroidal hemorrhage" are located at the same level of the SNOMED CT hierarchy, for example, "choroidal hemorrhage" is

more specialized than “hospital department” because “hospital department” has 100 subtypes while “choroidal hemorrhage” has 2 (Figure 7). By Equation (10), general concepts, which hold more subtypes, are penalized and get smaller weights.

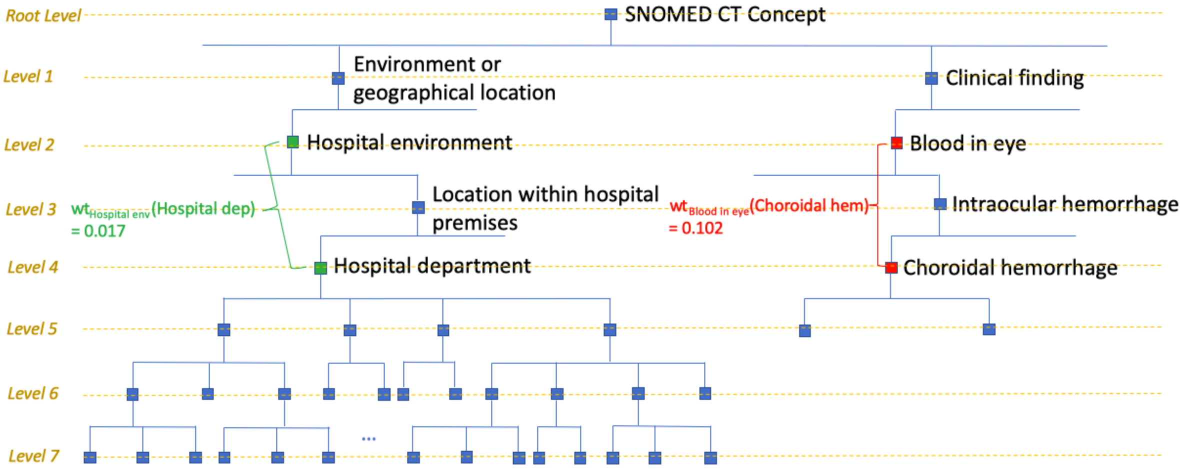


**Figure 7.** In the green case, the query concept is “hospital environment” and the patient record concept is “hospital department.” In the red case, the query concept is “blood in eye” and the patient record concept is “choroidal hemorrhage.” Intuitively speaking, the concept “choroidal hemorrhage” is concrete enough, while the concept “hospital department” is general and further definition—whether it is administrative departments, cardiology department, or intensive care unit—is needed in the lower levels of hierarchy. As shown in the figure, the concept “hospital department” has more subtypes than the concept “choroidal hemorrhage” does, and accordingly, it can be concluded that “choroidal hemorrhage” is semantically more specific than “hospital department” even though they are located at the same level in the SNOMED CT hierarchy. The calculated weights of the concepts “hospital department” and “choroidal hemorrhage” with respect to the query terms “hospital environment” and “blood in eye,” respectively, are 0.017 and 0.102, respectively, by Equation (10).

The final similarity score between a query and a patient record is the sum of the weight of every patient record concept with respect to each query concept. If a query contains 10 concepts and a patient record contains 1000 concepts, a total of 10,000 weight scores are calculated, and the sum of those scores is the final similarity of the patient record to the query.

$$sim(P,Q) = \sum_{i=1}^{m} \sum_{j=1}^{n} wt_{q_j}(p_i) \cdot \log\left(freq(p_i) + 1\right) \tag{11}$$

*sim(P,Q)* is the similarity score between Query Concept Feature Set *Q*, which contains *n* concepts, and Patient Record Concept Feature Set *P*, which contains *m* concepts, where usually $m >> n$. $freq(p_i)$ is the number of utterances of the *i*-th concept in the patient record. The larger the similarity score, the more semantically similar the query and the patient record are. Even though a single biomedical entity in a patient record is represented by multiple AUIs, only those which are the subtypes of the query concept contribute to the similarity score. Looking into the example of Table 6 in which two SNOMED CT concepts (**A2992622 | Anxiety disorder (disorder) |**, **A2878587 | Anxiety (finding) |**) were extracted from a single biomedical entity "anxiety" in a patient record, if the query concept is **A3322705 | Mental disorder (disorder) |**, only the concept **A2992622 | Anxiety disorder (disorder) |** will be considered for calculating similarity because it is directly subsumed by the query concept **A3322705 | Mental disorder (disorder) |** while the concept **A2878587 | Anxiety (finding) |** is not.

For the purpose of illustration, assume that we have the following Query Concept Feature Set:

*Q* = {**A2935334 | Ketoacidosis |**, **A7873407 | Ischemic heart disease |**}

and a Patient Record Concept Feature Set:

*P* = {**A2928669 | Diabetes mellitus |**, **A10865852 | Diabetic ketoacidosis |**, **A2938836 | Myocardial infarction |**}

For the calculation of similarity between the Query Concept Feature Set and the Patient Record Feature Set, we first compute weights of every patient record concept with respect to every query concept using Equation (10):

- $wt_{Ketoacidosis}(Diabetes\ mellitus) = 0$ (∵ “Diabetes mellitus” is not subsumed by “Ketoacidosis”)
- $wt_{Ketoacidosis}(Diabetic\ ketoacidosis) = \log\left(\frac{9}{10} + 1\right) \times \frac{1}{\sqrt{4}} = 0.321$
- $wt_{Ketoacidosis}(Myocardial\ infarction) = 0$ (∵ “Myocardial infarction” is not subsumed by “Ketoacidosis”)
- $wt_{Ischemic\ heart\ disease}(Diabetes\ mellitus) = 0$ (∵ “Diabetes mellitus” is not subsumed by “Ischemic heart disease”)
- $wt_{Ischemic\ heart\ disease}(Diabetic\ ketoacidosis) = 0$ (∵ “Diabetic ketoacidosis” is not subsumed by “Ischemic heart disease”)
- $wt_{Ischemic\ heart\ disease}(Myocardial\ infarction) = \log\left(\frac{6}{7} + 1\right) \times \frac{1}{\sqrt{88}} = 0.066$

According to Equation (11), the final similarity score between the Query Concept Feature Set and Patient Record Concept Feature Set is the sum of all participating weights.

$$\text{Sim}(P, Q) = 0 + 0.321 + 0 + 0 + 0 + 0.666 = 0.987$$

### 3.2.12 Determining Eligibility of Patient

If there was a defined number of items that should be met to satisfy eligibility, those items are also considered in the Query Concept Feature Set. In the ADVANCED-CAD example, “advanced cardiovascular disease” is defined as having two or more of the items listed in Table 8. To comply with that criterion, only those patients who have two or more eligibility subcriteria (out of four) were predicted to be eligible.

Since the gold-standard labels of the eligibility of each patient record in the test data of the 2018 n2c2 dataset were binary (e.g., “met” or “not met”) while the current framework yields numeric similarity scores, a cut-off similarity value needed to be established to determine the number of true positive (TP), true negative (TN), FP, and FN cases. Patient records were sorted by descending order of similarity score calculated in section 3.2.11. A cut-off similarity score, above which patients were deemed eligible, was established for each eligibility criterion (see section 3.2.13) and those patients whose similarity scores are greater than or equal to the cut-off similarity were predicted to be “eligible” by the current framework and the rest “not eligible.”

### 3.2.13 Evaluation

There are two types of classification errors in a cohort identification task: false negative (FN) (i.e., the current framework misses eligible patients) and false positive (FP) (i.e., the current framework includes patients who are not eligible for the study). The effort to reduce FN cases demands the post-hoc review of included patients, while the effort to minimize FP cases prevents recruiters from reaching out to potentially eligible patients. Therefore, FN and FS need to be balanced in a cohort identification task. The F1 score, a harmonic mean of precision and recall (Equation 12), is a desirable metric to assess errors in cases where FS and FP are equally undesirable.

$$F1\ Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{12}$$

F1 score is calculated from precision and recall which, in turn, are calculated on the predicted labels. The binary classification of patient eligibility predicted in the previous step resulted in the number of TP, true negatives TN, FP, and FN cases. Threshold cut-off similarity scores were adjusted to optimize the F1 score for each eligibility criterion. The performance of the current framework was evaluated in terms of the F1 score on the 2018 n2c2 test dataset, which is comprised of 86 patient records, and recall and precision obtained by the best possible F1 score were reported.

The 2018 n2c2 shared task has disclosed the performance of participating cohort identification systems in terms of recall, precision, and F1 score. Using the same data set and evaluation methods allows a legitimate evaluation of the current framework's performance by comparing it to those submitted by participants of the 2018 n2c2 shared task track 1.

# IV. RESULTS

## 4.1 Introduction

Chapter 3 described the methods used to prepare for predicting a patient's eligibility for each of the three criteria. In this chapter, the performance of the proposed framework is compared with systems submitted to the 2018 n2c2 shared task using the same test data and evaluation metrics.

## 4.2 Overall Performance

The experimental results, along with the best F1 score obtained by the participants of 2018 n2c2 in the three selection criteria, are displayed in Table 11. The macro average F1 score of the current framework was 0.933, exceeding that of the best of the 2018 n2c2. The performance of the proposed framework was higher than the best-performing systems of the n2c2 throughout the three selection criteria.

**Table 11**. The current framework's overall performance on test data set and comparison with n2c2 submissions.

| **Criteria** | **Current Framework** | | | | **Best n2c2 submission** | **Median n2c2 submission*** |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** (before eliminating patient records with spurious labeling) | **F1** (after eliminating patient records with spurious labeling) | **F1** | **F1** |
| ABDOMINAL | 0.964 | 0.900 | 0.931 | 0.931 | 0.912 | 0.889 |
| ADVANCED-CAD | 0.823 | 0.933 | 0.875 | 0.894 | 0.870 | 0.780 |
| MAJOR-DIABETES | 0.974 | 0.884 | 0.927 | 0.974 | 0.884 | 0.831 |
| Average | | | | 0.933 | 0.889 | 0.833 |

* Median among top 10 performers

### 4.3 Cohort Identification Performance: ABDOMINAL

The current framework's precision, recall, and F1 score for the ABDOMINAL criterion were 0.964, 0.900, and 0.931, respectively. The F1 score of the current framework for the ABDOMINAL criterion was s higher that of the best system submitted to the 2018 n2c2. The individual similarity score for each patient is presented in Appendix B.

The similarity score above which a patient record was predicted to be *met* was 0.001 for the ABDOMINAL criteria. The median number of relevant AUIs (e.g., AUIs that contributed to the similarity scores of each patient record) in the ground-truth *met* and *not met* patient records was 2 (mean = 1.900, range 0-7) and 0 (mean = 0.018, range 0-1), respectively.

There were one false-positive case and three false-negative cases among the 86 patient records in test data. In the false positive case, concepts for uterine myomectomy, **A3319398 | Uterine myomectomy |**, which is considered an intra-abdominal surgery by the current system, were appropriately extracted from the clinical note. However, the annotators of the 2018 n2c2 data labeled the case as “not met.” In the false-negative case, transurethral resection of the prostate (TURP) in patient record # 218 was not recognized by the system as an intra-abdominal surgery because the system was configured to exclude *Operative procedure on male genitourinary tract*. Cesarean section was once mentioned in patient record # 236, as in “midline scar in the lower abdomen is evident from her C section,” but the system translated it as “midline scar in the lower abdomen is evident from her C **A2895486 | Transection |**” and failed to recognize it as a cesarean section.

### 4.4 Cohort Identification Performance: ADVANCED-CAD

The system’s precision, recall, and F1 score for the ADVANCED-CAD criterion were 0.823, 0.933, and 0.875, respectively. The current system had a higher F1 score than the best system submitted to 2018 n2c2. After removing spurious patient records, the precision, recall, and F1 score for ADVANCED-CAD were 0.955, 0.840, and 0.894, respectively. The similarity scores for each patient record are presented in Appendix B.

The similarity score above which a patient record was predicted to be *met* was 1.0 for the ABDOMINAL criteria. The median number of relevant AUIs (e.g., AUIs that contributed to the similarity scores of each patient record) in the ground-truth *met* and *not met* patient records was 22 (mean = 23.111, range 7-61) and 11 (mean = 11.390, range 2-30), respectively. More than two subcriteria should be satisfied in order to be eligible for the ADVANCED-CAD study, but each *not met* patient record contained a minimum of two relevant AUIs which satisfied only one subcriterion. Examples of those cases include patients who are taking calcium channel blockers for the treatment of hypertension, but not for the treatment of coronary artery disease.

There were 9 false-positive cases and 3 false-negative cases wrongly predicted by the current system before removing spurious data points. The error analysis of the 9 false-positive cases showed:

- Unspecified or ambiguous entities:
  - The system considered simple utterances of *coronary artery disease* to be *myocardial infarction*, although the two entities are clinically different (patient records # 115, 156, 205)
- Flawed assertion detection
  - The system failed in assertion detection: it was not able to determine if diseases were present, conditional, suspected, or hypothetical. As a result, the system incorrectly identified suspicious cases as definite present problems (patient record # 396).
- Failed to recognize context around concepts:
  - The system ignored the context word *nonobstructive*, which makes the coronary artery disease it modifies less severe, from the word phrase *nonobstructive coronary artery disease*. This error caused it to falsely predict that the patient had advanced coronary artery disease (patient record # 194). It also ignored the word *stable*, followed by the concept “angina”, which had originally meant that the patient is currently not experiencing angina. (patient record # 119)
  - In some cases, doctors discussed the risk factors of coronary artery disease with a patient who had not yet been diagnosed with coronary artery disease. The current system incorrectly projected that those patients had a history of myocardial

infarction (patient records #266, #342).

- Discordance between the author and data annotators in the interpretation of a situation
    - A case in which a patient received coronary artery bypass graft surgery, which indicates severe myocardial infarction, was not labeled as “met” by the annotators (patient record # 277).

The errors that occurred in three false-negative cases are as below:

- Spurious labeling error
    - In patient record # 135, both I and my system could not find any clinical evidence of overt myocardial infarction in the patient. Though electrocardiography showed some ischemic changes (the attending physician then suspected this was a rate-related change rather than ischemia), cardiac catheterization revealed no evidence of coronary stenosis. The patient did not experience angina and denied chest pain or discomfort throughout the entire visit. The attending physician wrote in the patient record that “pt [patient] has already ruled out for MI [myocardial infarction].” The medication that the patient was taking was for treating cardiac valve diseases (mitral stenosis, mitral valve regurgitation, and aortic stenosis) and resultant heart failure, but not for coronary artery disease.
- Failure to detect relevant entities
    - Patient # 246 took two medications, atenolol and atorvastatin, for coronary artery disease. MetaMap failed to detect atenolol from the text.
    - The system failed to extract the concept “non-ST elevation myocardial infarction” from the sentence “she has completed a cardiac stress test as recommended after

suffering a NSTEMI during her last hospitalization." MetaMap extracted CUIs *C4255010*, *C1536222*, *C1536221*, and *C3537184* from the word phrase "non st elevation myocardial infarction," which was expanded from the trigger term "NSTEMI," and none of the extracted CUIs had a valid corresponding AUIs in SNOMED CT terms. Further scrutiny revealed that the January 2016 version of SNOMED CT has the concept **A3473213 | Acute non-ST segment elevation myocardial infarction |**, which only partially matched to but not exactly matched to the string "NSTEM" (patient record # 350).

### 4.5 Cohort Identification Performance: MAJOR-DIABETES

The system's precision, recall, and F1 score for the MAJOR-DIABETES criterion were 0.974, 0.884, and 0.927, respectively. After removing spurious patient records, the precision, recall, and F1 score for ADVANCED-CAD were 0.974, 0.974, and 0.974, respectively. The F1 score of the current framework for the MAJOR-DIABETES criterion was higher than that of the best system submitted to the 2018 n2c2. The similarity scores for each patient record are presented in Appendix B.

The similarity score above which a patient record was predicted to be *met* was 2.0 for the ABDOMINAL criteria. The median number of relevant AUIs (e.g., AUIs that contributed to the similarity scores of each patient record) in the ground-truth *met* and *not met* patient records was 16 (mean = 25.698, range 0-129) and 5 (mean = 6.279, range 2-19), respectively.

There were one false-positive case and 5 false-negative cases before the data with spurious labeling were removed. In the false-positive case, the current framework identified an (age-related) macular degeneration that was not associated with diabetes. Many false negatives were cases where coronary artery disease was considered by the annotators of the 2018 n2c2 dataset to be a major complication of diabetes (patient records # 115, 182, and 224). Although the MAJOR-DIABETES criterion defined major diabetes-related complications to be microvascular complications such as nephropathy, retinopathy, and neuropathy, there may have been an implicit agreement among the annotators of the 2018 n2c2 dataset to include macrovascular complications such as coronary artery disease in their definition of the major diabetes-related complications. As described in section 4.4, patient record # 135 had been mislabeled as "met" for the ADVANCED-CAD criterion, and this may have led to yet another misclassification of the patient record # 135 with respect to the MAJOR-DIABETES criterion by acknowledging that coronary artery disease was a major diabetic complication.

In another case, the current framework could not infer that the patient had diabetic retinopathy from the sentence "was recently hold he needs laser surgery (patient record # 166)."

# V. DISCUSSION

## 5.1 Primary Findings

This study demonstrated that a taxonomic structure of a biomedical ontology such as SNOMED CT can be utilized to identify eligible patients for clinical trials from free-text clinical narratives, especially to query medical knowledge by navigating hierarchical relations among entities. This is not surprising, given that SNOMED CT is commonly leveraged for storing and retrieving symptoms, disorders, tests, medications, and procedures in various types of clinical data models and data repositories.

The conventional use of SNOMED CT involves the retrieval of patient data using SNOMED CT codes input by clinicians; this use case is limited in retrieving cases from legacy systems or institutions where SNOMED CT codes are not used. Since SNOMED CT has not been adopted as a standardized terminology EHRs until 2013 in the U.S. and still not been available in many countries, SNOMED CT code data are not readily available for patient data retrieval in many cases. When SNOMED CT concepts were extracted from free-text clinical narratives, however, the advantage of using SNOMED CT for cohort identification tasks can extend to environments where the input of SNOMED CT codes is not supported.

The extraction of SNOMED CT concepts from the free text also allowed for more sophisticated queries of patients. From free-text clinical narratives, one can query diseases that a patient had been diagnosed with or procedures that they had undergone, but were not codified by clinicians due to the complexity in SNOMED CT code structure or the lack of time to input codes. Procedures such as small bowel obstruction can be seemingly unrelated to the current situation of a diabetic patient or irrelevant to the current practitioner's reimbursement, for example, and would be unlikely to be included on the patient's problem list by a doctor and lost due to discontinuity in care; such care-centric comments might only be found in free-text clinical notes. Since those major and minor problems that are more care-centric than cost-centric can be identified from free-text clinical notes only, cohort identification could be more sophisticated if SNOMED CT codes could be mapped from free-text clinical notes for cohort identification tasks.

Another benefit of using SNOMED CT, as shown in this study, is that it can be employed to measure similarities between queries (eligibility criteria) and patient records. The semantic similarity was useful in biomedical and computer science applications to determine how similar the two concepts are. The current study expanded the idea of semantic similarity between two individual concepts to the semantic similarity between queries and patient records. A unique challenge in measuring semantic similarity between queries and documents such as patient records was the disproportionate information quantity between the two, which was overcome by considering the patient record concepts that were the direct subtypes of a query concept when calculating similarity. This method prevented the current framework from violating some first principles of information retrieval: adding non-query terms to a document should not make it more relevant. By adding the logarithms of the frequency of concepts in a patient record, the method allowed more relevance to patient records that mentioned the query term more, while it gave diminishing marginal gain of relevance when it saw the same query term in the patient record.

Semantic similarity measurement between a query and patient records provides a quantitative measure of each patient's fitness in relation to the selection criteria. Quantified semantic similarity offers greater flexibility in recruiting patients than a binary determination of eligibility does; by calculating similarity and quantitative estimate of eligibility, clinical trial recruiters can control how extensively they will include potential subjects.

The proposed method is feasible since the UMLS Metathesaurus provides the taxonomic paths linked by *is_a* relations between concepts. By employing the Metathesaurus in the Python interface, navigating the paths between any concepts can be done to calculate the depths and number of subtype concepts. Adopting Python as an integrated processing interface also made the coordination among the various NLP subcomponents efficient. In this framework, regular expressions, document parsing, MySQL query, MetaMap interface, and similarity calculating were combined in the Python interface, uniting the whole NLP process of the current framework.

Preprocessing is a crucial step to successful cohort identification from free texts. Patient records often hold information about persons other than the patient (e.g., mother) and entities that have not yet occurred to the patient (e.g., treatment plans). Although the entities that were unrelated to the current situation of a patient could be fairly removed by empirical rules and regular expressions if the entities were syntactically located adjacent to the words, section division could warrant more reliable exclusion of entities that are unrelated to the current situation of patients. Even though the first attempt to exclude the concept of *myocardial infarction* extracted from the sentence "Family history: mother died of MI at age 59" failed, for example, it could be excluded from the final patient feature concept set by identifying and removing the entire section of *family history*.

Theoretically, the proposed similarity metric prevents the cohort identification framework from retrieving biomedical concepts that are irrelevant to queries by allocating zero weight to those concepts that are not subsumed by query concepts. This results in reduced processing time, which is of substantial advantage to cohort identification tasks where patient records inevitably contain much noisy information.

The performance of the proposed model is promising, yielding a higher F1 score than the best performer of the 2018 n2c2. The error analysis of the result shows that most errors were caused by the discordance between the author's and the data annotators' medical knowledge. For example, the author did not accept *myocardial infarction* as a major complication of diabetes while some n2c2 annotators inconsistently did so. Although this may cause a *prima facie* drop in the performance, it does not imply a flaw in the framework: a user may choose to include coronary artery disease in the definition of major complications of diabetes and, in that case, the performance of the framework will improve. Though knowledge engineering for the construction of Query Concept Feature Set is outside the current research's focus, it warrants additional independent research.

## 5.2 Additional Findings

The current research unexpectedly faced possible labeling errors in the test dataset of the 2018 n2c2. As discussed in section 1.1, labeling and annotation errors are not infrequent in publicly available data, which threatens the external validity of the models. Although deep learning is believed to be robust to label noise in *train* data, labeling errors in *test* datasets can destabilize benchmarks by which we measure progress in models (Northcutt et al. 2021). As Northcutt et al. (2021) put it, a greater deal of labeling errors is reflected in the high-capacity prediction models than in their counterparts with fewer parameters (like NasNet vs. ResNet-18, NasNet's fewer-parameter version). This is different from overfitting in which the quality of results worsens as the models try to learn too much from data; labeling errors are the problem of stability of the model, in which models with larger parameters produce worse predictions than their lower-capacity counterparts when evaluated on the corrected labels. This may lead to underestimation and abandonment of a useful model which could have been performed better in real-world settings.

The suspected labeling errors in the test data of the 2018 n2c2 were caused by the inconsistency in medical knowledge representation (patient records # 135 for DIABETES-MAJOR and # 277 for ADVANCED-CAD) or annotation policy (patient records # 115, 182, and 224 for DIABETES-MAJOR). Manual annotating and labeling of large-scale textual data are so extensive that annotators may lose track of what they have done and will suffer from inconsistent annotation practice. Labeling errors will be a perennial concern for the NLP community and lead to poor data utilization if not resolved.

## 5.3 Areas of Improvement

In this study, Query Concept Feature Sets were constructed manually at the medical expert's discretion. The automated generation of Query Feature Sets from natural language description of eligibility criteria was outside the current research's focus, though it would be worth additional independent research. A suggested alternative to automatic query generation is to build reference sets for each selected criterion. For example, "history of myocardial infarction" can be represented as a reference set that contains the element concepts **A7873496 |**

**Coronary arteriosclerosis |**, **A2938836 | Myocardial infarction |**, **A24246836 | Coronary artery stent |**, **A26575074 | EKG: myocardial infarction |**, and **A6919959 | Disorder of artery |**.

The workflow of cohort identification tasks presented in this study may differ from those in real-world settings. Query Concept Feature Sets in this study could not be formulated without reading through patient records in the training dataset to understand the implicit agreement on the more specified definition of each eligibility criterion. The 2018 n2c2 did not pronounce the specifics of each eligibility criterion (e.g., whether *dilation and curettage* is intra-abdominal surgery), and the designers of cohort identification systems had no choice but to induce the specifics of each criterion by examining the training dataset either manually or through machine learning algorithms. Since the eligibility criteria descriptions by themselves are limited in providing information about the implicit selection process, this study tried to translate the implied selection rationale into Query Concept Feature Sets by comparing biomedical entities that satisfied each eligibility criterion with those that did not. This process is not necessary in real-world settings; instead, recruiters will translate the eligibility criteria directly into Query Concept Feature Sets to define their needs. Therefore, manual construction of Query Concept Feature Sets does not significantly restrict the application of the current framework to cohort identification in real-world settings because users can control defining eligibility criteria using their own set of concepts under their own authority. For instance, instead of excluding "Endometrial scaping" from the definition of *intra-abdominal surgery*, a user may choose to exclude "Excision of pelvis" in order not to take *hysterectomy* as well as *dilation and curettage* into account.

### 5.4 Limitations

It should be acknowledged that the current study is focused more on testing the ability of SNOMED CT's semantic structure to query common EHR phenotypes (i.e., symptoms, diagnoses, procedures, medications, etc.) for the identification of patients from free-text clinical narratives, than it is about developing a cohort identification system that is ready to be deployed in real-world clinical or research settings. The current framework showed that care-centric terminologies such as SNOMED CT can represent clinical phenotypes embedded in the unstructured part of EHRs for a cohort identification task. This implies that SNOMED CT can make detailed information about a patient more readily sharable and discoverable, and that a cohort identification system based on SNOMED CT's semantic structure promises better accessibility to patients across multiple healthcare systems, leading to a broader patient cohort.

One of the limitations of the current study is that it was tested upon three highly medical criteria only. Those three are thought to be the best testbed for SNOMED CT in that they allowed it to leverage the hierarchical relations between query and patient records. It is expected that supervised machine learning methods, which generate classifiers by supervised learning from training data input, would have an advantage in inference—a process examined in the ENGLISH and MAKES-DECISION criteria (Table 1)—over ontologies. The proposed framework, in addition, did not experiment with a laboratory results extraction task—a process examined in the HBA1C and CREATININE criteria (Table 1)—where the rule-based approach may have outperformed an ontology-based system (Stubbs et al. 2019). These limitations suggest the extent to which SNOMED CT can be further used for cohort identification tasks—future work will provide a hybrid system where SNOMED CT is employed along with supervised machine learning and rules to demonstrate its performance in all thirteen criteria of the n2c2.

Another limitation is the fact that the proposed framework has not been tested on test data outside 2018 n2c2. A possible scenario of non-transferability of the current framework is mostly caused by a failure to appropriately preprocess free-text clinical notes produced at other institutions. The provenance of the current dataset was a single institution (i.e., UTHealth), and there may be a particular style of writing shared by clinicians in that institution, which makes the processing of unstructured narratives non-transferable to other hospitals. The organizers of 2018 n2c2 separated visit records by a line of asterisks (*) followed by "Record date: *YYYY-MM-DD*" to merge different visit records for research purposes. The method adopted in this study to separate a patient record by visit date may not be generalizable to other datasets. When the current framework is to be employed in a real-world setting, a distinct strategy for merging visit records within a single patient record needs to be established.

Writing styles may also vary by individual clinicians: one may prefer using unofficial abbreviations while others may not. While I have tried to cover all possible variations of clinicians' writing styles in the current framework, taking into account as many abbreviations I can imagine used in clinical settings as possible and as many ways to make sections (e.g., present illness, past medical history, etc.) discrete in free-text notes as possible, no preprocessing methods will be perfectly transferable to other institutional environments. However, owing to standardized medical education within the U.S. as well as worldwide, the discrepancy in writing styles among clinicians is not expected to be substantial. For example, the abbreviation "fh" represents "family history" both in New York and California. Rather, in this study, I am proposing the basic idea of exploiting SNOMED CT's semantic relations extracted from free-text clinical notes for identifying patient cohorts.

Although some noisy SNOMED CT concepts can arise from non-readily-transferable preprocessing of clinical notes when the proposed framework is tested upon other datasets, the performance is not expected to drop significantly, since the current framework considers the descendants of query concepts only. Therefore, even though the concept “Texas (state)” was mistakenly extracted from the trigger word “tx,” which originally meant *treatment*, this will not contribute to any errors unless the query contains any terms that refer to states.

While the ground-truth labels of the 2018 n2c2 dataset were binary (“met” and “not met”), the current framework yielded numeric similarity scores. To reconcile the binary classification with the numeric score, threshold scores were set at the point where the best F1 score could be achieved. This method of transforming numeric scores to binary classifications and finding the optimal threshold score may not be feasible in real-world settings. An optimal threshold score is subject to change by clinical trials, and it only provides information about the extent to which a patient record is semantically similar to an eligibility description (not eligibility class). Since the 2018 n2c2 dataset had provided no information on how much each patient is eligible for each clinical trial, the performance of the current framework could not be assessed to its greatest detail; only a crude estimation of classification performance could be measured by transforming numeric similarity scores to binary predictions. If the 2018 n2c2 dataset could be re-annotated with numeric figures that can tell how close each patient is to each clinical trial, the performance of the current framework could be measured in a much more sophisticated manner, ideally in terms of correlation coefficients.

## 5.5 Conclusion

The current study demonstrated that the semantic structure of SNOMED CT can be used for cohort identification task to query clinical phenotypes embedded in the unstructured part of EHRs. SNOMED CT was leveraged for cohort identification from free-text clinical notes, without referring to training data that requires extensive annotation and labeling. A novel similarity metric was developed using existing technology (i.e., MetaMap) and ontology (e.g., SNOMED CT), which can effectively and efficiently perform cohort identification. The hierarchical semantic relations of SNOMED CT measured the semantic similarity between eligibility criteria and each patient record and quantified how well the patient fits the eligibility criteria. Future research is suggested to develop a hybrid system that integrates ontology and machine learning-based approaches to enhance other NLP components such as inference in cohort identification tasks.

## **APPENDIX A**: AUIS WITH LOWER QUANTIFIED IMPORTANCE (<0.5), THEIR DESCRIPTIONS, DEPTH LEVELS, AND EXAMPLE OF SUBTYPE CONCEPTS.

| AUI | Description | Depth level | Example of subtype concept |
|---|---|---|---|
| A3241965 | Activity of daily living | 5 | Walking |
| A16369886 | Interpretation of findings | 3 | Diagnosis |
| A2873197 | Proton | 6 | Proton (to prevent mapping of a single character "h") |
| A3016529 | Drinks | 4 | Drinks |
| A2881738 | Foods | 4 | Egg fries |
| A3161951 | Natural material | 4 | Tree resins |
| A3738091 | Substance categorized by physical state | 3 | Liquid |
| A3394388 | Dietary product | 3 | Gluten-free cake mix |
| A29562204 | Medicinal product | 3 | Product containing aluminum |
| A9417638 | Vision test distance | 7 | 1/3 meter (to prevent mapping of fraction figures, e.g., 1/3) |
| A6917682 | Complaint | 6 | Chief Complaint |
| A3709940 | Social and personal history finding | 5 | Returning home |
| A2873229 | Problem | 5 | Problem |
| A3709956 | Social context finding | 5 | Financially poor |
| A13356695 | Finding related to biological sex | 4 | Male |
| A3469753 | Functional finding* | 4 | Functional finding |
| A3362993 | Color finding | 3 | Dark color |
| A3900203 | Adverse incident outcome categories | 3 | Adverse incident resulting in death |
| A3242802 | Administrative statuses | 3 | Report status |
| A3629808 | Physical examination procedure* | 6 | Physical examination procedure |
| A7881343 | Review of systems* | 5 | Review of systems |
| A3300149 | Therapeutic procedure* | 4 | Therapeutic procedure |
| A3092627 | Administrative procedure | 3 | Hospital admission |
| A2882492 | Homo sapiens | 13 | Homo sapiens (to prevent mapping of "man") |
| A3161523 | N+ | 6 | N+ (to prevent mapping of a single character "n") |

| AUI | Description | Depth level | Example of subtype concept |
|---|---|---|---|
| A3204387 | Staging and scales | 2 | Breslow system for melanoma staging |
| A3566528 | Monitoring* | 5 | Monitoring |
| A3192955 | Relative times | 5 | Before |
| A13009256 | Evaluation - action* | 4 | Evaluation - action |
| A3214246 | Types | 4 | Jones (to prevent mapping of person name) |
| A3091961 | Action* | 3 | Action |
| A10884570 | Clinical specialty | 3 | Cardiology |
| A10884570 | Certainties* | 4 | Certainties (to prevent mapping of "certain") |
| A3281759 | Finding status values | 3 | Worsening |
| A28896694 | Does form intended site | 3 | Enteral |
| A3060341 | Present* | 5 | Present |
| A3089702 | Absence findings | 4 | Negative (to preserve negation expressions in sentence) |
| A3132009 | Finding values | 3 | Positive |
| A3473454 | General clinical stage for disease AND/OR neoplasm | 3 | Late stage |
| A23454121 | Mechanism of disease spread value | 3 | Hematogenous spread |
| A2887024 | Surgery* | 4 | Surgery |
| A3157401 | Mechanisms | 3 | Barotrauma mechanisms |
| A3566118 | Modifiers of Analytes and Substances | 3 | Beta subunit (to prevent single-handed mapping of "beta" in "beta blockers") |
| A28878916 | Pharmaceutical dose form | 3 | Conventional release oral syrup |
| A24090101 | Precondition value | 3 | At rest |
| A27784444 | Process* | 3 | Process |
| A3137011 | Grades | 5 | Mild |
| A3110547 | Classes* | 4 | Classes |
| A3123635 | Editions* | 4 | Editions |
| A3137422 | Groups | 4 | Group A (to prevent mapping of letter "A") |
| A3152689 | Levels | 4 | Intermediate |
| A3198151 | Scores* | 4 | Scores |
| A3204385 | Stages | 4 | Stage 1A (to prevent mapping of number-letter combinations) |
| A3214246 | Types | 4 | Serotype H3N2 (to prevent mapping of number-letter combinations) |

| AUI | Description | Depth level | Example of subtype concept |
|---|---|---|---|
| A3194939 | Result comments | 3 | Normal gait (to prevent mapping of "normal") |
| A6947385 | Route of administration value | 3 | Intravenous route (to prevent mapping of "IV") |
| A3200743 | Side* | 5 | Side |
| A7873282 | Sport | 3 | Baseball |
| A28898878 | State of matter | 3 | Liquid |
| A24089189 | Technique | 3 | Mechanical |
| A3755612 | Time frame | 3 | Daily |
| A3393038 | Dialysis dosage form | 4 | Hemodialysis solution |
| A3657907 | Radiopharmaceutical dosage form | 4 | Radioactive implant |
| A7880590 | Percentage unit | 5 | % positive cells (to prevent mapping of "%") |
| A3087908 | per forty* | 5 | /40 (to prevent mapping of a numeric figure "40") |
| A29521976 | Unit of measure | 3 | Months |
| A3030642 | History of | 4 | History of |
| A3243208 | Times relative to admission | 7 | On admission |
| A3126026 | Event orders | 6 | Initially |
| A3134454 | Frequencies | 5 | Continuous |
| A3100796 | Behavior descriptors | 4 | Complicated |
| A3296880 | Pathogenesis | 7 | Drug-induced |
| A3192915 | Relationships | 4 | Autologous |
| A3199217 | Sensibilities | 4 | Olfactory |
| A3120607 | Directions | 4 | Elevation - value |
| A3269797 | Courses* | 5 | Courses |
| A3121378 | Distributions* | 4 | Distributions |
| A3122687 | Durations* | 4 | Durations |
| A3134454 | Frequencies* | 4 | Frequencies |
| A3171632 | Occurrences* | 4 | Occurrencies |
| A3188236 | Priorities* | 4 | Priorities |
| A3210757 | Time patterns | 4 | Acquired (to prevent single-handed mapping of "acquired" in "acquired immunodeficiency syndrome") |
| A3174565 | Ordinal number | 4 | Secondary |
| A3214887 | Uniformities | 4 | Irregular |
| A3214887 | Velocities* | 4 | Velocities |

| AUI | Description | Depth level | Example of subtype concept |
|---|---|---|---|
| A7873501 | Descriptor | 3 | More |
| A10885554 | Dosing instruction fragment | 3 | During - dosing instruction fragment |
| A3285870 | Healthcare professional | 5 | Doctor |
| A13021175 | Person in the healthcare environment | 4 | Patient |
| A6945981 | Person categorized by age | 4 | Women |
| A7881113 | Racial group | 4 | Caucasian |
| A10884564 | Clinical equipment and/or device | 4 | Audiometric room |
| A3299532 | Study* | 10 | Study |
| A2981767 | St. Lucia | 6 | St. Lucia (to prevent single-handed mapping of "ST" in "ST segment") |
| A3286724 | Location within hospital premises | 5 | Intensive care unit |
| A3819332 | Site of care | 4 | Medical center |

* The concept itself gets the quantified importance of < 0.5 and its quantified importance does not impact on its subtype concepts.

## **APPENDIX B**. SIMILARITY SCORES OF PATIENT RECORDS FOR EACH ELIGIBILITY CRITERIA

### 1. ABDOMINAL

| **Patient Record** | **Ground-Truth Label (0: *not met*, 1: *met*)** | **Similarity Score** | **Patient Record** | **Ground-Truth Label (0: *not met*, 1: *met*)** | **Similarity Score** |
|---:|---:|---:|---:|---:|---:|
| 108 | 1 | 0.19939875 | 269 | 1 | 0.30074149 |
| 115 | 1 | 0.2033543 | 270 | 0 | 0 |
| 118 | 1 | 0.32610288 | 274 | 0 | 0 |
| 119 | 0 | 0 | 276 | 0 | 0 |
| 120 | 0 | 0 | 277 | 0 | 0 |
| 131 | 0 | 0 | 282 | 0 | 0 |
| 135 | 0 | 0.40986719 | 285 | 1 | 0.45841977 |
| 137 | 0 | 0 | 293 | 0 | 0 |
| 140 | 1 | 0.61905311 | 294 | 0 | 0 |
| 141 | 0 | 0 | 296 | 1 | 1.34172682 |
| 153 | 0 | 0 | 305 | 0 | 0 |
| 155 | 0 | 0 | 309 | 0 | 0 |
| 156 | 1 | 0.25704762 | 314 | 1 | 0.22574903 |
| 158 | 1 | 0.32230894 | 317 | 0 | 0 |
| 165 | 1 | 0.09090992 | 320 | 0 | 0 |
| 166 | 0 | 0 | 321 | 0 | 0 |
| 167 | 1 | 0.28298893 | 322 | 1 | 0.18508333 |
| 171 | 1 | 0.73453488 | 323 | 1 | 1.264964 |
| 173 | 1 | 0.12819482 | 327 | 0 | 0 |
| 178 | 0 | 0 | 328 | 1 | 0.28867268 |
| 182 | 1 | 0.08587637 | 342 | 1 | 0.12220492 |
| 190 | 0 | 0 | 343 | 0 | 0 |
| 192 | 0 | 0 | 346 | 0 | 0 |
| 193 | 0 | 0 | 347 | 1 | 0 |
| 194 | 0 | 0 | 348 | 0 | 0 |
| 201 | 0 | 0 | 350 | 0 | 0 |
| 205 | 0 | 0 | 351 | 0 | 0 |
| 213 | 0 | 0 | 352 | 1 | 0.22198719 |
| 216 | 1 | 0.03389064 | 353 | 0 | 0 |
| 217 | 0 | 0 | 359 | 1 | 0.03389064 |
| 218 | 1 | 0 | 360 | 0 | 0 |
| 224 | 0 | 0 | 361 | 1 | 0.17175274 |
| 227 | 0 | 0 | 368 | 1 | 0.29212562 |
| 229 | 0 | 0 | 369 | 1 | 0.28923067 |
| 232 | 0 | 0 | 370 | 0 | 0 |
| 233 | 0 | 0 | 373 | 0 | 0 |
| 236 | 1 | 0 | 379 | 0 | 0 |
| 239 | 0 | 0 | 383 | 0 | 0 |
| 240 | 0 | 0 | 384 | 0 | 0 |
| 246 | 0 | 0 | 386 | 1 | 0.37656796 |
| 251 | 0 | 0 | 394 | 1 | 0.72672658 |
| 265 | 0 | 0 | 396 | 0 | 0 |
| 266 | 0 | 0 | 399 | 0 | 0 |

## 2. ADVANCED-CAD

| Patient Record | Ground-Truth Label (0: *not met*, 1: *met*) | Similarity Score | Patient Record | Ground-Truth Label (0: *not met*, 1: *met*) | Similarity Score |
|---:|---:|---:|---:|---:|---:|
| 108 | 1 | 5.65501471 | 269 | 1 | 3.81536883 |
| 115 | 0 | 1.62925934 | 270 | 1 | 4.14874033 |
| 118 | 1 | 6.32154951 | 274 | 1 | 4.03950843 |
| 119 | 0 | 3.19611291 | 276 | 1 | 4.33056497 |
| 120 | 0 | 0 | 277 | 0 | 1.84582156 |
| 131 | 1 | 6.64275676 | 282 | 1 | 2.62561885 |
| 135 | 1 | 0 | 285 | 1 | 1.72846692 |
| 137 | 1 | 7.86296857 | 293 | 1 | 4.59379671 |
| 140 | 0 | 0 | 294 | 1 | 4.26179305 |
| 141 | 1 | 5.76449943 | 296 | 1 | 1.51067821 |
| 153 | 1 | 3.14433263 | 305 | 0 | 0 |
| 155 | 1 | 4.20913292 | 309 | 0 | 0 |
| 156 | 0 | 3.68413988 | 314 | 1 | 4.64022968 |
| 158 | 1 | 5.69785006 | 317 | 1 | 2.17096375 |
| 165 | 1 | 2.85745935 | 320 | 0 | 0 |
| 166 | 1 | 5.3868063 | 321 | 0 | 0 |
| 167 | 1 | 2.49303301 | 322 | 0 | 0 |
| 171 | 1 | 5.03032112 | 323 | 0 | 0 |
| 173 | 1 | 1.13341559 | 327 | 0 | 0 |
| 178 | 1 | 4.24264911 | 328 | 0 | 0 |
| 182 | 1 | 5.71238305 | 342 | 0 | 2.8545091 |
| 190 | 1 | 1.70483762 | 343 | 0 | 0 |
| 192 | 0 | 0 | 346 | 0 | 0 |
| 193 | 0 | 0 | 347 | 0 | 0 |
| 194 | 0 | 4.79062329 | 348 | 0 | 0 |
| 201 | 1 | 7.82275028 | 350 | 1 | 0 |
| 205 | 0 | 2.23372449 | 351 | 0 | 0 |
| 213 | 1 | 1.9066364 | 352 | 1 | 4.32396341 |
| 216 | 1 | 7.65696804 | 353 | 0 | 0 |
| 217 | 1 | 4.15368854 | 359 | 0 | 0 |
| 218 | 1 | 6.71630761 | 360 | 0 | 0 |
| 224 | 1 | 5.9244181 | 361 | 0 | 0 |
| 227 | 1 | 7.04029076 | 368 | 0 | 0 |
| 229 | 1 | 5.14753675 | 369 | 0 | 0 |
| 232 | 1 | 3.38736823 | 370 | 0 | 0 |
| 233 | 1 | 2.30898206 | 373 | 0 | 0 |
| 236 | 1 | 2.5104864 | 379 | 0 | 0 |
| 239 | 1 | 2.23798318 | 383 | 0 | 0 |
| 240 | 1 | 5.55225805 | 384 | 0 | 0 |
| 246 | 1 | 0 | 386 | 0 | 0 |
| 251 | 0 | 0 | 394 | 0 | 0 |
| 265 | 1 | 3.12998901 | 396 | 0 | 2.10083378 |
| 266 | 0 | 3.10541547 | 399 | 0 | 0 |

## 3. MAJOR-DIABETES

| Patient Record | Ground-Truth Label (0: *not met*, 1: *met*) | Similarity Score | Patient Record | Ground-Truth Label (0: *not met*, 1: *met*) | Similarity Score |
|---:|---:|---:|---:|---:|---:|
| 108 | 0 | 0 | 269 | 1 | 1.61478181 |
| 115 | 1 | 0 | 270 | 1 | 6.55740748 |
| 118 | 0 | 0 | 274 | 0 | 0 |
| 119 | 0 | 0 | 276 | 0 | 0 |
| 120 | 0 | 0 | 277 | 0 | 0 |
| 131 | 1 | 1.78313927 | 282 | 0 | 0.12877817 |
| 135 | 1 | 0 | 285 | 0 | 0 |
| 137 | 0 | 0 | 293 | 0 | 0 |
| 140 | 1 | 3.46232907 | 294 | 0 | 0 |
| 141 | 0 | 0 | 296 | 1 | 0.82241593 |
| 153 | 1 | 0.20827518 | 305 | 0 | 0 |
| 155 | 0 | 0 | 309 | 1 | 0.54298671 |
| 156 | 1 | 0.78248673 | 314 | 1 | 1.09395318 |
| 158 | 0 | 0 | 317 | 1 | 1.64318162 |
| 165 | 1 | 1.07418725 | 320 | 0 | 0 |
| 166 | 1 | 0 | 321 | 1 | 2.02892486 |
| 167 | 0 | 0 | 322 | 1 | 1.85578314 |
| 171 | 0 | 0 | 323 | 1 | 0.52315068 |
| 173 | 1 | 2.30378125 | 327 | 1 | 0.22026415 |
| 178 | 0 | 0 | 328 | 0 | 0 |
| 182 | 1 | 0 | 342 | 0 | 0 |
| 190 | 1 | 0.66446982 | 343 | 1 | 0.54869589 |
| 192 | 1 | 0.60377184 | 346 | 1 | 0.29234639 |
| 193 | 0 | 0 | 347 | 0 | 0 |
| 194 | 0 | 0 | 348 | 0 | 0 |
| 201 | 1 | 0.64932599 | 350 | 1 | 1.53466451 |
| 205 | 1 | 0.70107906 | 351 | 1 | 3.21220359 |
| 213 | 0 | 0 | 352 | 1 | 0.9239485 |
| 216 | 1 | 0.32941007 | 353 | 1 | 1.15933638 |
| 217 | 0 | 0.27476711 | 359 | 0 | 0 |
| 218 | 0 | 0 | 360 | 0 | 0 |
| 224 | 1 | 0 | 361 | 1 | 6.01957264 |
| 227 | 1 | 2.91644345 | 368 | 1 | 1.00793085 |
| 229 | 0 | 0 | 369 | 0 | 0 |
| 232 | 1 | 4.59201478 | 370 | 1 | 0.64838881 |
| 233 | 0 | 0 | 373 | 0 | 0 |
| 236 | 1 | 0.51691065 | 379 | 0 | 0 |
| 239 | 1 | 3.86202308 | 383 | 0 | 0 |
| 240 | 0 | 0 | 384 | 0 | 0 |
| 246 | 0 | 0 | 386 | 0 | 0 |
| 251 | 1 | 0.31650687 | 394 | 1 | 0.38479741 |
| 265 | 0 | 0 | 396 | 1 | 3.43054753 |
| 266 | 0 | 0 | 399 | 1 | 1.92584209 |

# REFERENCES

Abhyankar S, Goodwin RM, Sontag M, Yusuf C, Ojodu J, McDonald CJ. 2015. An update on the use of health information technology in newborn screening. *Semin. Perinatol.* 39(3):188–93

Abidi SSR, Singh AK, Christie S. 2016. Transcription of case report forms from unstructured referral letters: A semantic text analytics approach. *Stud. Health Technol. Inform.* 228:322–26

Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, et al. 2018. Natural language processing of clinical notes for identification of critical limb ischemia. *Int. J. Med. Inform.* 111:83–89

Al-Mubaid H, Nguyen HA. 2009. Measuring semantic similarity between biomedical concepts within multiple ontologies. *IEEE Trans. Syst., Man, Cybern. C*. 39(4):389–98

Ali T, Hussain M, Ali Khan W, Afzal M, Hussain J, et al. 2017. Multi-model-based interactive authoring environment for creating shareable medical knowledge. *Comput. Methods Programs Biomed.* 150:41–72

Alonso I, Contreras D. 2016. Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach. *Expert Syst. Appl.* 44:386–99

Arbabi A, Adams DR, Fidler S, Brudno M. 2019. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR Med. Inform.* 7(2):e12596

Arguello-Casteleiro M, Stevens R, Des-Diz J, Wroe C, Fernandez-Prieto MJ, et al. 2019. Exploring semantic deep learning for building reliable and reusable one health knowledge from PubMed systematic reviews and veterinary clinical notes. *J. Biomed. Semantics*. 10(Suppl 1):22

Aronson AR, Lang F-M. 2010. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.* 17(3):229–36

Aronson AR. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proc. AMIA Symp.* 17–21

Asgari MM, Eide MJ, Warton EM, Fletcher SW. 2013. Validation of a large basal cell carcinoma registry. *J. Registry Manag.* 40(2):65–69

Asklund T, Malmström A, Bergqvist M, Björ O, Henriksson R. 2015. Brain tumors in Sweden: data from a population-based registry 1999-2012. *Acta Oncol.* 54(3):377–84

Bache R, Miles S, Taweel A. 2013. An adaptable architecture for patient cohort identification from diverse data sources. *J. Am. Med. Inform. Assoc.* 20(e2):e327-33

Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. 2016. A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci. Data*. 3:160026

Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, et al. 2019. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artif. Intell. Med.* 97:79–88

Barros JM, Duggan J, Rebholz-Schuhmann D. 2018. Disease mentions in airport and hospital geolocations expose dominance of news events for disease concerns. *J. Biomed. Semantics*. 9(1):18

Batet M, Sánchez D, Valls A, Gibert K. 2013. Semantic similarity estimation from multiple ontologies. *Appl. Intell.* 38(1):2013

Beauharnais CC, Larkin ME, Zai AH, Boykin EC, Luttrell J, Wexler DJ. 2012. Efficacy and cost-effectiveness of an automated screening algorithm in an inpatient clinical trial. *Clin. Trials*. 9(2):198–203

Becker M, Böckmann B. 2017. Personalized guideline-based treatment recommendations using natural language processing techniques. *Stud. Health Technol. Inform.* 235:271–75

Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021. On the dangers of stochastic parrots: can language models be too big? □. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–23. New York, NY, USA: ACM

Blobel B. 2010. Architectural approach to eHealth for enabling paradigm changes in health. *Methods Inf. Med.* 49(2):123–34

Budanitsky A, Hirst G. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*. 32(1):13–47

Butt L, Zuccon G, Nguyen A, Bergheim A, Grayson N. 2013. Classification of cancer-related death certificates using machine learning. *Australas. Med. J.* 6(5):292–99

Castro VM, Dligach D, Finan S, Yu S, Can A, et al. 2017. Large-scale identification of patients with cerebral aneurysms using natural language processing. *Neurology*. 88(2):164–68

Caviedes JE, Cimino JJ. 2004. Towards the development of a conceptual distance metric for the UMLS. *J. Biomed. Inform.* 37(2):77–85

Chakrabarti S, Sen A, Huser V, Hruby GW, Rusanov A, et al. 2017. An interoperable similarity-based cohort identification method using the OMOP Common Data Model version 5.0. *J. Healthc. Inform. Res.* 1(1):1–18

Chandar P, Yaman A, Hoxha J, He Z, Weng C. 2015. Similarity-based recommendation of new concepts to a terminology. *AMIA Annu. Symp. Proc.* 2015:386–95

Chang E, Mostafa J. 2021. The use of SNOMED CT, 2013-2020: a literature review. *J. Am. Med. Inform. Assoc.* 28(9):2017-26

Chan LWC, Wong SCC, Chiau CC, Chan T-M, Tao L, et al. 2017. Association patterns of ontological features signify electronic health records in liver cancer. *J. Healthc. Eng.* 2017:6493016

Chen C-J, Warikoo N, Chang Y-C, Chen J-H, Hsu W-L. 2019a. Medical knowledge infused convolutional neural networks for cohort selection in clinical trials. *J. Am. Med. Inform. Assoc.* 26(11):1227–36

Chen L, Gu Y, Ji X, Lou C, Sun Z, et al. 2019b. Clinical trial cohort selection based on multi-level rule-based natural language processing system. *J. Am. Med. Inform. Assoc.* 26(11):1218–26

Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. 2015. A study of active learning methods for named entity recognition in clinical text. *J. Biomed. Inform.* 58:11–18

Collins R. 1992. Ethics in clinical trials. In *Introducing New Treatments for Cancer: Practical, Ethical and Legal Problems*, ed. CJ Williams, pp. 49–56. Chichester: John Wiley & Sons

D'Amore JD, Mandel JC, Kreda DA, Swain A, Koromia GA, et al. 2014. Are Meaningful Use Stage 2 certified EHRs ready for interoperability? Findings from the SMART C-CDA Collaborative. *J. Am. Med. Inform. Assoc*. 21(6):1060-8

Danahey K, Borden BA, Furner B, Yukman P, Hussain S, et al. 2017. Simplifying the use of pharmacogenomics in clinical practice: building the genomic prescribing system. *J. Biomed. Inform.* 75:110–21

de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J. Am. Med. Inform. Assoc.* 18(5):557–62

Elkin PL, Mullin S, Sakilay S. 2018. Biomedical Informatics Investigator. *Stud. Health Technol. Inform.* 255:195–99

Fletcher B, Gheorghe A, Moore D, Wilson S, Damery S. 2012. Improving the recruitment activity of clinicians in randomised controlled trials: A systematic review. *BMJ Open*. 2(1):e000496

Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. 2016. Extracting information from the text of electronic medical records to improve case detection: A systematic review. *J. Am. Med. Inform. Assoc.* 23(5):1007–15

Geiger RS, Cope D, Ip J, Lotosh M, Shah A, et al. 2021. "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*. 1–33

Girardi D, Wartner S, Halmerbauer G, Ehrenmüller M, Kosorus H, Dreiseitl S. 2016. Using concept hierarchies to improve calculation of patient similarity. *J. Biomed. Inform.* 63:66–73

Gøeg KR, Cornet R, Andersen SK. 2015. Clustering clinical models from local electronic health records based on semantic similarity. *J. Biomed. Inform.* 54:294–304

Greenbaum NR, Jernite Y, Halpern Y, Calder S, Nathanson LA, et al. 2019. Improving documentation of presenting problems in the emergency department using a domain-specific ontology and machine learning-driven user interfaces. *Int. J. Med. Inform.* 132:103981

Greibe K. 2013. Development of a SNOMED CT based national medication decision support system. *Stud. Health Technol. Inform.* 192:1147

Gruber TR. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Systems Laboratory September 1992 Technical Report*. *KSL 92-71*, Knowledge Systems Laboratory, Stanford, CA

Grundel B, Bernardeau M-A, Langner H, Schmidt C, Böhringer D, et al. 2021. [Extraction of features from clinical routine data using text mining]. *Ophthalmologe*. 118(3):264–72

Guien C, Blandin G, Lahaut P, Sanson B, Nehal K, et al. 2018. The French National Registry of patients with facioscapulohumeral muscular dystrophy. *Orphanet J. Rare Dis.* 13(1):218

Haase P, Siebes R, van Harmelen F. 2004. Peer selection in peer-to-peer networks with semantic topologies. , pp. 108–25. Semantics for Grid Databases

Hadj Taieb MA, Ben Aouicha M, Ben Hamadou A. 2014. Ontology-based approach for measuring semantic similarity. *Eng. Appl. Artif. Intell.* 36:238–61

Hagen MS, Jopling JK, Buchman TG, Lee EK. 2013. Priority queuing models for hospital intensive care units and impacts to severe case patients. *AMIA Annu. Symp. Proc.* 2013:841–50

Harispe S, Sánchez D, Ranwez S, Janaqi S, Montmain J. 2014. A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain. *J. Biomed. Inform.* 48:38–53

Hendriks MP, Verbeek XAAM, van Vegchel T, van der Sangen MJC, Strobbe LJA, et al. 2019. Transformation of the national breast cancer guideline into data-driven clinical decision trees. *JCO Clin. Cancer Inform.* 3:1–14

Hernandez AF, Fleurence RL, Rothman RL. 2015. The ADAPTABLE trial and PCORnet: Shining light on a new research paradigm. *Ann. Intern. Med.* 163(8):635–36

Hier DB, Pearson J. 2019. Two algorithms for the reorganisation of the problem list by organ system. *BMJ Health Care Inform.* 26(1)

Hitchins ARC, Hogan SC. 2018. Outcomes of early intervention for deaf children with additional needs following an Auditory Verbal approach to communication. *Int. J. Pediatr. Otorhinolaryngol.* 115:125–32

Hostetter J, Wang K, Siegel E, Durack J, Morrison JJ. 2015. Using standardized lexicons for report template validation with LexMap, a web-based application. *J. Digit. Imaging*. 28(3):309-14

Hsieh S-L, Chang W-Y, Chen C-H, Weng Y-C. 2013. Semantic similarity measures in the biomedical domain by leveraging a web search engine. *IEEE J. Biomed. Health Inform.* 17(4):853–61

Hubálek Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biol. Rev.* 57(4):669–89

Jani BD, Pell JP, McGagh D, Liyanage H, Kelly D, et al. 2020. Recording COVID-19 consultations: review of symptoms, risk factors, and proposed SNOMED CT terms. *Br J Gen Pract Open*. 4(4):bjgpopen20X101125

Jensen PB, Jensen LJ, Brunak S. 2012. Mining electronic health records: Towards better research applications and clinical care. *Nat. Rev. Genet.* 13(6):395–405

Jiang JJ, Conrath DW. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the 10th Research on Computational Linguistics International Conference*, pp. 19–33

Jia Z, Lu X, Duan H, Li H. 2019. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Med. Inform. Decis. Mak.* 19(1):91

Ji X, Ritter A, Yen P-Y. 2017. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *J. Biomed. Inform.* 69:33–42

Kahn CE. 2014. Annotation of figures from the biomedical imaging literature: a comparative analysis of RadLex and other standardized vocabularies. *Acad. Radiol.* 21(3):384–92

Karimi S, Metke-Jimenez A, Kemp M, Wang C. 2015. Cadec: A corpus of adverse drug event annotations. *J. Biomed. Inform.* 55:73–81

Kate RJ. 2013. Towards Converting Clinical Phrases into SNOMED CT Expressions. *Biomed. Inform. Insights*. 6(Suppl 1):29–37

Kimia AA, Savova G, Landschaft A, Harper MB. 2015. An introduction to natural language processing: how you can get more from those electronic notes you are generating. *Pediatr. Emerg. Care*. 31(7):536–41

Knowledge Base workgroup of the Observational Health Data Sciences and Informatics (OHDSI) collaborative. 2017. Large-scale adverse effects related to treatment evidence standardization (LAERTES): an open scalable system for linking pharmacovigilance evidence sources with clinical data. *J. Biomed. Semantics*. 8(1):11

Konstantinidis S, Fernandez-Luque L, Bamidis P, Karlsen R. 2013. The role of taxonomies in social media and the semantic web for health education. A study of SNOMED CT terms in YouTube health video tags. *Methods Inf. Med.* 52(2):168–79

Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, et al. 2015. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med. Inform. Decis. Mak.* 15:53

Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. 2018. Extracting cancer mortality statistics from death certificates: a hybrid machine learning and rule-based approach for common and rare cancers. *Artif. Intell. Med.* 89:1–9

Köpcke F, Lubgan D, Fietkau R, Scholler A, Nau C, et al. 2013. Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. *BMC Med. Inform. Decis. Mak.* 13:134

Kossovsky MP, Sarasin FP, Bolla F, Gaspoz JM, Borst F. 1999. Distinction between planned and unplanned readmissions following discharge from a Department of Internal Medicine. *Methods Inf. Med.* 38(2):140–43

Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y. 2019. Measuring bias in contextualized word representations. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 166–72. Stroudsburg, PA, USA: Association for Computational Linguistics

Lamy J-B, Venot A, Duclos C. 2015. PyMedTermino: an open-source generic API for advanced terminology services. *Stud. Health Technol. Inform.* 210:924–28

Leacock C, Chodorow M. 1998. Combining local context and WordNet similarity for word sense identification. In *WrodNet: An Electronic Lexical Database*, eds. C Fellbaum, G Miller, pp. 265–83. Cambridge, MA: The MIT Press

Lee J, Maslove DM, Dubin JA. 2015. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLoS ONE*. 10(5):e0127428

Lee W-N, Das AK. 2010. Local alignment tool for clinical history: temporal semantic search of clinical databases. *AMIA Annu. Symp. Proc.* 2010:437–41

Lerner I, Paris N, Tannier X. 2020. Terminologies augmented recurrent neural network model for clinical named entity recognition. *J. Biomed. Inform.* 102:103356

Liaqat S, Pineda J, Velayutham J, Lee A, Reicher J, et al. 2020. Sharing is Caring: Exploring machine learning-enabled methods for regional medical imaging exchange using procedure metadata. *AMIA Jt Summits Transl Sci Proc*. 2020:383–92

Liaw S-T, Taggart J, Yu H, de Lusignan S, Kuziemsky C, Hayen A. 2014. Integrating electronic health record information to support integrated care: practical application of ontologies to improve the accuracy of diabetes disease registers. *J. Biomed. Inform.* 52:364–72

Lindman K, Rose JF, Lindvall M, Lundström C, Treanor D. 2019. Annotations, ontologies, and whole slide images - development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue. *J. Pathol. Inform.* 10:22

Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, et al. 2014. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J. Am. Med. Inform. Assoc.* 21(3):406–13

Lin C, Hsu C-J, Lou Y-S, Yeh S-J, Lee C-C, et al. 2017. Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. *J. Med. Internet Res.* 19(11):e380

Lin D. 1998. Automatic retrieval and clustering of similar words. *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics* -, pp. 768–74. Morristown, NJ, USA: Association for Computational Linguistics

Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, et al. 2013. An information extraction framework for cohort identification using electronic health records. *AMIA Jt Summits Transl Sci Proc*. 2013:149–53

Li J. 2018. *Ontology-based clinical information extraction using SNOMED CT*. Doctoral dissertation thesis

Lomonaco V, Martoglia R, Mandreoli F, Anderlucci L, Emmett W, et al. 2014. UCbase 2.0: ultraconserved sequences database (2014 update). *Database (Oxford)*. 2014:

Mabotuwana T, Lee MC, Cohen-Solal EV. 2013. An ontology-based similarity measure for biomedical data-application to radiology reports. *J. Biomed. Inform.* 46(5):857–68

Maheronnaghsh R, Nezareh S, Sayyah M-K, Rahimi-Movaghar V. 2013. Developing SNOMED-CT for decision making and data gathering: a software prototype for low back pain. *Acta Med. Iran.* 51(8):548–53

Manohar N. Adam TJ, Pakhomov SV, Melton GB, Zhang R. 2015. Evaluation of herbal and dietary supplement resource term coverage. *Stud. Health Technol. Inform. 216:785-9*

Ma H, Weng C. 2016. Prediction of black box warning by mining patterns of Convergent Focus Shift in clinical trial study populations using linked public data. *J. Biomed. Inform.* 60:132–44

Martínez García L, Sanabria AJ, Araya I, Lawson J, Solà I, et al. 2015. Efficiency of pragmatic search strategies to update clinical guidelines recommendations. *BMC Med. Res. Methodol.* 15:57

Martínez S, Sánchez D, Valls A. 2013. A semantic framework to protect the privacy of electronic health records with non-numerical attributes. *J. Biomed. Inform.* 46(2):294–303

McInnes BT, Pedersen T, Liu Y, Melton GB, Pakhomov SV. 2014. U-path: an undirected path-based measure of semantic similarity. *AMIA Annu. Symp. Proc.* 2014:882–91

McInnes BT, Pedersen T. 2015. Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. *J. Biomed. Inform.* 54:329–36

Mc Cord KA, Hemkens LG. 2019. Using electronic health records for clinical trials: Where do we stand and where can we go? *CMAJ*. 191(5):E128–33

Metke-Jimenez A, Steel J, Hansen D, Lawley M. 2018. Ontoserver: a syndicated terminology server. *J. Biomed. Semantics*. 9(1):24

Meystre SM, Ferrández Ó, Friedlin FJ, South BR, Shen S, Samore MH. 2014. Text de-identification for privacy protection: a study of its impact on clinical text information content. *J. Biomed. Inform.* 50:142–50

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013. Distributed representations of words and phrases and their compositionality. *arXiv*.

Millar J. 2016. The need for a global language - SNOMED CT introduction. *Stud. Health Technol. Inform.* 225:683–85

Minard A-L, Ligozat A-L, Ben Abacha A, Bernhard D, Cartoni B, et al. 2011. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *J. Am. Med. Inform. Assoc.* 18(5):588–93

Miotto R, Weng C. 2015. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *J. Am. Med. Inform. Assoc.* 22(e1):e141-50

Mitchell AP, Hirsch BR, Abernethy AP. 2014. Lack of timely accrual information in oncology clinical trials: A cross-sectional analysis. *Trials*. 15:92

Mohd Sulaiman I, Karlsson D, Koch S. 2017. Mapping acute coronary syndrome registries to SNOMED CT. A comparative study between Malaysia and Sweden. *Methods Inf. Med.* 56(4):330–38

Monsen KA, Rudenick JM, Kapinos N, Warmbold K, McMahon SK, Schorr EN. 2018. Documentation of social determinants in electronic health records with and without standardized terminologies: a comparative study. *Proceedings of Singapore Healthcare*. 28(1):201010581878564

Morash M, Mitchell H, Yu A, Campion T, Beltran H, et al. 2018. CATCH-KB: Establishing a pharmacogenomics variant repository for chemotherapy-induced cardiotoxicity. *AMIA Jt Summits Transl Sci Proc*. 2017:168–77

Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. 2018. Classification of forensic autopsy reports through conceptual graph-based document representation model. *J. Biomed. Inform.* 82:88–105

Müller L, Gangadharaiah R, Klein SC, Perry J, Bernstein G, et al. 2019. An open access medical knowledge base for community driven diagnostic decision support system development. *BMC Med. Inform. Decis. Mak.* 19(1):93

Napolitano G, Marshall A, Hamilton P, Gavin AT. 2016. Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artif. Intell. Med.* 70:77–83

National Library of Medicine. 2016. *Overview of SNOMED CT*. www.nlm.nih.gov/healthit/snomedct/snomed_overview.html. Accessed 26 October 2020

National Library of Medicine. 2021. *UMLS Release File Archives.* https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html. Accessed 3 April 2022

Nguyen HA, Al-Mubaid H. 2006. New ontology-based semantic similarity measure for the biomedical domain. *2006 IEEE International Conference on Granular Computing*, pp. 623–28. IEEE

Ni Y, Bermudez M, Kennebeck S, Liddy-Hicks S, Dexheimer J. 2019. A real-time automated patient screening system for clinical trials eligibility in an emergency department: Design and evaluation. *JMIR Med. Inform.* 7(3):e14185

Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, et al. 2015. Increasing the efficiency of trial-patient matching: Automated clinical trial eligibility pre-screening for pediatric oncology patients. *BMC Med. Inform. Decis. Mak.* 15:28

Northcutt CG, Athalye A, Mueller J. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks . *arXiv.org*. 2103.14749v3

Norton S, Cordery DV, Abbenbroek BJ, Ryan AC, Muscatello DJ. 2016. Towards public health surveillance of intensive care services in NSW, Australia. *Public Health Res. Pract.* 26(3):

Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. 2019. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *J. Am. Med. Inform. Assoc.* 26(11):1247–54

Oluoch T, de Keizer N, Langat P, Alaska I, Ochieng K, et al. 2015. A structured approach to recording AIDS-defining illnesses in Kenya: a SNOMED CT based solution. *J. Biomed. Inform.* 56:387–94

Pacheco JA, Rasmussen LV, Kiefer RC, Campion TR, Speltz P. et al. 2018. A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments. *J. Am. Med. Inform. Assoc.* 25(11):1540-6

Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. *AMIA Annu. Symp. Proc.* 2010:572–76

Pandey A, MacNamara J, Sarma S, Velasco F, Kannan V, et al. 2019. Rapid-cycle implementation of a multi-organization registry for heart failure with preserved ejection fraction using health information exchange standards. *Stud. Health Technol. Inform.* 264:1560–61

Park HS, Cho H, Kim HS. 2016. Development of an integrated biospecimen database among the regional biobanks in Korea. *Healthc. Inform. Res.* 22(2):129–41

Patrick J, Wang Y, Budd P. 2007. An Automated system for conversion of clinical notes into SNOMED clinical terminology. *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers*. 68:219–226

Pearce C, McLeod A, Patrick J, Ferrigi J, Bainbridge MM, et al. 2019. Coding and classifying GP data: the POLAR project. *BMJ Health Care Inform.* 26(1):

Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *J. Biomed. Inform.* 40(3):288–99

Pedrera M, Serrano P, Terriza A, Cruz J, Varela C, et al. 2020. Defining a standardized information model for multi-source representation of breast cancer data. *Stud. Health Technol. Inform.* 270:1243–44

Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. 2012. Effort required in eligibility screening for clinical trials. *J. Oncol. Pract.* 8(6):365–70

Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. 2018. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Jt Summits Transl Sci Proc*. 2017:188–96

Peterson KJ, Liu H. 2020. Automating the transformation of free-text clinical problems into SNOMED CT expressions. *AMIA Jt Summits Transl Sci Proc*. 2020:497–506

Petrova A, Ma Y, Tsatsaronis G, Kissa M, Distel F, et al. 2015. Formalizing biomedical concepts from textual definitions. *J. Biomed. Semantics*. 6:22

Plaza L. 2014. Comparing different knowledge sources for the automatic summarization of biomedical literature. *J. Biomed. Inform.* 52:319–28

Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, et al. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J. Am. Med. Inform. Assoc.* 22(1):143–54

Przybyła P, Brockmeier AJ, Ananiadou S. 2019. Quantifying risk factors in medical reports with a context-aware linear model. *J. Am. Med. Inform. Assoc.* 26(6):537–46

Pundt H, Bishr Y. 2002. Domain ontologies for data sharing–an example from environmental monitoring using field GIS. *Comput. Geosci.* 28(1):95-102

Rasmussen LA, Jensen H, Virgilsen LF, Jensen JB, Vedsted P. 2018. A validated algorithm to identify recurrence of bladder cancer: a register-based study in Denmark. *Clin. Epidemiol.* 10:1755–63

Reátegui R, Ratté S. 2018. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Med. Inform. Decis. Mak.* 18(Suppl 3):74

Resnik P. 1995. Using information content to evaluate semantic similarity in a taxonomy . *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–53. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc

Rios AM. 2020. *pymetamap: Python wraper for MetaMap*. GitHub. https://github.com

Roldán-García MDM, García-Godoy MJ, Aldana-Montes JF. 2016. Dione: an OWL representation of ICD-10-CM for classifying patients' diseases. *J. Biomed. Semantics*. 7(1):62

Ruch P, Gobeill J, Lovis C, Geissbühler A. 2008. Automatic medical encoding with SNOMED categories. *BMC Med. Inform. Decis. Mak.* 8(suppl. 1):S6

Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. New York, NY, USA: ACM

Sanchez Bocanegra CL, Sevillano Ramos JL, Rizo C, Civit A, Fernandez-Luque L. 2017. HealthRecSys: A semantic content-based recommender system to complement health videos. *BMC Med. Inform. Decis. Mak.* 17(1):63

Sánchez D, Batet M, Isern D. 2011. Ontology-based information content computation. *Knowledge-Based Systems*. 24(2):297–303

Sánchez D, Batet M, Viejo A. 2014. Utility-preserving privacy protection of textual healthcare documents. *J. Biomed. Inform.* 52:189–98

Sánchez D, Batet M. 2011. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. *J. Biomed. Inform.* 44(5):749–59

Schroen AT, Petronic GR, Wang H, Gray R, Wang XF, et al. 2010. Preliminary evaluation of factors associated with premature trial closure and feasibility of accrual benchmarks in phase III oncology trials. *Clin. Trials.* 7(4): 312-21

Schulz S, Daumke P, Romacker M, López-García P. 2019. Representing oncology in datasets: standard or custom biomedical terminology? *Informatics in Medicine Unlocked*. 15:100186

Schulz S, Rodrigues J-M, Rector A, Chute CG. 2017. Interface terminologies, reference terminologies and aggregation terminologies: A strategy for better integration. *Stud. Health Technol. Inform.* 245:940–44

Segura-Bedmar I, Raez P. 2019. Cohort selection for clinical trials using deep learning models. *J. Am. Med. Inform. Assoc.* 26(11):1181–88

Shen F, Liu S, Wang Y, Wen A, Wang L, Liu H. 2018. Utilization of electronic medical records and biomedical literature to support the diagnosis of rare diseases using data fusion and collaborative filtering approaches. *JMIR Med. Inform.* 6(4):e11301

Shobhana B., Radhakrishnan R. 2016. Estimation of semantic similarity between concepts and fuzzy rules optimization with modified genetic algorithm (MGA). *IIOAB J.* 7(7):52–60

Siefridt C, Grosjean J, Lefebvre T, Rollin L, Darmoni S, Schuers M. 2020. Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations. *Int. J. Med. Inform.* 133:104009

Silva Layes E, Bondarenco M, Machiavello D, Frola F, Lemos M. 2019. Implementation of a terminology server with SNOMED CT in graph databases. *Stud. Health Technol. Inform.* 264:1584–85

Small AM, Kiss DH, Zlatsin Y, Birtwell DL, Williams H, et al. 2017. Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease. *J. Biomed. Inform.* 72:77–84

SNOMED CT International. 2021. *SNOMED - Why SNOMED CT?* www.snomed.org

Sohn S, Liu H. 2014. Analysis of medication and indication occurrences in clinical notes. *AMIA Annu. Symp. Proc.* 2014:1046-55

Sohn S, Wang Y, Wi C-I, Krusemark EA, Ryu E, et al. 2018. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *J. Am. Med. Inform. Assoc.* 25(3):353–59

Sohn S, Wu S, Chute CG. 2012. Dependency parser-based negation detection in clinical narratives. *AMIA Jt Summits Transl Sci Proc*. 2012:1–8

Song T-M, Park H-A, Jin D-L. 2014. Development of health information search engine based on metadata and ontology. *Healthc. Inform. Res.* 20(2):88–98

Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, et al. 2018. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J. Am. Med. Inform. Assoc.* 25(3):331–36

Spasic I, Krzeminski D, Corcoran P, Balinsky A. 2019. Cohort selection for clinical trials from longitudinal patient records: text mining approach. *JMIR Med. Inform.* 7(4):e15980

Spasic I, Nenadic G. 2020. Clinical text data in machine learning: Systematic review. *JMIR Med. Inform.* 8(3):e17984

Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. 2019. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *J. Am. Med. Inform. Assoc.* 26(11):1163–71

Stubbs A, Uzuner Ö. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *J. Biomed. Inform.* 58 Suppl:S20–29

Sun M, Zhu W, Tao S, Cui L, Zhang G-Q. 2015. COBE: a conjunctive ontology browser and explorer for visualizing SNOMED CT fragments. *AMIA Annu. Symp. Proc.* 2015:2092–2100

Tahmasebi AM, Zhu H, Mankovich G, Prinsen P, Klassen P, et al. 2019. Automatic normalization of anatomical phrases in radiology reports using unsupervised learning. *J. Digit. Imaging*. 32(1):6–18

Tannier X, Paris N, Cisneros H, Bozec CL, Doutreligne M, et al. 2019. Hybrid approaches for our participation to the n2c2 challenge on cohort selection for clinical trials. *arXiv*

Ternois I, Billard-Pomares T, Carbonelle E, Franchinard L, Duclos C. 2019. Using SNOMED-CT to help the transition from microbiological data to ICD-10 sepsis codes. *Stud. Health Technol. Inform.* 264:1604–5

Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. 2009. Electronic screening improves efficiency in clinical trial recruitment. *J. Am. Med. Inform. Assoc.* 16(6):869–73

Tomic K, Sandin F, Wigertz A, Robinson D, Lambe M, Stattin P. 2015. Evaluation of data quality in the National Prostate Cancer Register of Sweden. *Eur. J. Cancer*. 51(1):101–11

Treweek S, Lockhart P, Pitkethly M, Cook JA, Kjeldstrøm M, et al. 2013. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. *BMJ Open*. 3(2):

U.S. Food & Drug Administration. 2017. *Drugs@FDA Data Files* . www.fda.gov

Valtchinov VI, Lacson R, Wang A, Khorasani R. 2020. Comparing artificial intelligence approaches to retrieve clinical reports documenting implantable devices posing MRI safety risks. *J. Am. Coll. Radiol.* 17(2):272–79

Velupillai S, Skeppstedt M, Kvist M, Mowery D, Chapman BE, et al. 2014. Cue-based assertion classification for Swedish clinical text--developing a lexicon for pyConTextSwe. *Artif. Intell. Med.* 61(3):137–44

Vydiswaran VGV, Strayhorn A, Zhao X, Robinson P, Agarwal M, et al. 2019. Hybrid bag of approaches to characterize selection criteria for cohort identification. *J. Am. Med. Inform. Assoc.* 26(11):1172–80

Wang Y, Luo J, Hao S, Xu H, Shin AY, et al. 2015. NLP based congestive heart failure case finding: a prospective analysis on statewide electronic medical records. *Int. J. Med. Inform.* 84(12):1039–47

Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, et al. 2018. Clinical information extraction applications: A literature review. *J. Biomed. Inform.* 77:34–49

Weeber M, Klein H, Aronson AR, Mork JG, de Jong-van den Berg LT, Vos R. 2000. Text-based discovery in biomedicine: the architecture of the DAD-system. *Proc. AMIA Symp.* 903–7

Wei DH, Fu G. 2017. Using SNOMED distance to measure semantic similarity of clinical trials. *Stud. Health Technol. Inform.* 245:1341

Wongthongtham P, Zadjabbari B. 2012. Ontology based Knowledge Transferability and Complexity Measurement for Knowledge Sharing. *Proceedings of the International Conference on Knowledge Management and Information Sharing*, pp. 5–14. SciTePress - Science and and Technology Publications

Wu Y, Jiang M, Xu J, Zhi D, Xu H. 2017. Clinical named entity recognition using deep learning models. *AMIA Annu. Symp. Proc.* 2017:1812–19

Wu Z, Palmer M. 1994. Verbs semantics and lexical selection. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pp. 133–38. Morristown, NJ, USA: Association for Computational Linguistics

Xu H, Fu Z, Shah A, Chen Y, Peterson NB, et al. 2011. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu. Symp. Proc.* 2011:1564–72

Ye Y, Wagner MM, Cooper GF, Ferraro JP, Su H, et al. 2017. A study of the transferability of influenza case detection systems between two large healthcare systems. *PLoS ONE*. 12(4):e0174970

Zare M, Pahl C, Nilashi M, Salim N, Ibrahim O. 2015. A review of semantic similarity measures in biomedical domain using SNOMED-CT. *J Soft Comput Decis Support Syst*. 2(6):

Zolnoori M, Fung KW, Patrick TB, Fontelo P, Kharrazi H, et al. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. *J. Biomed. Inform.* 90:103091

Zvára K, Tomečková M, Peleška J, Svátek V, Zvárová J. 2017. Tool-supported interactive correction and semantic annotation of narrative clinical reports. *Methods Inf. Med.* 56(3):217-29.