A Corpus-Driven Standardization Framework for Encoding Clinical Problems with SNOMED CT Expressions and HL7 FHIR

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Kevin Peterson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Hongfang Liu, Yuk Sham

December, 2020

Acknowledgements

I thank Dr. Hongfang Liu for her guidance, support, and patience with me through the years. I could ask for no better example of what it means to be a scientist and a leader. It has been my greatest privilege and honor to study under the guidance of such a wonderful mentor.

To my full committee, Dr. Hongfang Liu, Dr. Serguei Pakhomov, Dr. Yuk Sham, Dr. Chih-Lin Chi, and Dr. Guoqian Jiang, a heartfelt thank you for your time and effort as I have navigated through this process. Your comments, suggestions, and feedback have not only helped to shape this dissertation, but have inspired several interesting ideas and topics for future studies.

I am grateful for the guidance of Dr. Yuk Sham, Dr. Chad Myers, and the entire BICB program leadership. Without the BICB program and its accommodations for working students, pursuing this degree would not have been possible for me – and I am forever grateful for the opportunity.

I thank my parents for providing me the foundation through which all this is possible. To my mother, thank you for teaching me to be kind and to listen, and to put other people first. Thank you for having patience when I took apart the radio, and the toaster, and the lawnmower ... it is that wonder about how the world works that has brought me to this dissertation. Thank you to my father, for showing me how to be self-confident yet humble, tough but kindhearted. I am quite aware that I have been afforded opportunities to do what I love only through the sacrifices of my parents.

Thank you to the Mayo Clinic for their support and investment in me, both as an employee and a student. They have stood by me at every turn, and have provided me every opportunity to be successful.

I'd like to thank my wife Laura for being there for me through this whole process. I hope that whenever you need it, I can be as supportive to you as you were to me. As with most other things in life, I wouldn't have been able to do this without you. To Valerie and Cameron, I hope that when you grow up you can find things in your life that you can be passionate about. And don't be afraid to take a leap of faith from time to time – you might accomplish more than you think.

In memory of Kimel L. Watt. I'm sorry you didn't get a chance to write a dissertation of your own.

Dedication

To Laura, Valerie, and Cameron

Abstract

Free-text clinical problem descriptions are used throughout the medical record to communicate patients' pertinent conditions. These summary-level representations of diagnoses and other clinical concerns underpin critical aspects of the modern patient record such as the problem list, and are key inputs to predictive models and clinical decision support applications. Given their importance to both clinical care and downstream analytics, representations of these clinical problems must be amenable to both human interpretation and machine processing. While free-text is expressive and provides the most transparent and unbiased view into the intent of the clinician, standardized and consistent representations of the semantics of these problem descriptions are necessary for contemporary data-driven healthcare systems.

Free-text problem descriptions may be standardized and structured in a variety of ways. First, they may be encoded using a controlled terminology such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT). Even though a single code may inadequately capture the context, modifiers, and related information of a problem, codes may be combined, or "post-coordinated" into more complex structures called SNOMED CT Expressions. Next, alignment to standardized semantic and data models such as Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) allows for the most structured representation, but with higher implementation complexity.

Competing usage priorities introduce a fundamental optimization problem in representing these entries – free-text is the most natural and useful form for clinicians, while structured and codified forms are computable and better suited for data analytics and interoperability. In this study, we introduce methods to minimize this conflict between structured and unstructured forms by proposing a framework for capturing the semantics of free-text clinical problems and transforming them into codified, structured formats using Natural Language Processing (NLP) techniques.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Healthcare data is inherently difficult to analyze and share at scale due to its heterogeneity, ambiguity, and the complexity of the domain itself.[1] While advances in informatics research and the contributions of standardizing bodies have driven change over time, organizations still struggle to reach a meaningful consensus on how to organize and standardize patient data.[2] While the potential and promise of large-scale data analytics and data-driven insights enabled by the modern electronic health record (EHR) is undeniable, access to large amounts of healthcare information does not imply data that is understandable or actionable. The practical challenges of managing and organizing the patient record have only become more pronounced as our ability to capture and store healthcare data expands.[3]

One significant organizational shift to the patient record came in the 1960s when Lawrence Weed proposed orienting the data around a list of each patient's current conditions, or the "problem list."[4] Entries into this problem list are designed to succinctly reflect the current clinical state of a patient, and practically function as an index into the larger clinical narrative.[5] This orientation around problems gave rise to the problem-oriented medical record (POMR),[4] a fundamental realignment of healthcare data with clinical problems as the focus.

This strategy did, necessarily, place a newfound emphasis on the structuring and encoding of the clinical problems themselves.[6] These problem list entries anchoring the POMR are descriptions of clinical diagnoses and other patient issues, and may be encoded in several different ways. First, clinicians often enter these problem descriptions as free-text due to the expressiveness inherent in natural language.[7] This unstructured form is prevalent and often preferred by clinicians, but is the most difficult to standardize and process.[8] Next, problems may be codified by selecting concepts from a controlled terminology or vocabulary – either by using institution-specific lists or an internationally recognized standard such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT).[9] While this provides a more structured representation, there are significant practical issues in narrowing problem choices to static code lists, and even greater questions regarding how this codification impacts clinicians' workflows.[10] Finally, data standards such as Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR)[11] provide models for healthcare data that include format, semantic constraints, and serializations, but can be difficult to adopt and implement due to the infrastructure and expertise needed to utilize them effectively.[12,13]

The purpose of this work is to introduce a framework to reconcile these different clinical problem representations. In the following chapters, we propose methods to transform between the three problem encodings described above using Natural Language Processing (NLP) methods, automating the process starting from free-text and culminating in standard HL7 FHIR representations. We also include in our study a focus on two peripheral issues regarding implementation: the organization of value sets and the tracking of clinical language change over time. We include these topics as an acknowledgment that the success of standardization is often determined by factors other than the quality of the models themselves – such as the cost and complexity of adoption and integration.[13]

1.1 Motivation

Clinical problem descriptions, as with many aspects of the patient record, are highly complex and variable, and frequently captured as unstructured text.[3,14] This not only makes them challenging to analyze and share, but also increases the complexity and cost of healthcare software systems that must interpret them.[15,16] The ability to unambiguously share healthcare data, or data *interoperability*, requires agreed-upon conventions, or *standardization*. Data standardization can be syntactic (data format and encoding) as well as semantic (data meaning and organization of concepts), and the standardization of clinical problem descriptions is the focus of this study.

Despite standardization efforts that have been ongoing for decades in healthcare, the goal of consistent data semantics has remained elusive.[14,17–19] In the case of problem lists (and problem descriptions in general), a major contributing factor is the fundamental tension between clinician-preferred free-text and computationally friendly structured representations. To bridge this gap, many healthcare standards organizations and professional associations maintain standard vocabularies and terminologies which define the important concepts of the healthcare domain and link them to standard terms and phrases.[20] While these techniques can be effective at normalizing clinical language, many problem descriptions are too expressive to be captured by a single standardized concept.[21] This implies that some high-level structuring of semantics is needed. While the FHIR specification and SNOMED CT post-coordinated expressions fill that role, no robust solutions exist to automatically transform free-text problem descriptions into these more standardized representations. It is this transformation from free-text to SNOMED CT expressions and FHIR Resources that forms the bulk of this study.

Standards by their nature tend to force some degree of conformance to an idealized model. As a consequence, a standards-based view of healthcare data semantics

is often disconnected from the realities of the actual data. While advances in data-driven techniques have shown promise for eliciting robust semantic models from existing EHR data,[3,22–24] existing healthcare data standards are generally dictated by large standards bodies. As such, mismatches in granularity, scope, and purpose may occur between the standards and real-world usage scenarios. Furthermore, there are few available techniques and tools to align or compare the data-derived models to externally-defined standards. Blueprints on how to align the "top-down" development approach employed by standards organizations with "bottom-up" data-driven approaches are sparsely described in literature.[25] The incorporation of data-driven approaches with domain-specific curated knowledge is an important facet of this study.

Standardization is the goal, but must be done pragmatically and with a focus on end users. Standards that are too complex, abstract, or ambiguous lead to implementation challenges and high costs – or the abandonment of standards altogether in favor of one-off approaches.[12,13] FHIR is a movement towards more pragmatic standards and leverages FHIR Profiles as a mechanism to define computable, unambiguous semantic contracts. Profiles are necessary because many of the base FHIR resources allow for significant variation in how information is structured. They also address the interoperability challenge of iso-semantic resource representations, or similar conceptual data represented differently,[26] by establishing unambiguous contracts for a specific use case. They do, however, have two important deficiencies: (1) they are generally created and curated by hand, and (2) there are no known computational methods to compare the associated value sets of Profiles – specifically, to organize them into hierarchies or similarity clusters. This makes FHIR Profiles difficult to maintain and catalog, and could limit their usability at scale. In this work, we introduce two organizational techniques specifically targeted toward managing FHIR Profile value sets.

Finally, we recognize that standardization efforts are not point-in-time events, but ongoing commitments to maintaining running systems and infrastructure over time. Given that our system is based on processing clinical *language*, we must therefore also

acknowledge that the characteristics of this language could change over time. Failing to detect and quantify this change would steadily degrade the performance of deployed NLP systems,[27,28] with future implementations of our framework being no exception. With this in mind, a robust study of how clinical text changes over time is included in this work.

1.2 Research Summary

In this study, we have developed methods to transform free-text clinical problems into three progressively standardized representations. Our approach focuses on a combination of data-driven techniques and external domain knowledge bases to produce a standardization framework for encoding clinical problems. As this is a corpus-driven framework, we derive our methods mainly from a large corpus of problem descriptions extracted from over 14 million clinical documents.[29] Using this corpus as our base, we present the following main research foci, and briefly summarize their corresponding experiments and findings below.

1.2.1 A corpus-driven framework to learn semantic patterns from clinical problem summaries.

We first introduce methods to discover what medical concepts are used in free-text problem descriptions, what modifiers the concepts have, and how these concepts are related. The emphasis is on capturing these expressive semantic patterns in a way that is amenable to computation and analysis. This approach is broken into three main parts: (1) learning semantic patterns based on Unified Medical Language System (UMLS)[30] and SNOMED CT concepts, (2) organizing these patterns into a Web Ontology Language (OWL)[31] and Resource Description Framework (RDF)[32] framework, and (3) using OWL inference and the SPARQL query langauge[33] to analyze our processed

corpus. Our methodology is heavily based on a combination of Open Information Extraction (OIE)[34] and dependency parsing techniques to build a full RDF representation of our corpus. The following experiments were conducted to examine our approach, with Chapter 2 devoted to describing these techniques in full.

- **Question:** What semantic patterns exist in summary-level clinical problem descriptions?
  - **Experiments:** Using methods to parse clinical text into clinical concepts in the form of *subject-predicate-object* triples, we processed our clinical corpus using OpenIE[35] into an OWL/RDF representation. We then used the RDF query language SPARQL to mine prominent semantic patterns from this large normalized knowledge base.
  - **Results:** We produced a set of frequent semantic patterns found in our data set, as well as developed a SPARQL framework to query our corpus for prominent modifiers and qualifiers that will be used in our next experiment.
- **Hypothesis:** Our mined semantic patterns will show a correlation to SNOMED CT CORE Problem List Subset pre-coordinated concepts.
  - **Experiments:** We compared attribute usage in pre-coordinated concepts from the SNOMED CT CORE Problem List Subset[36] with concepts mined from our RDF corpus. We used the SNOMED CT CORE Problem List Subset as the basis for comparison as this set of concepts represents an industry standard specifically designed to represent clinical problems.[37]
  - **Results:** We found a moderate correlation between the attribute usage of the SNOMED CT CORE Problem List Subset and our data-driven concepts. This indicates that the composition of concepts (along with their modifiers) included in the SNOMED CT CORE Problem List Subset is relatively congruent with our clinical corpus.

1.2.2 A pipeline to encode free-text clinical problem summaries using SNOMED CT Expressions.

Summary level problem descriptions often describe complex clinical conditions with important supporting context (such as severity/stage, body location, related or contributing conditions, and so on). Even with extensive clinical ontologies such as SNOMED CT, many problem descriptions are too expressive for the complete meaning to be captured using one standard concept. For these more complex problem descriptions, we transform free-text representations to SNOMED CT expressions, a compositional grammar for SNOMED CT concepts. We accomplish this by introducing a Bidirectional Long Short-Term Memory (BiLSTM)[38,39] deep learning classifier used to determine the relationship type that holds between source and target entities – an important facet of building SNOMED CT expressions. A series of experiments that test the performance of our automated process, described in Chapter 3, are summarized below.

- **Hypothesis 1:** A deep learning model will outperform a Naïve Bayes model for relation identification on SNOMED CT stated relationship test data.
  - **Experiments:** The BiLSTM model was compared against a baseline Naïve Bayes[40] model trained and tested on the asserted SNOMED CT Relationship data set. Data was prepared for this experiment by partitioning 25% of the SNOMED CT Relationship set for testing and 75% for training. The Naïve Bayes and the BiLSTM classifiers were both then trained and tested on the same data, with the exception of a further 20% of the BiLSTM training data being withheld for validation. For testing, the F_1 scores for each individual attribute as well as overall averages were recorded for both classifiers.
  - **Results:** We find that the Naïve Bayes model baseline was outperformed by the BiLSTM architecture. This provides evidence that a deep learning approach is viable for this task.

- **Hypothesis 2:** A deep learning model will outperform the state of the art (Kate's Support Vector Machine (SVM) model[41]) for relation identification on SNOMED CT stated relationship test data.

  - **Experiments:** We compared our results to previously reported results of Kate's SVM model.[41] We followed Kate's evaluation procedures in order to replicate his experiment using our BiLSTM model: For each attribute, 5000 relationships of the desired type were randomly selected from the SNOMED CT Relationship set along with an equal number of negative examples. Given this test set, the ability of the classifier to correctly determine whether or not the chosen relationship was present was recorded.
  - **Results:** Our model outperformed Kate's results in four of the five relationship types under test. This demonstrates that our deep learning approach not only outperforms a Naïve Bayes baseline (see previous experiment), but also a more sophisticated SVM model. A comparison to Kate's work is important to this study as it is the only known previously published results for this specific task.

- **Hypothesis 3:** Training on SNOMED CT stated relationships will allow the model to generalize to actual clinical free-text diagnosis statements.

  - **Experiments:** To evaluate model performance in real-world scenarios and test generalizability to different data sets, we utilized our clinical problem list corpus to create a human-annotated gold standard test set of clinical relationships. A random sample of problem descriptions were extracted for manual annotation with each entry annotated with the focal concept and all associated relationships. We then evaluated our BiLSM model using this test corpus.
  - **Results:** Our results showed a significant decrease in performance when

generalizing our model to real clinical text. This indicates that training on the SNOMED CT Relationship set alone is not sufficient to generalize to real-world clinical text. This deficiency is further addressed below (see Section 1.2.3, Hypothesis 1).

- **Hypothesis 4:** A dependency parse-based method is effective at locating the focus concept of the clinical problem.
  - **Experiments:** The "focus" concept of a problem is the primary clinical condition or issue to which the problem description is referring. We hypothesize here that the `ROOT` of the problem dependency parse will match what a human annotator would classify as the focus. To test this method, a set of clinical problem descriptions was randomly sampled from our corpus and manually annotated with their focus concept. Results were gathered using several dependency models, ranging from generic English to models specifically trained for the biomedical domain.
  - **Results:** We find that our dependency parsed-based method has good alignment with our gold standard annotated test set, with F_1 scores > 0.9. We also find that dependency parse models based on generic English perform poorly compared to those tailored to the biomedical domain.

1.2.3 The standardization of free-text clinical problem summaries using the HL7 FHIR Specification.

HL7 FHIR is an emerging standard for healthcare data, providing both a model for organizing data (*Resources*) and a mechanism for modeling their semantic constraints (*Profiles*). We begin by transforming the free-text representations into FHIR `Condition` resources using an expanded form of the deep learning methods used to build SNOMED CT Expressions (see above). Next, we addressed an implementation challenge in the standard: the lack of a built-in Profile value set organizational model. We applied

a novel organizational scheme to FHIR Profile value sets that allowed us to discover, manage, and group them based on hierarchies and similarity clusters.

To examine our framework's ability to align free text to FHIR `Condition` resources, as well as our ability to address the implementation concern above, the following series of experiments were conducted. Details pertaining to this approach can be found in Chapters 4 & 5.

- **Hypothesis 1:** A hybrid rule-based/deep learning approach to generating FHIR `Conditions` from free-text problem descriptions will generalize better than a rule-based approach alone.
  - **Experiments:** The purpose of this experiment was to address the lack of generalizability found in our previous methods used for the SNOMED CT expression processing (Section 1.2.2). To create a more generalizable model, our existing deep learning model was augmented with data programming,[42] a weak-labeling system. Data programming allowed us to incorporate existing rules and knowledge bases into our training pipeline, allowing us to create a model better aligned to actual clinical text while avoiding the need to manually annotate training data. To test the effectiveness of data programming, we again compared against our gold standard annotated corpus, where six hundred random problem descriptions were extracted for manual annotation. Attributes of the FHIR `Condition` resource were annotated to create the test set. For completeness, other attributes outside of the base FHIR specification were also annotated including attributes from common FHIR extensions,[†] the Clinical Element Model (CEM) - *ClinicalAssert* model,[43] and the openEHR - *Problem/Diagnosis* archetype.[44] For the data programming approach to be effective, it needs to show improvement over a rule-based solution alone. In this experiment we first tested the rule-based approach, and then the full

[†]FHIR extensions: https://www.hl7.org/fhir/condition-extensions.html

data programming pipeline, and compared the results.

  - **Results:** For all tested attributes relevant to creating FHIR `Condition` resources, the data-programming approach outperformed the rule-based system alone. This indicates that a deep learning framework built on top of a rule-based foundation (enabled via data-programming) using no manually-annotated training data can be an effective approach for this task.

- **Hypothesis 2.** Value sets found in the extracted FHIR Profiles will self-organize into multi-level containment hierarchies.

  - **Experiments:** In this experiment we intend to verify that the extracted FHIR Profile value sets do in fact organize into containment hierarchies. We begin by running the containment organization algorithm on all extracted value sets. Once complete, we counted all value sets that fit into some hierarchy (in other words, any value set whose codes were are not a strict subset of at least one other value set), as well as the average number of levels in the hierarchy. We hypothesize that this structuring will produce some form of a hierarchy. The null hypothesis is that no hierarchical relationships are found, meaning there is no practical information gained via this structuring. We ran this experiment on both the FHIR Profile value sets and value sets extracted from the Value Set Authority Center (VSAC),[45] a large public repository of curated value sets.

  - **Results:** Hierarchies for both the FHIR Profile and VSAC value sets were detected via our methods, at times multiple levels deep. Although hierarchy was observed for FHIR Profile value sets, it was less than what was observed for the VSAC value sets, the other value set corpus studied in our experiment.

- **Hypothesis 3.** Value sets found in the extracted FHIR Profiles will organize into clusters (a clustering coefficient $>$ than that of a random graph).

- **Experiments:** This experiment explores whether or not extracted FHIR Profile value sets naturally organize into similarity clusters. To test this, first the clustering algorithm was run over the entire value set graph. Next, an Erdős-Rényi random graph[46] was created based on the characteristics of the value set graph (i.e. similar number of nodes and edges). After the random graph has been constructed, the same clustering algorithm was applied to find the clustering coefficient. As with the previous experiment, we also applied these methods to the VSAC value sets.
- **Results:** Our experiment showed that the extracted value sets exhibited higher clustering coefficients than would be expected from a similar random graph. Similar groupings were observed for the VSAC value sets as well.

1.2.4 Quantifing diachronic change of clinical problem summary language.

Linguistic patterns and word meanings are subject to change over time.[47] This can be in response to societal and cultural factors, or stem from new technology advances and knowledge within a specific domain. The biomedical domain is not exempt from this type of change,[48] and accounting for this variance is an important aspect of operating a clinical NLP system.

We describe four different methods to quantify diachronic change of clinical problem descriptions. Our methods are designed to explore different aspects of change, including grammatical/lexical differences, word sense change, variation in relation to different versions of standardized domain vocabularies, and changes to pragmatic diagnosis interpretation. Chapter 6 details the following experiments used to explore linguistic change in our clinical corpus.

- **Hypothesis 1:** Language model perplexity will increase as the number of years between training and testing data increases.

  - **Experiments:** Language models are a way to represent linguistic patterns as statistical models. Language models can be compared using *perplexity*, or a measure of how well a model predicts some unseen corpus.[49,50] In this experiment, we segment our clinical corpus by year, and train a language model for each year. For each model, we then measure perplexity for text of each subsequent year. We expect a language model to have increasing difficulty (or, greater perplexity) as the number of years between the train and test text increases.
  - **Results:** We show that the year distance between trained language model and test data is positively correlated with perplexity, supporting our hypothesis. This means that a model trained on clinical text one year will have an increasingly difficult time predicting text of future years, implying change over time.

- **Hypothesis 2:** Semantic changes in words can be detected through word embedding changes over time.
  - **Experiments:** To test if word embeddings can detect diachronic change, we trained a word2vec[51] embedding model for each year of our clinical corpus. We then compared how these word embeddings vary over time for a selected subset of words.
  - **Results:** We find that word embeddings can be effective for detecting shifting meaning for words. We show that certain words subject to drastic meaning change due to changes in technology or medical practice show correspondingly large changes in their embedding vector representation.

- **Hypothesis 3:** The same diagnosis text string, when codified using two different UMLS distributions, will result in progressively different codings as the years between the UMLS distributions increase.

  - **Experiments:** We installed sixteen previous years of the Unified Medical Language System (UMLS),[30] and codified a random subset of clinical problems from our corpus with each version using QuickUMLS,[52] a biomedical concept extraction tool based on the UMLS. We then compared how differently the same text was codified over all of the years.
  - **Results:** Our experiment showed that given a free-text clinical problem description, its UMLS codification differences are in fact correlated with how far apart the UMLS distributions are in years. From a practical perspective, this can be used to quantify the semantic "cost" of upgrading to a new UMLS version, given the number of years since the last upgrade.

- **Hypothesis 4:** Even if the summary level diagnosis linguistic patterns remain stable over time, how the condition is described in the clinical note narrative will progressively evolve.
  - **Experiments:** As problem descriptions are meant to be succinct summarizations of the clinical problems, they are in effect proxies for the clinician's full interpretation treatment plan for the disease. As such, we suspect that it is possible that even if the problem description remains the same, its interpretation in the full clinical narrative may change over time, reflecting changing medical practices. For this experiment, we selected the top ten most frequently occurring diagnosis strings, and gathered their attendant Impression, Report and Plan (IRP) sections, segmented by year, from the clinical narrative. We then computed Term Frequency Inverse Document Frequency (TF-IDF) differences between each year to quantify changing clinical *intent* even given unchanged diagnosis strings.
  - **Results:** We show that for our selected diagnoses, the amount of TF-IDF change is correlated to the number of years between text. This indicates that even if clinical problem descriptions remain relatively static over time, they

may be being interpreted differently by clinicians.

1.3 Research Contributions

The following list summarizes our contributions to this research area:

- Using a data-driven approach to mine semantic patterns increases efficiency over what is often a manual process.[53] We contribute to this area of research by expanding on the *frame detection* NLP task.[54] Specifically, we expand on the work of the FRED project[55] by (1) incorporating external biomedical ontologies for reasoning and inference, (2) integrating healthcare-specific entity detection, (3) implementing a custom relation extraction mechanism, and (4) introducing a SPARQL[33]-based framework for mining prominent patterns.

- Previously, the automated conversion of summary-level problem description text to SNOMED CT expressions remained largely unexplored. Existing approaches[41] had focused on only portions of the task and did not incorporate recent advances in deep learning. To our knowledge, this effort was the first to address the full range of tasks required to transform free text into SNOMED CT expressions.

- Although research into the transformation of the broad clinical narrative into FHIR is ongoing via the NLP2FHIR project,[56] our approach is focused on problem encoding (via FHIR `Condition` resources) and could enhance that effort by providing a more robust capture of modifiers vs. the ConText algorithm[57] used by NLP2FHIR.

- There are no known ways to organize the value sets used by FHIR Profiles to reduce management burden. Existing approaches require manual metadata curation or indexing by subject matter experts. This work represents the first attempt to overlay organizational structure over FHIR Profile value sets through unsupervised algorithms.

- While diachronic linguistic change has been studied extensively, this is the first effort to our knowledge that examines the change of clinical problem descriptions over time. We also detail two novel change metrics specific to clinical problem descriptions – change in UMLS codifications over time and pragmatic change relative to the clinical narrative. We believe these are important considerations for a full accounting of problem list change.

1.4 Outline

A general outline of the ensuing chapters is shown below:

- Chapter 2 outlines techniques for analyzing the semantic patterns of clinical problem descriptions.
- Chapter 3 describes methods to convert free-text problem descriptions into SNOMED CT expressions, the next tier of standardization.
- Chapter 4 extends the standardization approach of Chapter 3 to the HL7 FHIR specification.
- Chapter 5 introduces two new organizational methods for value sets – important semantic artifacts used in FHIR Profiles to scope the semantics of FHIR Resources.
- Chapter 6 explores an important aspect of free-text problem descriptions – detecting how they linguistically evolve over time.
- Chapter 7 finalizes the discussion of our framework and proposes some future research and implementation directions.

Note that where indicated, previously published material has been incorporated into this dissertation, used with permission. For these chapters, Kevin Peterson's CRediT roles include Conceptualization, Formal Analysis, Investigation, Methodology, Software, Writing - Original Draft, and Writing - Review & Editing.

Chapter 2

The Sublanguage of Clinical Problem Lists: A Corpus Analysis

Kevin J Peterson and Hongfang Liu. The sublanguage of clinical problem lists: a corpus analysis. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1451. American Medical Informatics Association, 2018

Abstract

Summary-level clinical text is an important part of the overall clinical record as it provides a condensed and efficient view into the issues pertinent to the patient, or their "problem list." These problem lists contain a wealth of information pertaining to the patient's history as well as current state and well-being. In this study, we explore the structure of these problem list entries both grammatically and semantically in an attempt

to learn the specialized rules, or "sublanguage" that governs them. Our methods focus on a large-scale corpus analysis of problem list entries. Using Resource Description Framework (RDF), we incorporate inferencing and reasoning via domain-specific ontologies into our analysis to elicit common semantic patterns. We also explore how these methods can be applied dynamically to learn specific sublanguage features of interest for a particular concept or topic within the domain.

2.1 Introduction

The amount of data being produced by the healthcare industry is ever-increasing, with much of that data being in the form of unstructured or semi-structured clinical notes.[3] These notes serve an important role in capturing the overall context and content of the patient-provider encounter, and give the providers a mechanism to capture information with an expressivity and flexibility that structured forms may not.[7] Consequently, clinical notes often contain information that may not be found elsewhere in more structured parts of the medical record.[24]

Unstructured clinical notes are often accompanied by summaries of pertinent information – or the "problem list." This high-level synopsis of the patient's pertinent data is akin to a "table of contents" for the entire patient's record.[4] These relatively short, information-rich summaries of the patient state are the focus of our study, and extracting information from them using Natural Language Processing (NLP) techniques presents some interesting challenges.

Problem list data in the medical record comes in various forms – all with distinct processing challenges. Narrative-based summary level clinical data is common, but is often terse and relies heavily on the semantics of the included terms and their mappings to conceptual types or categories to convey information.[59] This alignment of the lexical representation to domain-specific categorizations is an important NLP technique

referred to as building the *semantic lexicon.*[60] Simply assigning semantic types to textual phrases is not sufficient to capture the full meaning of the data, however. While generally concise, the phrases in summary level data often include nuanced and important semantic relationships. That is why flat lists of codes, another common problem list information structure, often incur semantic loss[61] and may place the codification burden on providers.[62] More formal problem list representations such the emerging Fast Healthcare Interoperability Resources (FHIR)[11] standard show promise in terms of interoperability, but come with their own sets of semantics and structure that must be aligned to the existing data representations.

In terms of NLP, it remains a challenge in practice to bind healthcare semantics to a uniform lexicon.[63] Lexical variations for similar semantic concepts are prevalent and degrade the ability to uniformly process data from disparate sources. At a semantic level, however, the data often appears much more homogeneous. In fact, Sohn et al. found in their recent comparison of clinical notes across institutions that while the lexicon showed considerable variability, the overarching semantics were very similar at a conceptual level.[64] Defining this semantic layer for a domain of interest – the concepts, modifiers, and relationships – is an important step in formalizing the language.[65]

In their study, Liu et al. provided a comprehensive analysis of the semantic characteristics of summary-level data in a large clinical corpus.[29] Our goal for this work is to extend that effort and build on its findings by examining the lexical and semantic constraints of summary-level problem list data. By conducting a data-driven analysis, we aim to learn these domain-specific language constraints, or the *sublanguage*[66] of the domain.

2.2 Background & Significance

Zellig Harris in his theory of sublanguages posited that specialized domains use specialized language, the rules and constraints of which can be ascertained via analysis of a

representative corpus.[66] The work of Friedman has helped to extend the sublanguage theory into the biomedical field, and provides a thorough explanation of how this theory can be applied to the sublanguages of the biomolecular and clinical domains,[53] with the latter being our focus for this study. Below is a brief synopsis of the high-level tenets of Harris' sublanguage theory and how they scope this work.

- **Dependency relations.** Given all the words that make up a language, some words may depend on a subset or category of other words for their meaning. This brings an ordering and mathematical structure to a language and how it conveys information.[67] This is relevant to our study as problem list entries often have clinically important modifiers that scope a primary topic. For example, in the phrase "The patient has increased pain," 'patient' and 'pain' have no dependencies, but the adjective "increasing" depends on the noun "pain."

- **Paraphrastic reductions.** Sentences may be restructured many times while still conveying the same information, and may undergo several transformations for the sake of efficiency or brevity. Since we are analyzing summary-level text that has been derived from a larger body of notes, we note that extensive paraphrastic reductions have already taken place in order for the problem list entry to be available for our analysis.

- **Inequalities of likelihood.** The dependency relations mentioned above tend to be more constrained in sublanguages than in general language. Using this property, we can meaningfully categorize modifiers that are appropriate given a certain context – for example "The patient has (**type I** or **type II** or **gestational**) diabetes" would be an appropriate set of modifiers for diabetes in this context, while "The patient has (**macular** or **cerebral** or **pulmonary**) diabetes" would not. Extracting these modifier sets is an important focus of this study.

- **Sublanguage grammar.** As the domain of interest becomes more specialized,

> the constraints around words and patterns used to describe it become more pronounced. The key to understanding a sublanguage is to understand the word classes (the semantic lexicon) and the relationships that allow the sublanguage to efficiently convey the information of the domain.

But why is learning the characteristics of a sublanguage important? If the language of a domain varies more lexically as opposed to semantically,[64] capturing and modeling its semantic rules and constraints could allow for a more appropriate layer of abstraction on which downstream tooling can be built. Ideally, NLP tools could be architected in such a way that processing machinery (software) and the knowledge artifacts (sublanguage models) could both be separable, sharable, and reusable, and information extracted from these domains could be more readily mapped to iso-semantic formats and representations.[68]

In this work, we explore methods to elicit these sublanguage rules from a large clinical corpus using Resource Description Framework (RDF).[32] We begin by parsing summary-level clinical phrases into RDF *subject-predicate-object* triples. Next we attach semantics to these triples by linking the lexical phrases and/or terms to controlled vocabularies. We then apply domain knowledge in the form of ontologies to incorporate reasoning and inference into our analysis. Our methods aim to extract several pertinent aspects of sublanguage theory from the corpus, such as prevalent semantic rules and patterns, allowable modifiers for given concepts, and the binding of concepts to their lexical variations. This analysis is not static, however – we also propose a mechanism to allow domain experts to inspect the corpus for sublanguage characteristics dynamically. Examples of this could include searching for frequent semantic patterns related to a disease or diagnosis, extracting a set of allowable values for a data element of a proposed data model, or augmenting existing pattern-matching or rule-based NLP systems with further lexical variants for a given concept.

2.3 Related Work

The Linguistic String Project[69] and the Medical Language Extraction and Encoding System (MedLEE)[70] are notable implementations of Harris' theories applied to the clinical domain, and have proven successful in several applications.[71,72] MedLEE, along with the FrameNet project,[73] accomplishes this by binding the grammatical structure of the clinical phrases to defined patterns or *frames* where the data fits into relationships called *slots*.[74] These frames allow not only primary topics to be extracted, such as problems or findings, but their modifiers as well. In general, the notion of linguistic forms and semantic structures being separate (albeit related) layers of abstraction over clinical text[75] is an important point of emphasis for this work.

As an information storage medium, RDF has been leveraged by the NLP community in several different ways. Because of its flexibility, RDF has been used to promote tool interoperability[76] and as a schema for representing the information output by NLP systems.[77] Unstructured text to RDF parsers also have seen increasing use in question answering systems,[78,79] with implementations such as K-Extractor,[80] MEANS,[81] and BmQGen.[82]

Finally, in this study we build primarily on the previous work of Liu et al. which explored the alignment of summary level problem list data with standard medical vocabularies.[29] That study observed a rich set of semantic underpinnings present in problem list entries, and we extend that work by exploring in greater detail the sublanguage characteristics to which the domain conforms.

2.4 Methods

Our methods center around a data-driven analysis of a large corpus of problem list entries extracted from clinical notes. The data set used was the same Mayo Clinic corpus examined by Liu et al.[29] Derived from over 14 million clinical documents, the set contained 35,962,088 problem list entries, 9,157,136 of which were distinct. The

entries themselves on average contained eight words with a mean sentence length of 59 characters (including spaces).

2.4.1 Preprocessing and Parsing

The Stanford CoreNLP natural language processing toolkit[83] was used throughout the preprocessing and parsing phase. First, Stanford CoreNLP's built-in tokenizer and lemmatizer were used in the preprocessing step. After preprocessing, relationships were extracted based on related clauses in the sentences. This technique is called Open Information Extraction (OIE) and is used to extract relationship triples from a corpus without the need for domain-specific knowledge or training.[84] Stanford OpenIE,[35] an implementation of OIE, was used to elicit *subject-predicate-object* triples that denote relationships within the text. An example problem list entry is shown in Table 2.1 with its resultant parsing into triples.

Table 2.1: Parsing *subject-predicate-object* triples from the problem list phrase: "Patient has well-controlled chronic hypertension."

Subject	Predicate	Object
patient	has	well-controlled chronic hypertension
patient	has	well-controlled hypertension
patient	has	chronic hypertension
patient	has	hypertension
hypertension	is	well-controlled
hypertension	is	chronic

2.4.2 Concept Detection, Composition, and Standardization

Each subject and object was processed using MetaMap,[85] a National Library of Medicine (NLM) suite of tools designed to normalize free-text clinical terms to Unified Medical Language System (UMLS)[30] concepts. MetaMap was used in an attempt to map each phrase to a UMLS concept, thus establishing our link between the lexical representations

and the semantics. Table 2.2 shows the detected concepts that have been attached to the triples from Table 2.1.

Table 2.2: UMLS concept detection of the subject and object components of the RDF triple using MetaMap.

Subject Concept	Subject	Predicate	Object	Object Concept
C0030705	patient	has	well-controlled chronic hypertension	None
C0030705	patient	has	well-controlled hypertension	None
C0030705	patient	has	chronic hypertension	C0745114
C0030705	patient	has	hypertension	C0020538
C0020538	hypertension	is	well-controlled	C3853142
C0020538	hypertension	is	chronic	C0205191

Many phrases, however, cannot be semantically encompassed by a single UMLS concept. For example, the phrases *well-controlled chronic hypertension* and *well-controlled hypertension* in Table 2.2 may only be fully represented by multiple UMLS concepts composed together. We handle these compositional phrases by relating them hierarchically to phrases that have been normalized to a UMLS concept. To accomplish this, first we process all of the OpenIE-derived phrases for a given problem list entry using the Standford Dependency Parser.[86] Next, we compare the dependency tree for each phrase not successfully mapped using MetaMap to ones that have been codified. We define a *subClassOf* relationship only if the following conditions hold: (1) the root nouns for the two phrases are the same, and (2) the potential subclass contains a proper superset of any words attached via `mod - modifier` dependencies to the parent's root noun.* For example, Figure 2.1 shows a scenario where *well-controlled chronic hypertension* is considered a subclass of *chronic hypertension* because they share the same root noun (*hypertension*) and the set of adjectives for the child is a proper superset of the parent, or {'well-controlled', 'chronic'} $\supset$ {'chronic'}. Conversely, *well-controlled hypertension* is not considered a subclass of *chronic hypertension* because although they share the

*See the full list of `mod - modifier` dependency types: https://nlp.stanford.edu/software/dependencies_manual.pdf

same root noun, the child adjectives are not a proper superset: {'well-controlled'} $\not\supset$ {'chronic'}.

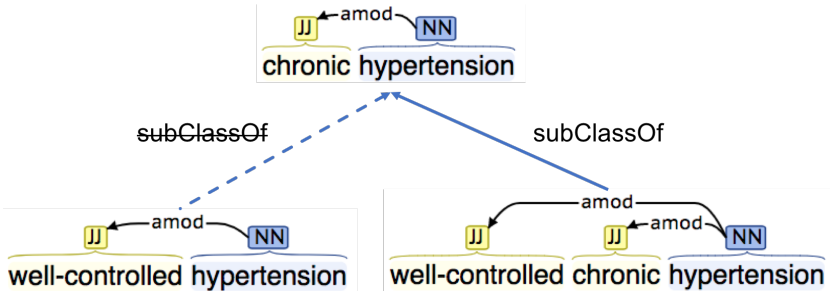


Figure 2.1: Asserting hierarchical relationships between phrases. Here, we attempt to infer the semantics of complex phrases in terms of simpler ones by comparing modifiers of a root noun. Hierarchical relationships can only be assumed if the modifiers of the child are a proper superset of the parent.

Along with the UMLS concepts, we also standardized to Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT),[87] an industry-standard vocabulary for clinical data. This was done by inspecting the UMLS Metathesaurus for the SNOMED CT concept(s) included in the detected UMLS concept. SNOMED CT as a standardization target was chosen for three main reasons: (1) Liu found promising overlap between SNOMED CT and our target corpus,[29] (2) it has precedent for being used to encode problem lists,[21,88] and (3) SNOMED CT is formally described as an ontology and may be utilized for reasoning and inference, as will be described further in the following section.

2.4.3 Reasoning and Inference

In the biomedical field, data-driven approaches provide the promise of less manual intervention and curation. In practice, however, curated domain knowledge artifacts are important parts of the data-driven approach as they make the data more computable and actionable.[89] An ontology, or the formal representation of a domain's concepts and relationships, is one such knowledge artifact. Since the data analysis of our study centers

around RDF, Web Ontology Language (OWL)[31] ontologies are readily incorporated into our methods. Fortunately, the biomedical community has been active in producing these knowledge artifacts. Two OWL ontologies relevant to our analysis are described below.

SNOMED CT encodes a wealth of domain-specific knowledge – such as multiple hierarchies and context-dependent concepts – that can be effectively formalized using OWL.[90] The SNOMED CT OWL ontology used in this study was included in the July 2017 SNOMED CT release.

The UMLS Semantic Network[91] defines a robust set of semantic categories, or *semantic types*, that are used to classify UMLS concepts. The Semantic Types Ontology is an OWL representation of the UMLS Semantic Network and can be obtained through the National Center for Biomedical Ontology (NCBO) BioPortal[92] site. This ontology arranges the individual semantic types into a hierarchy that can be then leveraged for reasoning-based queries.

2.4.4 RDF Analysis Techniques

With the RDF data generated, our analysis was focused in three main directions. First general corpus semantic characteristics were gathered, including concept mention frequencies and co-occurrence rates. Next, our analysis broadened beyond co-occurrences to a deeper exploration of the concepts of the domain and their relationships. This was accomplished by searching for frequently occurring patterns of RDF triples. From a frame semantics point of view, this technique is analogous to finding the most probable combination of slot types in a frame. We repeated this process twice, once to elicit representative frames based on SNOMED CT and once for the UMLS Semantic Types. Finally, we applied our methods in reverse – given a semantic frame of interest, what meaningful data can the corpus provide? For this analysis, we extracted both the sets of allowable modifiers and lexical variants for a frame. All analysis was conducted via SPARQL[33] queries into the RDF triple store.

2.4.5 Comparison to SNOMED CT CORE Problem List Subset

The next step in our analysis is a comparison of the extracted semantic patterns to an existing standard – specifically, the SNOMED CT Clinical Observations Recordings and Encoding (CORE) Problem List Subset.[36] The CORE Problem List Subset is an aggregation of commonly used SNOMED CT problem terms developed in collaboration with several large medical institutions.[37] We use this subset to compare the extracted semantic patterns from our corpus, and leverage the analysis to calculate how closely our corpus aligns to an industry standard.

For this analysis, we aim to compare differences in how attribute values are specified in SNOMED CT compared to how they are found in the data, where "attribute value" is the SNOMED CT term for a modifying or qualifying concept that changes the semantics of the main concept. For example, the concept 281000119103|*Severe recurrent major depression*| has the following two attribute/value pairs:

263502005|*Clinical course (attribute)*| → 255227004|*Recurrent (qualifier value)*|

246112005|*Severity (attribute)*| → 24484000|*Severe (qualifier value)*|

In this example, 255227004|*Recurrent (qualifier value)*| and 24484000|*Severe (severity modifier) (qualifier value)*| are the attribute values, or concepts that qualify the main concept. By comparing how these attribute values are used in SNOMED CT, we can compare general usage patterns in the data-driven patterns vs. SNOMED CT CORE Subset concepts. This can be used to show if there are discrepancies between the modifier patterns used to model SNOMED CT vs. what is actually found in the data.

To execute this comparision, we first examined every concept in the SNOMED CT CORE Subset and extracted a distribution of its attribute usage. This distribution will represent the pre-coordination (or semantic pattern) tendencies of the SNOMED CT CORE Subset. We then compared this to the attribute usage distribution from our data-driven analysis. We hypothesize that the attribute usage will be similar between the two. If this is true, it would give us evidence that the concepts in the SNOMED CT

CORE Subset have similar modifier usage patterns compared to what was found in the data. If the distributions are drastically different, we may suspect that in real-world scenarios, clinical problems are recorded with modifiers not included by SNOMED CT's pre-coordinated concepts, suggesting a mismatch.

To capture attribute usage patterns, we created a SPARQL query to extract the modifying concpets for all clinical problems in our RDF corpus. We modeled the usage distribution of these modifying attributes at two levels of granularity – first, we created a distribution over all SNOMED CT attributes (or, the 6482 descendants of `362981000`|*Qualifier value (qualifier value)*|). Next, we created a less-granular distribution, rolling up all qualifiers to the seventy immediate children of `362981000`.

2.5 Results

In total, 356,018,528 RDF triples were loaded into an OpenLink Virtuoso triple store.[93] Overall, 47.7% of the parsed RDF *entities* (subjects or objects of the triples) were able to be mapped to one or more SNOMED CT concepts. Table 2.3a shows the most commonly found SNOMED CT concepts, and Table 2.3b the most common predicates by which these concepts are related.

Table 2.3: The most frequent SNOMED CT concepts and RDF predicates.

(a) SNOMED CT concepts

#	Concept	Label
1916823	`392521001`	History of
1278195	`288563008`	After values
1276184	`237679004`	Status post
1276184	`255234002`	After
1260033	`38341003`	Hypertensive disorder
1089666	`90734009`	Chronic
1000691	`64572001`	Disease
869678	`22253000`	Pain
831439	`24028007`	Right
815479	`264180000`	Right sided

(b) RDF predicates

#	Predicate
6.50618×10^7	has
2.49417×10^7	is
3.34899×10^6	isWith
2.52518×10^6	isOf
1.84909×10^6	hasLeft
633927	isOn
628796	left
620415	isTo
590532	isFor
565410	isIn

We next examined the frequency at which any two SNOMED CT codes would appear in the same problem list entry. The ten most common SNOMED CT concept co-occurrences are shown in Table 2.4. We observe co-occurrences that appear semantically related, for example 22253000|*Pain*| and 90734009|*Chronic*|, an observation in line with what Liu observed.[29] Note that this metric analyzes concept co-occurrences within the context of an entire single problem list entry. Co-occurrences that occur within the context of a grammatical or semantic structure (such as the subject and object of a triple linked by a predicate) are discussed further below.

Table 2.4: SNOMED CT concept co-occurrences

#	Concept 1 (Label)	Concept 2 (Label)
334073	39823006 Generalized atherosclerosis	80891009 Heart structure
334072	359557001 Disorder of artery	80891009 Heart structure
235629	40593004 Fibrillation	59652004 Atrial structure
172240	22253000 Pain	113345001 Abdominal structure
172240	22253000 Pain	277112006 Abdominal
163742	73211009 Diabetes mellitus	258195006 Type 2
140399	87828008 Insufficiency	64033007 Kidney structure
135360	258158006 Sleep, function	263821009 Obstructed
120383	22253000 Pain	90734009 Chronic
113919	64572001 Disease	33359002 Degeneration

Tables 2.5 and 2.6 highlight our results from the analysis of semantic patterns, or frames, for the corpus. Generally, frames can be thought of as patterns of variable attributes that fit in predictable ways around a topic or concept.[74] As Lassila observed, this notion is very much analogous to RDF and its system of concepts and links between them.[94] As such, by examining the RDF association patterns, we aim to gain insight into the prominent semantic frames of our domain.

Table 2.5 depicts the most commonly found SNOMED CT concept patterns in which a focal concept is modified by two other concepts. By examining the prominent concepts and their relationships, we begin to see the common semantic patterns in the data. Table 2.6 shows results from the same analysis technique using UMLS Semantic Type

concepts instead of SNOMED CT. The UMLS Semantic Types allow for courser-grained classification and can be useful for deriving high-level relationship patterns.

Table 2.5: Frequent focal concept/two modifier patterns (SNOMED CT)

#	Focus Concept (Label)	Modifier 1 (Label)	Modifier 2 (Label)
105141	64572001 Disease	107669003 Degenerative abnormality	39352004 Joint structure
105141	64572001 Disease	33359002 Degeneration	39352004 Joint structure
105137	64572001 Disease	33359002 Degeneration	81087007 Articular sys. [...][1]
105137	64572001 Disease	107669003 Degenerative abnormality	81087007 Articular sys. [...][1]
105132	64572001 Disease	107669003 Degenerative abnormality	302536002 Entire joint
105132	64572001 Disease	33359002 Degeneration	302536002 Entire joint
56718	87828008 Insufficiency	90734009 Chronic	64033007 Kidney structure
55235	64572001 Disease	64033007 Kidney structure	42796001 End-stage
47047	40593004 Fibrillation	26593000 Paroxysmal	59652004 Atrial structure
38620	56246009 Hypertrophy	279689003 Prostatic gland structure	30807003 Benign

Full Label: [1] Articular system structure

Table 2.6: Frequent focal concept/two modifier patterns (Semantic Type)

#	Focus Concept (Label)	Modifier 1 (Label)	Modifier 2 (Label)
231728	T061 Therapeutic or [...][1]	T082 Spatial Concept	T079 Temporal Concept
167646	T023 Body Part, Organ, [...][2]	T082 Spatial Concept	T082 Spatial Concept
160216	T047 Disease or Syndrome	T023 Body Part, Organ, [...][2]	T079 Temporal Concept
146614	T169 Functional Concept	T079 Temporal Concept	T023 Body Part, Organ, [...][2]
137935	T047 Disease or Syndrome	T080 Qualitative Concept	T023 Body Part, Organ, [...][2]
134090	T061 Therapeutic or [...][1]	T082 Spatial Concept	T080 Qualitative Concept
129202	T061 Therapeutic or [...][1]	T079 Temporal Concept	T080 Qualitative Concept
122552	T047 Disease or Syndrome	T080 Qualitative Concept	T079 Temporal Concept
121814	T047 Disease or Syndrome	T080 Qualitative Concept	T080 Qualitative Concept
113590	T061 Therapeutic or [...][1]	T082 Spatial Concept	T081 Quantitative Concept

Full Labels: [1] Therapeutic or Preventive Procedure [2] Body Part, Organ, or Organ Component

Figure 2.2 compares the distribution of all SNOMED CT qualifier values between the SNOMED CT CORE Subset and the data-driven analysis. As shown, there appears to be at least some alignment of the distributions, meaning qualifier patterns commonly found in SNOMED CT tend to be similarily common in our data-derived semantic patterns.

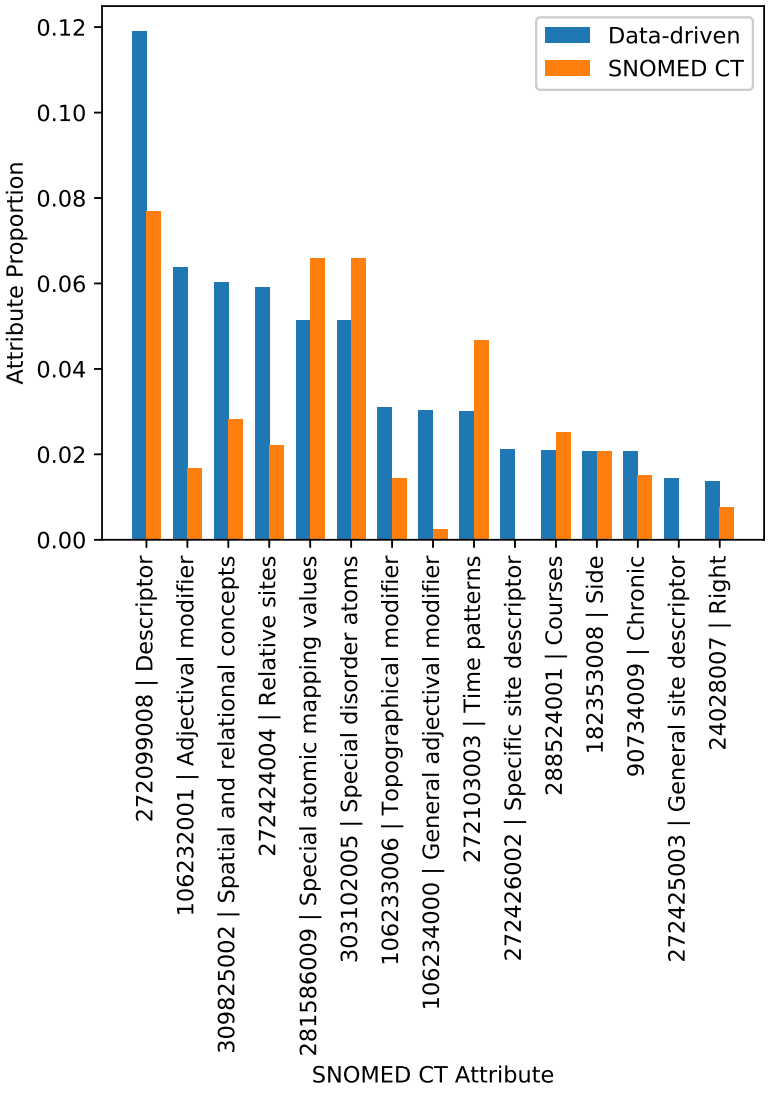


Figure 2.2: Comparing the frequencies of all SNOMED CT qualifier values between the data-driven semantic patterns and the SNOMED CT CORE Subset concepts.

Figure 2.3 again compares the SNOMED CT qualifier distribution, but at the level of the seventy qualifier roots. As with Figure 2.2, some congruence is noted between qualifier usage in SNOMED CT and the RDF corpus. Note that Figures 2.2 & 2.3 show only the top fifteen pairwise attribute comparisons by total frequency.

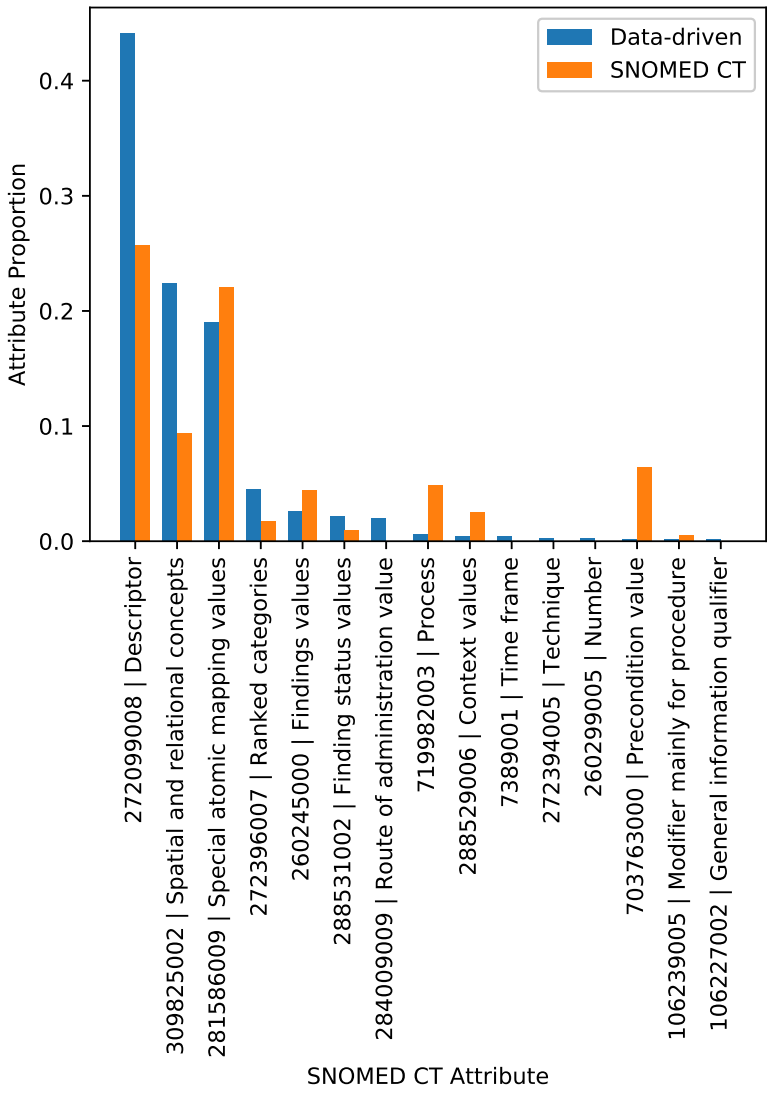


Figure 2.3: Comparing the frequencies of the top-level SNOMED CT qualifier values between the data-driven semantic patterns and the SNOMED CT CORE Subset concepts.

Table 2.7 further compares the attribute usage frequency between the SNOMED CT CORE Subset concepts and the data-derived semantic patterns. The Pearson product-moment correlation coefficients shown in this table quantify the attribute usage correlation between these two data sources for all 6482 qualifiers (*All Qualifiers*), and the seventy *Qualifier Roots*. As seen in this table, there is a moderate correlation between the SNOMED CT and data-driven qualifier frequencies.

Table 2.7: Correlation coefficient values of attribute usage frequency betweeen the data-driven semantic patterns and the SNOMED CT CORE Subset. Counts were computed from two levels of granularity – *All Qualifiers*, and only the top-level *Qualifier Roots*.

	correlation coefficient
All Qualifiers	0.76
Qualifiers Roots	0.82

2.6 Discussion

Tables 2.5 & 2.6 show that semantic relationship patterns can be derived automatically from the corpus based on frequency of occurrence. While important for understanding the overall sublanguage characteristics of the corpus, it is also our goal to allow these language characteristics and constraints to be leveraged dynamically for knowledge engineering tasks. To this end, our methods support dynamic introspection of specific semantic aspects of the corpus via SPARQL. These queries can be targeted for a specific task or area of semantic interest. For example, *Recent myocardial infarction* is a relatively highly occurring entry in the problem list. A semantic frame describing it may either be derived automatically based on frequency of occurrence or curated by domain experts. Either way, given a semantic frame of interest we are able to leverage the corpus to query for additional details. One such example of this is to list all possible lexical variants for a given frame. This information can be used to facilitate the creation of patterns for rule-based NLP systems, or as training data for machine-learning

or other data-driven NLP approaches. Because our RDF-based approach maintains relationships between concepts, we can effectively filter out lexical entries that contain the desired concepts but incorrect relationships, such as "***Recent*** *back and arm pain with no* ***myocardial infarction***." Figure 2.4 shows an example frame and its lexical variants extracted via a SPARQL query.

Recent Myocardial Infarction
`22298006`|*Myocardial infarction (disorder)*| → modifiedBy → `263852005`|*Recent episode (qualifier value)*|

Sample of Lexical Variants Found
Recent peri-operative myocardiac infarction — Recent non-ST elevated myocardiac infarction — Recent heart attack — Recent myocardiac infarct — Recent NQWMI

Figure 2.4: An example semantic frame, **Recent Myocardial Infarction**, and its lexical variants.

Furthermore, a listing of allowable modifiers or attributes for a concept, or a *value set*,[95] can be easily generated at runtime using SPARQL queries. This again allows us to inspect sublanguage characteristics dynamically and on demand. Figure 2.5 shows an example SPARQL query used to derive all the possible clinical course modifiers of kidney insufficiency. In this example, `87828008`|*Insufficiency*| with an `isOf` relationship to `64033007`|*Kidney structure*| can be modified by a given clinical course modifier. Because we have applied the SNOMED CT inference rules as defined by the SNOMED CT OWL ontology (as indicated by the Virtuoso-specific[†] statement `define input:inference ''http://snomed.info_rule''`), any subtype of `288524001`|*Courses (qualifier value)*| will be returned. As the results show, this technique allows for the extraction of specifically typed modifiers used in the corpus for a given concept. This could be practically leveraged for auto-completion of free text in data entry use cases, an application that has already shown promise in lowering the data-entry burden for providers.[96]

[†]http://docs.openlinksw.com/virtuoso/rdfsparqlrule/

```
define input:inference "http://snomed.info_rules"
prefix pl:     <http://mayo.edu/problemlist/>
prefix snomed: <http://snomed.info/id/>
prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#>

select distinct concat(?modifier2class, ' - ', ?label) where {
   ?focus pl:isOf       ?modifier1 .
   ?focus pl:modifiedBy ?modifier2 .
   ?focus     a snomed:87828008 .  # Insufficiency
   ?modifier1 a snomed:64033007 .  # Kidney structure
   ?modifier2 a snomed:288524001 . # Courses (qualifier value)
   ?modifier2 a ?modifier2class .
   ?modifier2class rdfs:label ?label .
}
### Results (ranked by # of occurrences decreasing) ###
# http://snomed.info/id/90734009  - Chronic (qualifier value)
# http://snomed.info/id/424124008 - Sudden onset AND/OR short duration [...]
# http://snomed.info/id/14803004  - Transitory (qualifier value)
# http://snomed.info/id/7087005   - Intermittent (qualifier value)
# http://snomed.info/id/255227004 - Recurrent (qualifier value)
# ... further results omitted for brevity
```

Figure 2.5: A SPARQL query used to derive a specific data-driven modifier list for a given condition.

These methods may also serve as a bridge between unstructured or semi-structured problem lists and FHIR resources – in particular, the FHIR `Condition` resource.[‡] Inherent in the `Condition` resource are modifying attributes such as `severity` and `bodySite`, all centered around a focal concept describing the disease or diagnosis. This paradigm is not dissimilar to the semantic frames-based approach that was explored in this work, and by understanding the problem list sublanguage we believe mappings from narrative-based representations to FHIR resources may be facilitated.

Figures 2.2 & 2.3 demonstrate that the distributions of attribute values between the SNOMED CT stated relationships vs. the patterns extracted via the data-driven techniques are similar. This is evidence that in general, the modeling of concepts in the SNOMED CT CORE Subset matches generally what is found in the data in terms of modifier patterns. Table 2.7 reinforces this claim, as the moderately high correlation coefficients indicate that commonly used attributes in SNOMED CT are also commonly found in the data. When interpreting this finding, it is important to consider that not all meaningful relationships are stated for SNOMED CT concepts – meaning, the textual description of the SNOMED CT concept may indicate more modifiers than are stated via SNOMED CT relationships. As this is an important facet of the semantics of each concept, this distinction is noted on each SNOMED CT concept as the "definition status," which can take one of two values:

- **Defined.** A SNOMED CT concept with stated relationships that are necessary and sufficent to convey the meaning of the concept.
- **Primitive.** A SNOMED CT concept lacking stated relationships such that its full meaning is not expressed via composition of its relationships.

Figure 2.6 shows the difference between a primitive and defined concept, illustrating how some concepts may have modifiers that are only described in the textual description – including concepts in the SNOMED CT CORE Subset.[97] In this example, we

[‡]https://www.hl7.org/fhir/condition.html

would expect the concept 310495003|*Mild depression (disorder)*| to be at least partially "defined" through a stated relationship to the modifier 255604002|*Mild*|. It does not – meaning the "mild" modifier is only stated through the text decription, making the concept primitive. Overall, for all of the SNOMED CT CORE Subset concepts analyzed, 55% were marked as fully defined.

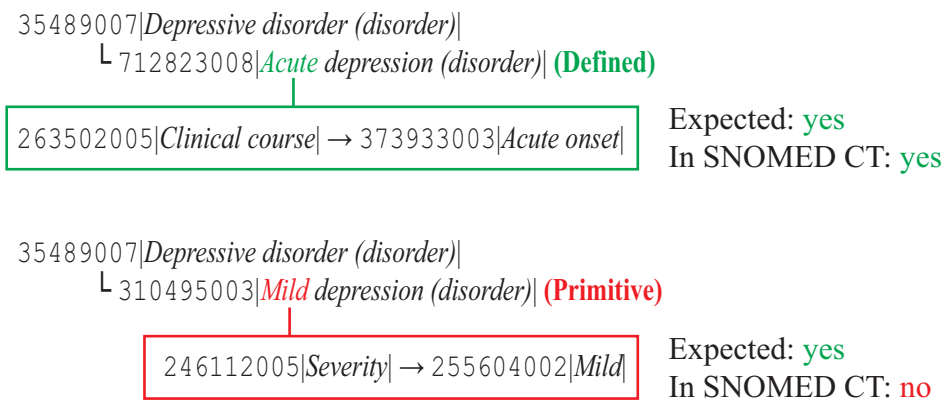


Figure 2.6: An illustration of the difference between SNOMED CT primitive and defined concepts.

2.7 Conclusion

In this work we have explored the sublanguage characteristics of a large-scale clinical problem list corpus. We have demonstrated that by parsing the text using Open Information Extraction techniques, RDF triples can be extracted to represent the lexical relationships in the entries. By standardizing these related text fragments to shared vocabularies via MetaMap, we show how a semantic lexicon can be built to bind the clinical concepts to their representations in the text. Next, by incorporating governed ontologies using OWL, we provide a mechanism to incorporate human-curated domain knowledge into an otherwise completely data-driven technique. We also show that our approach enables real-time inspection of the sublanguage characteristics of our corpus through SPARQL queries, providing knowledge engineers a potential tool for extracting

sets of frequent concept modifiers, composing complex concepts and semantic frames, or searching for possible lexical variants for a given topic. Finally, we present evidence that the SNOMED CT CORE Subset has a moderate amount of conguence with our data-derived concepts in terms of modifier and attribute usage, suggesting that the the SNOMED CT CORE Subset can be an effective base for representing post-coordinated concepts derived from our corpus.

2.8 Limitations and Future Work

The predicates used in the RDF triples (see Table 2.3b) were not semantically categorized, and were generally all considered as RDF sub-properties of a generic `modifiedBy` relationship. Alignment of these predicates to an ontology such as SNOMED CT or the relationship types of the UMLS Semantic Network would allow for more powerful queries and analysis.

In hierarchical concept processing (see Figure 2.1), all `mod - modifier` dependencies were considered during the dependency parsing phase. Precision may be increased by selecting a subset of these dependency types, but that aspect was not explored. Also, dependencies were parsed using the default model of the Standford Dependency Parser. Differences between the default model and a model trained specifically for our corpus were not quantified.

The SNOMED CT Expression Constraint Language[98] is a powerful formalism for creating post-coordinated expressions using a combination of one or more SNOMED CT concepts. As we now have evidence that SNOMED CT is an effective semantic match to our data set, the next chapter of this dissertation extends this work and explores programmatic ways to automatically generate SNOMED CT expressions from text-based problem list entries.

Acknowledgments

This study was funded by the PCORI CDRN-1501-26638-1 and the Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, NCATS U01TR002062, and NIBIB R01EB019403.

Chapter 3

Automating the Transformation of Free-Text Clinical Problems into SNOMED CT Expressions

Kevin J Peterson and Hongfang Liu. Automating the transformation of free-text clinical problems into SNOMED CT expressions. In *AMIA Summits on Translational Science Proceedings*, pages 497–506. American Medical Informatics Association, 2020

Abstract

An important function of the patient record is to effectively and concisely communicate patient problems. In many cases, these problems are represented as short textual summarizations and appear in various sections of the record including problem lists,

diagnoses, and chief complaints. While free-text problem descriptions effectively capture the clinicians' intent, these unstructured representations are problematic for downstream analytics. We present an automated approach to converting free-text problem descriptions into structured Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) expressions. Our methods focus on incorporating new advances in deep learning to build formal semantic representations of summary level clinical problems from text. We evaluate our methods against current approaches as well as against a large clinical corpus. We find that our methods outperform current techniques on the important relation identification sub-task of this conversion, and highlight the challenges of applying these methods to real-world clinical text.

3.1 Introduction

As the healthcare industry increasingly embraces the promise of new data-driven approaches, the challenges of managing and organizing complex patient data become more pronounced.[3] Even before the deployment of the first electronic health record (EHR), healthcare organizations struggled to establish a structured, organized, and standard representation of patient data.[2] A major advance in this area came in the 1960s when Lawrence Weed proposed orienting the data in the patient record around a list of current conditions, or the "problem list."[4] This emphasis on centralizing and enumerating relevant clinical problems enabled patient information to be consumed in a more systematic way, and helped to standardize physicians' interaction with the patient record.[5] A major advantage of this problem-oriented approach is that concise descriptions of clinical problems can summarize and emphasize sections of the larger clinical note narrative. These short phrases describing diagnoses and other patient issues are not limited to the problem list, however. "Summary level" descriptions of clinical problems are also found in diagnosis statements, chief complaints, and reasons for visit,[29,37] and all provide a concise way of expressing pertinent patient conditions.

There are a variety of ways in which these summary level problem descriptions are captured. Free-text is the most expressive form of these problem summaries, capable of capturing the clinical state directly as intended by the clinician.[7] While clinician-friendly, this unstructured representation presents significant problems for downstream analytics.[8] In contrast, a problem may be represented as codes chosen from a controlled terminology. This applies more structure, but limits expressiveness.[61] Even if codified capture is the goal, many systems still allow for free-text entry as a backup if the correct code cannot be readily found.[100,101] These competing representational priorities introduce a fundamental optimization problem in representing these entries – free-text maximizes usefulness for clinicians,[62,102] while structured and codified forms are more amenable to data analytics,[103] standardization activities,[104] and EHR secondary use.[29]

In this study we introduce a method to minimize this conflict between structured and unstructured forms by proposing a framework for converting free-text clinical problem descriptions to codified, structured formats using Natural Language Processing (NLP) techniques. The advantage of structured representations to downstream analytics primarily motivates this effort.[103] By leveraging deep learning methods, we aim to automatically translate text-based problems into Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) Expressions,[105] a structured representation capable of capturing the semantics of summary level problem descriptions in a computable way.

3.2 Background & Related Work

The selection and use of a controlled vocabulary to codify free-text clinical problems has been an active area of research.[37,106] In particular, SNOMED CT[9] has been shown both in principle and in practice to be an effective standard for capturing the semantics of these clinical conditions.[88,97,107] Generally, clinical problems can be represented using SNOMED CT concepts in one of two ways:[108]

- **Pre-Coordinated Concept:** A concept represented as an atomic unit with a

single identifier.

Example: `370221004`|*Severe asthma (disorder)*|

- **Post-Coordinated Concept:** A concept represented as the composition of multiple pre-coordinated concepts that in aggregate define the intended semantics.

  Example: `195967001`|*Asthma (disorder)*| + `24484000`|*Severe (severity modifier)*|

Although pre-coordination has the advantage of simplicity,[109] even summary level problem descriptions are often too expressive to be captured by a single, pre-coordinated SNOMED CT concept. Elkin et al. found that SNOMED CT could only represent 51.4% of problem list entries without composition compared to 92.3% with composition.[21] Liu also found that composition was necessary, observing that 53% of summary level data required two or more SNOMED CT concepts.[29]

Post-coordinated concepts can be readily represented in SNOMED CT via the SNOMED CT Compositional Grammar,[105] a formal specification for representing SNOMED CT post-coordinated expressions. For example,[†] "Severe asthma" can be represented via the following expression consisting of a main focal concept optionally qualified by attribute/value pairs:

```
195967001|Asthma (disorder)|:
        246112005|Severity (attribute)| = 24484000|Severe (severity modifier)|
```

Previous work on converting text to SNOMED CT expressions has focused on the identification and classification of the attribute relationships – for example, what (if any) SNOMED CT attribute best describes the relationship between "Severe" and "asthma." One general approach to this task is to iteratively learn the lexical patterns around how entities relate for a given relationship type.[110] Miñarro-Giménez et al. utilized this technique to fit extracted problem list concepts into learned SNOMED CT relationship patterns.[111] This work leveraged the fact that lexical patterns in pre-coordinated

[†]See: https://confluence.ihtsdotools.org/display/DOCSTART/7.+SNOMED+CT+Expressions for more examples.

SNOMED CT terms are known and relatively predictable.[112]

Kate proposed a different approach to this task – not as a relation extraction task between two concepts within the context of a sentence, but as an attempt to identify if a relationship holds between the entire problem phrase and the concept, or "relation identification."[41] To illustrate the difference, take the "Severe asthma" example above. The Miñarro-Giménez et al. approach would attempt to find the relationship between "Severe" and "asthma," whereas Kate would take the entire phrase "Severe asthma" and attempt to determine which SNOMED CT concepts relate and how. Kate used a Support Vector Machine (SVM)[41] model trained separately for each relationship type to determine if a given relationship held between a concept and the full text entry. Our contributions in this study are focused on extending Kate's work in the following ways: First, we present an end-to-end process for converting free-text summary level problem descriptions to SNOMED CT expressions, enumerating the sub-tasks and incorporating additional NLP techniques such as dependency parsing. Next, we leverage deep learning techniques to increase relation identification performance as compared to the SVM model. Finally, we begin an initial evaluation of model performance in real-world scenarios using summary level problem descriptions extracted from clinical notes.

3.3 Methods

At a high level, our methods are broken into three sequential processing steps as shown in Figure 3.1. First, given a summary level problem description (in text form), the relevant biomedical concepts are recognized and extracted. Next, one of the extracted concepts is chosen as the main semantic focal point, or the *focus concept* of the expression. Following this, the remaining concepts are attached to the focus concept by inferring their role in the expression as a whole. This relation identification task is a critical step in the formation of the SNOMED CT expression and constitutes the bulk of our contributions to this research area. These three steps are explained in detail below.

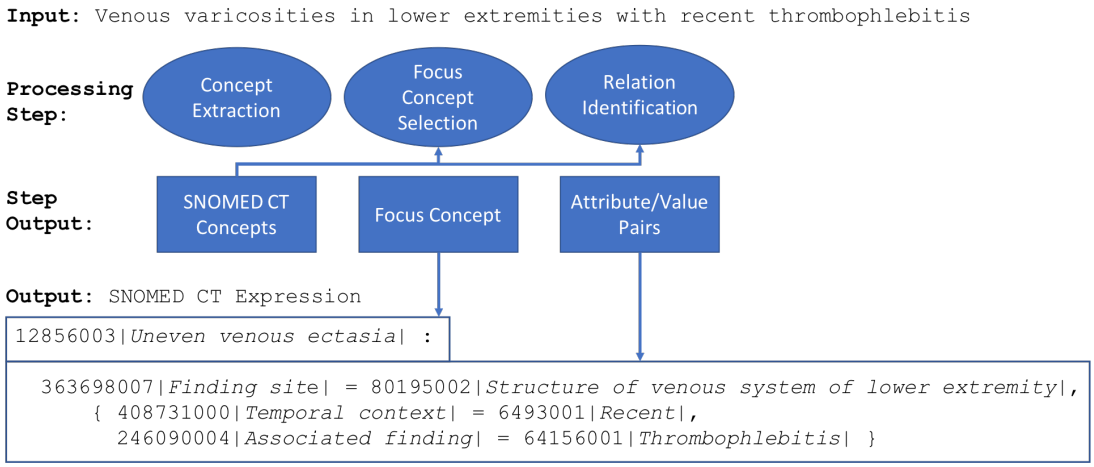


Figure 3.1: Overall steps to convert summary level problem text to a SNOMED CT expression. The processing steps are executed sequentially from left to right with arrows indicating where output from one step feeds into a subsequent step. The output of each processing step is shown linked to the portion of the SNOMED CT expression to which it contributes.

Concept Extraction

The first step in converting a free-text summary level problem description to a SNOMED CT expression is to extract a list of all relevant concepts from the text. To accomplish this, we leveraged MetaMap, a named-entity recognition tool developed by the National Library of Medicine (NLM) to extract Unified Medical Language System (UMLS) concepts from text.[85] An example output of the MetaMap application given the input problem "Venous varicosities in lower extremities with recent thrombophlebitis" is shown below:

```
C0226813:Vein of lower extremity [Body Part, Organ, or Organ Component]
C0042345:Varicosities (Varicosity) [Disease or Syndrome]
C0332185:Recent [Temporal Concept]
C0040046:Thrombophlebitis [Disease or Syndrome]
```

In this example, the extracted concepts are shown with their UMLS Concept Unique Identifiers (CUIs) and textual descriptions. As our goal is to construct SNOMED CT

expressions, we must additionally map the UMLS concepts to SNOMED CT. By configuring MetaMap to match only on SNOMED CT terms, we ensured that each returned UMLS concept encompassed at least one SNOMED CT concept. If the UMLS concept included a single SNOMED CT concept, a direct mapping was made. There are, however, cases where multiple SNOMED CT concepts are incorporated into a UMLS concept.[113] In these cases, all matching SNOMED CT concepts were considered.

Focus Concept Selection

Choosing the focus concept given the list of extracted SNOMED CT concepts is the next step. Here we utilized dependency parsing to align the root node of the problem dependency tree with an extracted MetaMap concept, a technique inspired by Spasić's use of dependency trees to determine the semantic similarity of clinical terms.[114] First, dependency parsing was conducted on the input clinical problem description. Next, the word with the `ROOT` dependency was compared to all extracted MetaMap concepts. Finally, if one of the extracted MetaMap concepts was triggered by the root word, that concept was then chosen as the focus concept.[‡] Figure 3.2 shows the dependency parse with the root word *varicosities*, which is then matched to the relevant concept. We used the spaCy open-source NLP toolkit for dependency parsing along with specifically trained biomedical models from the scispaCy project.[115]

Relation Identification

Identifying the relationships between the problem text and the extracted concepts is the next step. We build primarily on the relation identification task definition as described by Kate[41] and formalize it for our purposes as such: Given an input summary level problem description and a concept extracted via MetaMap, compute the appropriate SNOMED CT attribute (or relationship type) to connect them.

[‡]See https://metamap.nlm.nih.gov/Docs/MMI_Output_2016.pdf for details on how trigger words for MetaMap concepts were obtained.

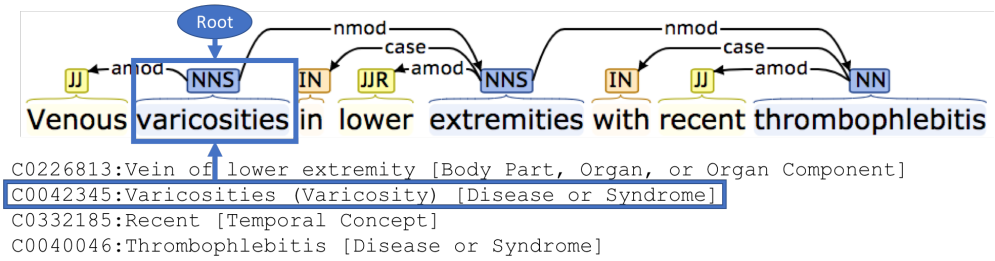


Figure 3.2: Extracting the focus concept from summary level problem descriptions using dependency parsing. The focus concept is selected via alignment to the `ROOT` of the dependency parse.

Figure 3.3 outlines the high-level steps of the relation identification algorithm. First, the problem text and the text of each concept extracted via MetaMap are input into a classifier, the details of which are described further in the following sections. Next, the classifier outputs a probability for each SNOMED CT attribute type indicating the likelihood that a particular relationship holds between the problem text and the extracted concept. Finally, candidate relationships are pruned based on the stated domain and ranges of the SNOMED CT Machine Readable Concept Model (MRCM) Attribute Range Reference Set.[116] For example, if the extracted MetaMap concept of interest is `80195002`|*Structure of venous system of lower extremity (body structure)*|, any SNOMED CT attribute with a range that is incompatible would be removed (such as `424226004`|*Using device (attribute)*|, whose range is limited to children of `49062001`|*Device (physical object)*|), as shown in Figure 3.3. After the pruning, the SNOMED CT attribute with the highest remaining probability is chosen.

A Deep Learning Approach to Relation Identification. Deep learning architectures have shown promise in a variety of NLP tasks,[117] and in this study we compare two popular models for the relation identification classifier. We first consider a Bidirectional Long Short-Term Memory (BiLSTM)[38,39] deep learning architecture. At its base level, a BiLSTM is a specialized type of Recurrent Neural Network (RNN),[118]

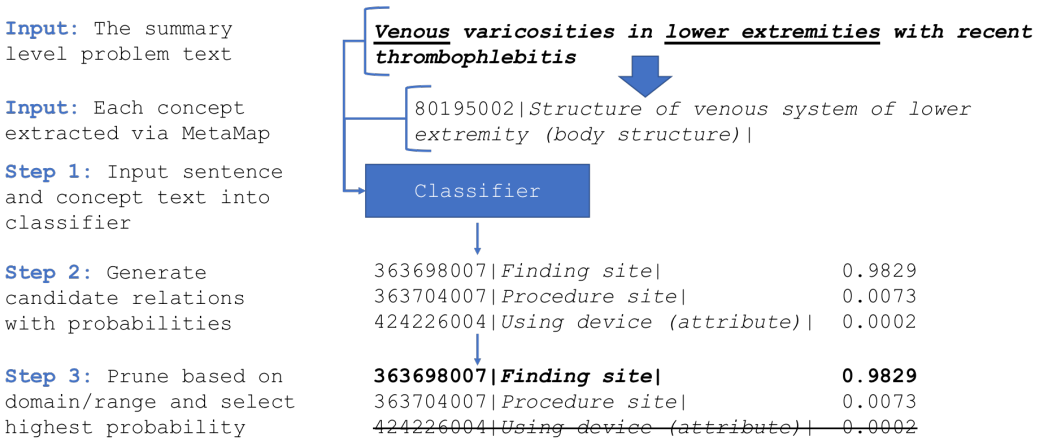


Figure 3.3: Steps to identify relationships between concepts extracted from a free-text summary level problem description.

an artificial neural network architecture that processes information sequentially, factoring in previous input at each current step. This makes RNNs specifically applicable to NLP tasks as text is processed much like a human would – reading words sequentially and inferring the semantics of the current word based on the previous ones.[119] The Long Short-Term Memory (LSTM) facet of the architecture allows for finer control over what information is retained and forgotten by employing more sophisticated feedback loops layered on top of the RNN framework.[38] The bidirectional extension to the LSTM completes our architecture, allowing context to be built not only forward but in the reverse direction as well.[120] In general, the LSTM family of models has shown promising results for NLP relationship extraction tasks.[121]

Convolutional Neural Networks (CNN)[122] are another deep learning architecture. Like LSTMs, CNN models also can recognize spatially related features of the data. For these models, input is filtered via sliding windows which are then pooled to create a subsampled representation of the input sequence. Although used heavily for image processing, CNNs have shown promise in a variety of NLP related tasks including relationship classification.[123] In this study, we compare both models for our relationship identification task.

Both of our deep learning architectures use embedding models based on Bidirectional Encoder Representations from Transformers (BERT), a context-aware language model with state-of-the-art performance on a variety of NLP tasks.[124] For our experiments we leveraged Clinical BERT,[125] a pre-trained BERT model fine-tuned on a clinical text corpus.

Architecture. The high-level architecture and data flow for both the BiLSTM and CNN classifiers were similarly structured. First, two inputs are passed to the `Input` layer – the full problem text and the text of the extracted concept. Next, each input is passed to an `Embedding` layer to create a vector representation of the text input using the BERT model. The vectorized input is then processed by either a BiLSTM with 100 hidden units or a CNN with two convolutional layers. Both models were configured for 20% dropout to avoid overfitting. Finally, a fully-connected `Dense` layer with a softmax activation function is used to output the probabilities for each SNOMED CT attribute type.

Training. In concordance with Kate's approach,[41] stated concept relationships from SNOMED CT US Edition, September 2018 Release[9] were used to train the classifier. Training set construction began with all SNOMED CT stated concept-to-concept relationships excluding `116680003`$|Is\ a|$ relationships. The reasoning for excluding "Is a" relationships is that once the concepts of the expression are known (see the concept extraction step), any "Is a" relationship for these concepts can be directly inferred from the SNOMED CT hierarchy. Next, because our classifier inputs are text, we use the SNOMED CT concept labels for training. As SNOMED CT concepts may contain multiple labels, given each relationship we created training records for all possible pairs of source and target labels. Finally, we excluded relationship types with less than 125 instances, leaving a total of 1,526,043 training records and 78 relationship types available for training. All experiments below used this data set for training, and

Experiments 1 & 2 used a held-out portion of this set for testing. We refer to this data set as the **SNOMED CT Relationship** data set.

Evaluation

System performance was measured for both the Focus Concept Selection and Relation Identification steps of the architecture. The Concept Extraction phase was not directly evaluated, as for this task MetaMap was used without modification (see Reátegui et al.[126] for a recent analysis of MetaMap performance on clinical text). Four experiments were conducted to evaluate model performance.

Experiment 1: First, both the BiLSTM and CNN models were compared against a baseline Naïve Bayes[40] model. All models were trained and tested on the same SNOMED CT Relationship data set. Data was prepared for this experiment by partitioning 25% of the SNOMED CT Relationship set for testing and 75% for training. The Naïve Bayes, BiLSTM, and CNN classifiers were all then trained and tested on the same data, with the exception of a further 20% of the training data being withheld from the BiLSTM and CNN models for validation. For testing, we recorded the F_1 scores for each individual attribute as well as overall averages for all classifiers.

Experiment 2: Next, we compared our results to previously reported results of Kate's SVM model.[41] We followed Kate's evaluation procedures in order to replicate his experiment using the best performing model from Experiment 1: For each attribute, 5000 relationships of the desired type were randomly selected from the SNOMED CT Relationship set along with an equal number of negative examples. Given this test set, the ability of the classifier to correctly determine whether or not the chosen relationship was present was recorded.

Experiment 3: To evaluate relation identification model performance in real-world

scenarios and test generalizability to different data sets, we utilized a large data set of summary level clinical problems extracted from a Mayo Clinic corpus of over 14 million clinical documents. This corpus has been extensively analyzed by Liu et al. and is a rich source of diverse summary level problem descriptions.[29] Three trained annotators examined a random subset of 401 summary level clinical problem descriptions from this corpus. First, the annotators were asked to select the word or words that represented the focus concept of the problem. Next, the annotators we tasked with connecting the focus to relevant modifiers using one of twenty-one relationship types. These relationship types were chosen via analysis of common attributes across prominent clinical problem models including the Fast Healthcare Interoperability Resources (FHIR) - `Condition` resource,[11] Clinical Element Model (CEM) - `ClinicalAssert` model,[43] and openEHR - `Problem/Diagnosis` archetype.[44] For more information, see Goossen[127] for details regarding these models. Only relationships supported by >20 annotations in the test corpus are evaluated in this study. This test corpus was then used to evaluate both the Focus Concept Selection and Relation Identification parts of the pipeline. The three pair-wise Cohen's kappa inter-annotator agreement values for each annotator pair were 0.78, 0.85, 0.84 for the focus concept annotations and 0.76, 0.76, 0.82 for the relationship annotations.

Experiment 4: Finally, we evaluated how effective the dependency parse-based method is at locating the focus concept of the clinical problem. Using the test data set described in Experiment 3, we evaluated the ability to correctly identify the focus using three spaCy dependency parse models: (1) general-purpose English, (2) scispaCy biomedical, and (3) a custom model based on scispaCy biomedical fine-tuned with 150 manually-annotated clinical problem text examples.

3.4 Results

Experiment 1: Overall results of the BiLSTM and CNN models compared to the Naïve Bayes baseline for the SNOMED CT relation identification task are shown in Table 3.1. While both deep learning models significantly outperformed the baseline, the BiLSTM model slightly outperformed the CNN model. Because of this result, downstream experiments focus on the BiLSTM model.

Table 3.1: Comparing overall SNOMED CT relation identification model performance.

	Accuracy	**F_1 score (macro avg)**	**F_1 score (weighted avg)**
Naïve Bayes (baseline)	0.720	0.460	0.665
CNN + Clinical BERT	0.886	0.822	0.880
BiLSTM + Clinical BERT	**0.888**	**0.851**	**0.888**

Experiment 2: Figure 3.4 shows the results of the BiLSTM classifier compared to the SVM classifier results as reported by Kate.[41] Note we did not evaluate `116680003`|*Is a*| relationships, so we do not fully correspond to Kate's results. Also, Kate reported two results for `363698007`|*Finding site*| based on two different domains/ranges. The F_1 score reported in Figure 3.4 represents the highest of the two scores.

Experiment 3: The results of the evaluation of the BiLSTM classifier against the annotated test corpus are shown in Table 3.2. This table contrasts two main data points: (1) the **Clinical Text F_1 score**, which measures the classifier's ability to predict the correct relationship type given the focus and a modifier in the clinical text corpus, and (2) the **SNOMED CT Relationship F_1 score**, which is the corresponding value derived from Experiment 1 for the given attribute. The $\mathbf{F_1\Delta}$ value shows the difference between the two scores, illustrating the difference in performance when testing on relatively predictable SNOMED CT terms vs. real clinical text. The eight relationships supported by >20 annotations in the clinical text test dataset are displayed.

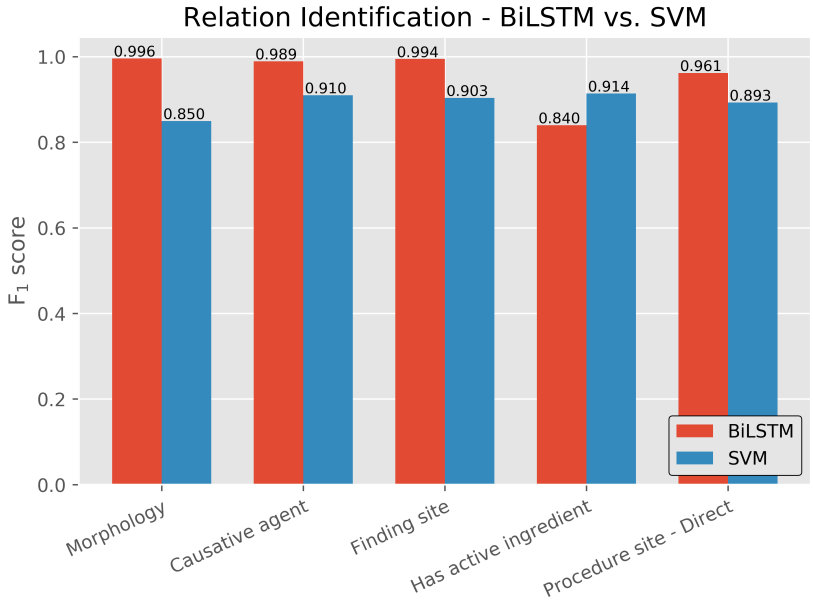


Figure 3.4: Comparing the BiLSTM + Clinical BERT model for SNOMED CT relation identification to the SVM model as reported by Kate[41] for five SNOMED CT attributes.

Table 3.2: Comparing BiLSTM relation identification scores using two different test data sets: real-world clinical text (Clinical Text F_1 score) and relationships from SNOMED CT (SNOMED CT Relationship F_1 score). The $F_1\Delta$ value equals Clinical Text F_1 score minus SNOMED CT Relationship F_1 score.

Attribute	Clinical Text Precision	Clinical Text Recall	Clinical Text F_1 score	SNOMED CT Relationship F_1 score	F_1 Δ
Severity	1.000	0.880	0.936	0.882	0.054
Laterality	0.990	0.950	0.970	0.999	-0.029
Clinical course	1.000	0.800	0.889	0.994	-0.105
Finding site	0.956	0.790	0.865	0.988	-0.123
Due to	0.528	0.487	0.507	0.830	-0.323
Has interpretation	0.577	0.714	0.638	0.992	-0.354
Following	0.733	0.306	0.431	0.837	-0.406
Associated with	0.579	0.134	0.218	0.738	-0.520

Experiment 4: Finally, Table 3.3 shows the effectiveness of using a dependency parse-based method for selecting the focus concept of a problem description. For each dependency parse model the accuracy is shown, where accuracy in this context reflects the number of times the model selected the same focus span as the human annotators over the 401 total entries in the test set.

Table 3.3: Evaluating the performance of the dependency parse-based method for selecting the focus concept of the clinical problem. Three different dependency parse models were evaluated.

Model	**Accuracy**
Default spaCy English (baseline)	0.68
ScispaCy Biomedical	0.75
ScispaCy Biomedical + fine-tuning	0.91

3.5 Discussion

Overall, both the CNN and BiLSTM significantly outperformed the Naïve Bayes classifier in identifying relationships between two SNOMED CT concepts, as shown in Table 3.1. This comparison provides evidence that a deep learning architecture is a viable approach and can outperform a simple Naïve Bayes baseline. The BiLSTM did also outperform the CNN model slightly. The difference was most evident in the F_1 macro avg score, which is important as this metric gives equal weight to each relationship type and disregards any SNOMED CT relationship class imbalance.

Figure 3.4 shows that a deep learning approach can outperform SVM classifiers at the relation identification task with SNOMED CT relationships. The BiLSTM scored higher for four of the five attributes tested, while the SVM outperformed the BiLSTM for one attribute: "Has active ingredient." It is worth noting that we cannot directly compare Kate's results[41] with our BiLSTM model beyond these five attributes listed in

Figure 3.4. Overall F_1 scores are not comparable because Kate's model was trained and tested using the top 14 SNOMED CT attributes only, while our overall F_1 scores (see Table 3.1) are derived from a classifier trained and evaluated on 78 relationship types. Also, Kate's overall results factor in scores for `116680003`|*Is a*| attributes – relationships that we omitted.

Table 3.2 highlights the challenges that come with applying these methods to clinical text. For several attributes, relation identification was significantly worse against real-world clinical problem descriptions compared to SNOMED CT relationships (as shown by $F_1\Delta$). These differences are not unexpected – the SNOMED CT text used for training is relatively structured,[112] but actual clinical problem descriptions are not. As shown, performance degradation for several relationship classes is pronounced, with a highly negative $F_1\Delta$ score signifying low model generalizability from the SNOMED CT text corpus to the clinical text.

Finally, the focus concept selection results in Table 3.3 show not only that a dependency parse-based method of focus concept selection is an effective technique, but that using a domain-specific pre-trained model does boost performance noticeably. Even more, these results indicate that even minimal fine-tuning (150 manual annotations) can have a fairly large impact on overall performance.

3.6 Conclusion

The goal of this work was to present an end-to-end system for converting unstructured summary level problem descriptions into SNOMED CT expressions. Our contribution focused primarily on introducing a new deep learning method for relation identification between concepts and problem phrases. We show that our method outperforms current approaches to identifying relationships between clinical phrases and SNOMED CT concepts, a fundamental part of building SNOMED CT expressions. We also show that a model trained exclusively on SNOMED CT stated relationship text does not transfer

to clinical text without performance degradation.

3.7 Limitations & Future Work

Our study has several limitations. First, there is no available gold-standard test set for evaluating the full conversion of text-based problem descriptions to SNOMED CT expressions – we must evaluate individual steps of the pipeline independently. Furthermore, such a test set is challenging to construct as there may exist more than one way syntactically to represent the same conceptual expression. Also, it has been shown that codification of problems by physicians is subject to considerable variation.[128] All these factors together make quantitative evaluation of this task difficult.

In the following chapter we address one of the main challenges of this study: generalizability to clinical text. We also target the FHIR `Condition` resource as our target representation, representing the last level of standardization that will be discussed in this dissertation.

Acknowledgments

This study was funded by grant NCATS U01TR02016. We thank Donna Ihrke, Luke Carlson, and Sunyang Fu for assisting in the test corpus annotation and guideline development.

Chapter 4

A Corpus-Driven Standardization Framework for Encoding Clinical Problems with HL7 FHIR

Kevin J Peterson, Guoqian Jiang, and Hongfang Liu. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. *Journal of Biomedical Informatics*, page 103541, 2020

Abstract

Free-text problem descriptions are brief explanations of patient diagnoses and issues, commonly found in problem lists and other prominent areas of the medical record. These compact representations often express complex and nuanced medical conditions, making their semantics challenging to fully capture and standardize. In this study, we describe a framework for transforming free-text problem descriptions into standardized Health

Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) models. This approach leverages a combination of domain-specific dependency parsers, Bidirectional Encoder Representations from Transformers (BERT) natural language models, and cui2vec Unified Medical Language System (UMLS) concept vectors to align extracted concepts from free-text problem descriptions into structured FHIR models. A neural network classification model is used to classify thirteen relationship types between concepts, facilitating mapping to the FHIR `Condition` *resource. We use data programming, a weak supervision approach, to eliminate the need for a manually annotated training corpus. Shapley values, a mechanism to quantify contribution, are used to interpret the impact of model features. We found that our methods identified the focus concept, or primary clinical concern of the problem description, with an F_1 score of 0.95. Relationships from the focus to other modifying concepts were extracted with an F_1 score of 0.90. When classifying relationships, our model achieved a 0.89 weighted average F_1 score, enabling accurate mapping of attributes into HL7 FHIR models. We also found that the BERT input representation predominantly contributed to the classifier decision as shown by the Shapley values analysis.*

4.1 Introduction

The problem-oriented medical record (POMR) was a significant change in how the clinical patient record was structured.[4] Introduced in 1968, this strategy involves using concise descriptions of a patient's current health concerns to serve as indexed headings into the larger medical chart.[130] These summary level "problem descriptions" describe complex clinical conditions with important supporting context such as severity/stage, body location, related or contributing conditions, and so on, and are an integral part of the POMR as a whole.

By orienting the record around clinical problems, the POMR is by definition predicated on the ability to accurately and succinctly describe a patient's pertinent issues.[131]

Furthermore, it also places a greater burden on ensuring that problems are described comprehensively and in a standardized way.[6] Although these challenges pre-dated the widespread implementation of the electronic health record (EHR), the structure inherent in EHRs did not alleviate issues regarding how clinical problems are represented.[132] Specifically, while the expressiveness of free-text is required by clinicians to convey their impressions and reasoning regarding a patient's problems,[7] structured representation and standardization are beneficial for processing and analytics.[133]

Codification, or the assignment of codes or terms from a controlled terminology, is a common strategy for capturing and standardizing the semantics of a clinical problem. This can be done by the clinician directly, but requires significant time and effort[134] and adds to an already full clinical workload.[135] Alternatively, this coding may be accomplished using automated or semi-automated Natural Language Processing (NLP) techniques. Even if automated, codification often fails to capture the entirety of the clinician's intent, a situation known as the "content completeness problem."[136,137] This issue is rooted in the fact that natural language descriptions of medical problems are often too expressive to be fully represented via a finite set of terms.[138]

The content completeness problem is of particular importance to clinical problem descriptions, as it has been shown empirically that clinical problems often cannot be sufficiently described by a single concept, but instead require a set of concepts to capture modifiers and other related context.[21,29,139] To account for this, logical models can be paired with codification to create a more robust standard for data representation and exchange.[140] Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR), an emerging specification for representing clinical data, is a prominent example of this type of standardization.[11] FHIR specifies several models (or "Resources") for many types of healthcare data, including representations specifically suited for clinical problems.

The goal of this study is to introduce a framework for encoding free-text clinical problem descriptions using HL7 FHIR. Our methods focus on combining machine learning techniques with rule-based methods and domain-specific knowledge bases to

map free-text problem descriptions to FHIR-based structured representations.

4.2 Background and Significance

The standardization of free-text clinical problems has long been a focus of research. Concept extraction, or mapping text mentions to standardized terminologies or ontologies, is a fundamental clinical NLP task and an important step towards a standardized problem representation. Prominent implementations such as MetaMap[85] and Clinical Text Analysis and Knowledge Extraction System (cTAKES)[141] have been widely adopted and used for a variety of standardization applications.[142]

While concept extraction is an essential first step, further standardization may be applied by organizing the extracted concepts into logical structures that better capture their full context and semantics. The Medical Language Extraction and Encoding System (MedLEE)[70] accomplishes this using frames,[74] or structures that link concepts to their modifiers and related terms. Similar notions of combining, or "post-coordinating" concepts are natively built into the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) ontology via the SNOMED CT compositional grammar.[105]

In their most structured form, clinical problems may be represented via common information models.[143] Efforts such as the Clinical Element Model (CEM)[43,140] and the Clinical Information Modeling Initiative (CIMI)[144] aim to define a standard set of attributes and modifiers for clinical data exchange. HL7 FHIR is the latest of these efforts, and for this study the `Condition` resource of the FHIR specification is the chosen target for clinical problem representation.

The use of NLP to extract information from clinical text as FHIR resources is a growing field of study,[56,145,146] driven in part by the increasing prominence of FHIR in the healthcare information landscape.[147] The NLP2FHIR project, based on several Unstructured Information Management Architecture (UIMA)[148] tools, extracts a broad range of FHIR resources from unstructured clinical notes.[149] In contrast to the broader

scope of NLP2FHIR, our study exclusively focuses on encoding summary level problem descriptions into FHIR `Condition` resources. While care was taken to ensure our techniques could generally be applied to other free-text clinical data (such as procedures, labs, medications, and so on), the nuances of those transformations are beyond the scope of this work. While a focus on clinical conditions narrows our purview, several new challenges are introduced:

- Although free-text problem descriptions are generally terse, they are surprisingly expressive, and routinely encompass semantics outside the bounds of single concepts from a controlled terminology or vocabulary.[21,29]

- They are often phrased as collections of medical terms, which generally have quite different grammatical and linguistic characteristics as compared to the larger clinical note narrative.[150] Specifically, these problems are generally represented as noun phrases as opposed to full sentences.

- It is known that non-standard grammar, sentence structure, and word usage, or *non-canonical text*, poses significant problems for NLP model reuse.[151] As stated, these problem descriptions do not follow a canonical notion of grammar or structure – and complex noun phrases have proven to be especially difficult for many NLP parsing tasks.[152,153] Given this, existing NLP models may perform poorly when applied to these problems.

Our main contribution in this study is a standardization framework for clinical problem descriptions using HL7 FHIR, an expansion of our previous work in Chapter 3 on codifying problems using the SNOMED CT compositional grammar.[99] We extend this previous study in the following areas: First, we update our target representation to the FHIR `Condition` resource to take full advantage of the growing FHIR healthcare ecosystem. Next, we increase the performance of our previous methods through the addition of training techniques based on incorporating rule-based methods, weak labeling,

and distant supervision. This was necessary to make our previous methods more resilient to the linguistic heterogeneity seen in real-world clinical text – a main limitation of our previous work. Finally, we add a thorough analysis of our model features using the latest neural network explainability methods, giving us important insight into what model features are important and why.

4.3 Methods and Materials

We define our clinical problem description standardization task as such: Given a free-text description of a patient's clinical problem, output an HL7 FHIR `Condition` resource representing the codified problem and all relevant modifiers and context. We account for the specific challenges of processing problem descriptions using three general methodological foci: (1) an emphasis on leveraging existing pre-trained models to maximize transfer learning, using fine-tuning where necessary, (2) the incorporation of rule-based methods with neural network models to avoid manual training data annotation, and (3) the usage of recent advances in neural network explainability to examine the importance of the features in our model. Our methods are broadly segmented into five subtasks that gradually build an increasingly structured and standardized representation of the clinical problem. Figure 4.1 highlights the high-level steps of the standardization framework, the details of which are explained further below.

4.3.1 Preprocessing: Dependency Parsing

Dependency parsing is the formalization of text into a graph of words and their syntactic relationships. It is an important input into many clinical NLP tasks such as concept extraction,[154] semantic parsing,[155] and negation detection,[156] and contributes prominently to several of the subtasks described in our methods below.

For all clinical problem dependency parsing we used the spaCy NLP platform with a custom parsing model fine-tuned from the pre-trained ScispaCy Biomedical model.[115]

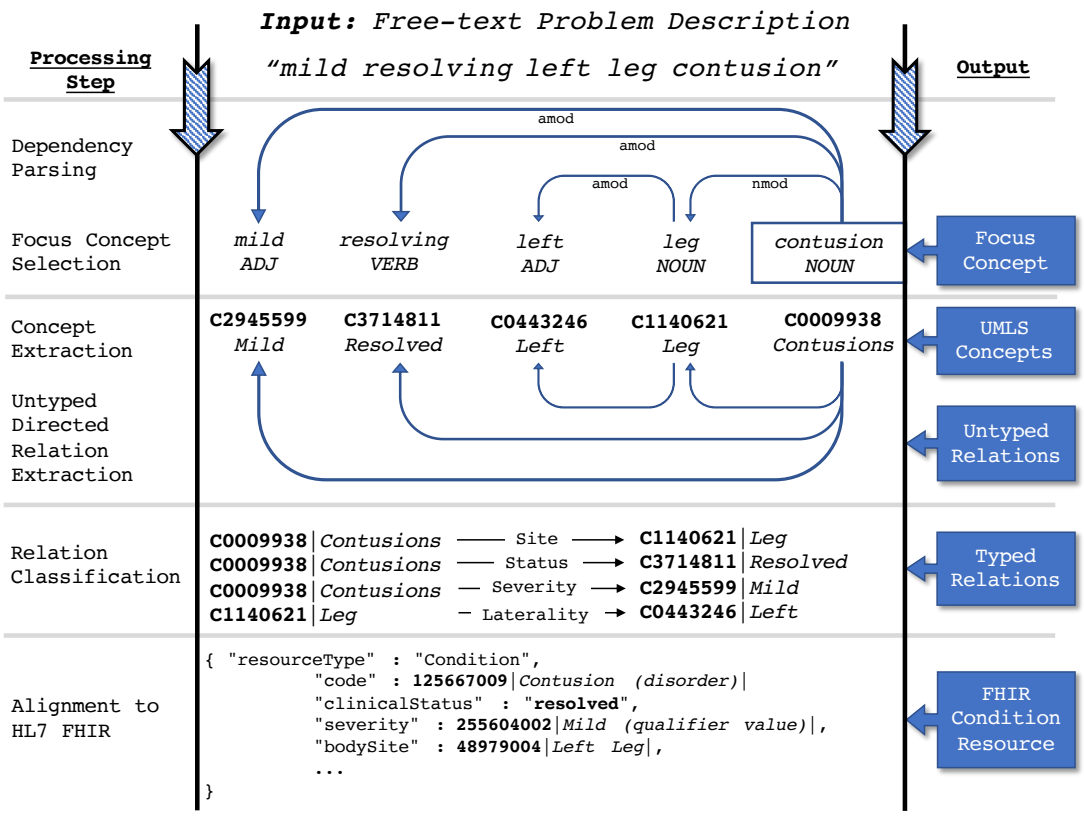


Figure 4.1: The high-level processing steps for encoding a free-text clinical problem description into an HL7 FHIR `Condition` Resource.

To fine-tune the model, we manually annotated 141 problem descriptions with their correct dependency parses. To keep the amount of manual annotation as low as possible, we used data augmentation, a strategy to increase the size and heterogeneity of training data. While many NLP data augmentation algorithms focus on expanding synonyms,[157,158] we used SNOMED CT to expand SNOMED CT `Qualifier Value` terms matched in the text, similar in concept to what Kobayashi describes as "contextual augmentation."[159] For example, given a problem *"severe contusion"*, we recognize that the term *severe* is a child of the SNOMED CT concept `272141005`|*Severities*|. Given this, we can expand this training example with other *Severities*, yielding *"mild contusion"*, *"moderate contusion"*, and so on. With data augmentation, we expanded our training set to 349 entries.

4.3.2 Subtask: Focus Concept Selection

We define the "focus concept" of a problem description as the semantic root, or primary concept from which the remaining concepts are either directly or indirectly connected. As summary level problem descriptions are primarily noun phrases, we hypothesize that the `ROOT` word of the dependency parse will align with the focus concept, a hypothesis based primarily on the work of Spasić et al.[114] Representing our text as a set of tokens S, this step aims to learn a function that inputs the problem description tokens S and outputs the root token r such that $r \in S$. This technique closely aligns with methods used in our previous work to select the focus concept for a SNOMED CT expression.[99] Note that this approach assumes each problem description primarily describes one and only one clinical problem. While a single problem description may contain several mentions of different conditions, signs, or symptoms, it is assumed that all serve to modify or add context to a single focus problem. Other formats of problem descriptions, such as concatenations of multiple, unrelated problems (for example: "Tonsillitis; fracture of the femur") are beyond the scope of this subtask.

4.3.3 Subtask: Concept Extraction

We extracted concepts from the text using MetaMap,[85] a tool based on the Unified Medical Language System (UMLS)[30] used to link free-text mentions of biomedical concepts to their corresponding UMLS concepts via Concept Unique Identifiers (CUIs). Given problem description text S composed of $s_1, ..., s_{|S|}$ tokens, we used MetaMap to implement the mapping $S \rightarrow E$ where E is a set of UMLS concepts.

4.3.4 Subtask: Untyped Directed Relation Extraction

It has been shown by Reichartz et al.[160] that using a dependency parse tree can be an effective way to extract semantic relationships between entities in text. Several biomedical relation extraction systems leverage the dependency parse tree, incorporating it into a wide variety of model architectures including rule-based approaches,[161,162] kernel-based methods,[163,164] and recently deep learning models.[165,166] The goal of this subtask is similarly to extract a set of untyped, directed entity relationships between a source and target entity, or (e_1, e_2). We implemented this by connecting pairs of entities via their shortest path in the dependency parse. Entities e_1 and e_2 are considered connected if (1) a path exists from e_1 to e_2, and (2) no other entity exists on the path between them.

4.3.5 Subtask: Relation Classification

Classifying relationships between biomedical concepts is an important task with wide-reaching applications,[167] with use cases including chemical-disease relations,[168,169] disease-symptom relations,[170] and protein–protein relations.[171] In this subtask we aim to classify the untyped relationships extracted via the Untyped Directed Relation Extraction subtask, or $r(e_1, e_2)$, where r represents the relationship type, and e_1 and e_2 represent the source and target entity, respectively. Further details regarding the methodology for this subtask are detailed below.

Relation Type Selection

We selected twenty-one relation types for inclusion in this study. Relation types were chosen via satisfaction of one or more of the following criteria: (1) they can be directly mapped to an attribute of the FHIR v4.0.1 `Condition` resource, (2) they are associated with a standard FHIR `Condition` extension, or (3) they are included in prominent clinical data models other than the FHIR specification. This was done to give our classifier a broader semantic range given that the FHIR specification allows for extensibility. Relationship types that did not come directly from the FHIR specification were obtained via a survey of the following models: the Clinical Element Model (CEM) - *ClinicalAssert* model,[43,140] the openEHR - *Problem/Diagnosis* archetype,[44] and the Clinical Information Modeling Initiative (CIMI) *FindingSiteConditionTopic* logical model.[144]

Table 4.1 shows these relationship types and their mapping to the FHIR model. Note that as shown in the table, some map directly to attributes in the FHIR `Condition` resource, some map to standard FHIR extensions, and some have no direct mapping to FHIR at all. For those with no FHIR mapping, the standardization framework(s) from which they were selected are listed.

Table 4.1: The set of twenty-one relation types considered in the Relation Classification subtask with their mappings to the FHIR `Condition` resource.

Relation Type	FHIR Mappings
	Base FHIR Condition
clinicalStatus	Condition.clinicalStatus
verificationStatus	Condition.verificationStatus
severity	Condition.severity
bodySite	Condition.bodySite
stage	Condition.stage

Table 4.1 (Continued): The set of twenty-one relation types considered in the Relation Classification subtask with their mappings to the FHIR `Condition` resource.

Relation Type	FHIR Mappings
	Standard FHIR Extensions
dueTo	condition-dueTo
ruledOut	condition-ruledOut
occurredFollowing	condition-occurredFollowing
associatedSignAndSymptom	condition-related
laterality	BodyStructure
anatomicalDirection	BodyStructure
	Non-FHIR Attributes
course	CIMI:FindingSiteAssertion - clinicalCourse
	OpenEHR:Problem/Diagnosis Archetype - Course label
periodicity	CIMI:FindingSiteAssertion - periodicity
exacerbatingFactor	CIMI:FindingSiteAssertion - exacerbatingFactor
interpretation	CIMI:FindingSiteAssertion - interpretation
findingMethod	CIMI:FindingSiteAssertion - method
historicalIndicator	CEM:ClinicalAssert - historicalInd
	OpenEHR:Problem/Diagnosis Archetype - Current/Past?
certainty	OpenEHR:Problem/Diagnosis Archetype - Diagnostic certainty
	CEM:ClinicalAssert - likelihood
risk	CEM:ClinicalAssert - riskForInd
negatedIndicator	generic negation modifier
otherwiseRelated	any other non-specified relationship

Relation Classification Model Architecture

An artificial neural network model was chosen for the relation classification task. The architecture consists of a fully-connected neural network with one hidden layer containing 256 nodes using ReLU activation functions. The output layer of this network contains one node for each relation class (see Table 4.1) using softmax activation functions. Dropout rates of 0.5 were used to prevent overfitting. The ultimate output of the model is a probability for each relationship class. We used the Keras framework[172] to implement our model.

A variety of input representations ranging from text embeddings to facets of the extracted UMLS concepts were selected as input features. Special emphasis was placed on incorporating features based on pre-trained, domain-specific models where transfer learning could be leveraged. The full feature set is described below.

Source/Target Entity Text Embedding (BERT). We transformed text into a suitable input format via Bidirectional Encoder Representations from Transformers (BERT),[124] a deep learning-based language representation aimed at capturing the semantic intent of words in context. We specifically used the pre-trained ClinicalBERT model, a BERT-based model trained on clinical notes.[173] For each of the source and target entities we obtained a 768-dimensional vector using mean pooling of the second-to-last BERT layer. BERT embeddings were incorporated into our pipeline via the bert-as-service project.[174]

Source/Target Entity Concept Embedding (cui2vec). We incorporated vectors from the extracted source and target entity concepts using cui2vec, a UMLS concept embedding model.[175] The concept embedding of cui2vec was used specifically for transfer learning of UMLS semantics into our model. The UMLS concepts of the source and target entities were mapped to 500-dimensional vectors from a pre-trained cui2vec model.[175]

Dependency Parse Shortest Path. The shortest path through the dependency

parse between the source and target entities is hypothesized to be helpful in determining their semantic relationship.[163] Furthermore, it has been shown that incorporating this as a feature in a machine learning model can improve relation extraction performance.[176] A BERT vector of the path was used to represent this feature.

Source/Target Entity Semantic Type. For both the source and target concepts extracted during the Concept Extraction step, MetaMap also assigns concepts one or more of the 127 UMLS semantic categories called *semantic types.*[177] These categories were used as features to represent the high-level semantics of the concepts.

Source/Target Entity Semantic Type Group. The source and target entity semantic types are additionally grouped into fifteen even broader categories called *semantic groups*,[178] representing the coarsest level of semantics in our feature set.

Data Programming

We merged rule-based and neural network NLP approaches by using data programming,[42] a technique for creating a weakly-labeled training data set given a set of "labeling functions," or domain-specific rules crafted by subject matter experts or other domain-specific oracles. We used the Snorkel framework to train a generative model from our labeling functions, and used that model to generate training data for the downstream neural network model.[179] Data programming allows us to address two main challenges: (1) we avoid the cost and time of using specially trained clinical informaticians to hand-annotate training data,[180] and (2) we incorporate domain knowledge via symbolic approaches and distant supervision,[181,182] highlighting the importance of leveraging domain expertise via rule-based approaches.[183,184] For our data programming implementation, we created approximately thirty labeling functions using both hand-crafted rules and distant supervision using SNOMED CT. Our neural network training data set was generated by applying the generative data programming model to 100,000 problem descriptions extracted from a large clinical corpus.[29]

Interpretability & Feature Attribution

Shapley values are a mechanism to quantify the contributions (in terms of gains or losses) of members of a coalition cooperating toward a common goal.[185,186] Shapley values have been applied to determining feature importance of machine learning models,[187] and are used here to gain insight into our relationship classification model.

Feature attribution can be cast as a game theory problem as such: First, given a set of model features F, assume we wanted to determine the contribution of some feature j where $j \in F$. Next, we generate a subset of F as S such that S does not include the feature of interest j. We test the contribution of feature j by measuring its contribution $v(S \cup \{j\}) - v(S)$ where v denotes the *characteristic function*, or the total contribution of a set of features toward the end goal. In our case, the characteristic function input is a set of features, and the output is the probability of the chosen relationship label. By repeating this for all subsets of F such that $S \subseteq F \setminus \{j\}$, we compute the Shapley value ϕ_j as:

$$\phi_j(v) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(v(S \cup \{j\}) - v(S)\right)$$

We then explain the approximate contribution of a feature $\hat{\varphi}_j$ by averaging all Shapley values over a random sampling of n training samples.[188]

$$\hat{\varphi}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j^{(i)}$$

4.3.6 Subtask: Alignment to HL7 FHIR

Table 4.1 shows the basic mappings of our chosen relation types to the FHIR `Condition` resource. Any relationship that does not map directly to a FHIR `Condition` attribute will be added as a FHIR extension, and the `Condition.code` attribute of the FHIR `Condition` resource will be set to the focus concept extracted via the Focus Concept

Selection subtask. FHIR alignment also involves mapping each extracted UMLS concept to SNOMED CT. To do this, we used the UMLS Metathesaurus to find the SNOMED CT concept associated with the given UMLS concept. If the UMLS concept maps to more than one SNOMED CT concept, each SNOMED CT concept will be added as a FHIR `Coding` for the particular FHIR attribute.

4.3.7 Evaluation & Experiments

Six hundred problem descriptions extracted from a large clinical notes corpus were manually annotated by three annotators. Annotation consisted of finding the focus concept of the problem (i.e. the primary disease or finding), all related diseases, findings, or modifiers, and the relationship types that connect them (see Table 4.1). BRAT,[189] a freely-available annotation tool, was used to conduct the annotation. Inter-annotator agreement was measured via Krippendorff's alpha score.[190] We conducted the following experiments to analyze the performance of our framework.

Untyped Directed Relation Extraction & Focus Concept Selection

Because both Untyped Directed Relation Extraction and Focus Concept Selection subtasks are based on dependency parsing, we evaluated them in tandem. For Untyped Directed Relation Extraction, we evaluated the effectiveness of using dependency parsing to determine related concepts (regardless of relationship type). We used the evaluation corpus detailed above with the following experiment: First, we extracted from the evaluation corpus all annotated relationships and retained a list of all source/target tuples. Then, we compared the annotated relationships with those asserted from the dependency parse. Figure 4.2 illustrates this test. In this example, **Dependency Parse A** produces three incorrect relationships, while **Dependency Parse B** fully corresponds to the human-annotated example.

In our Focus Concept Selection subtask, we hypothesize that the `ROOT` dependency of the dependency parse will correspond to what a human annotator would specify as

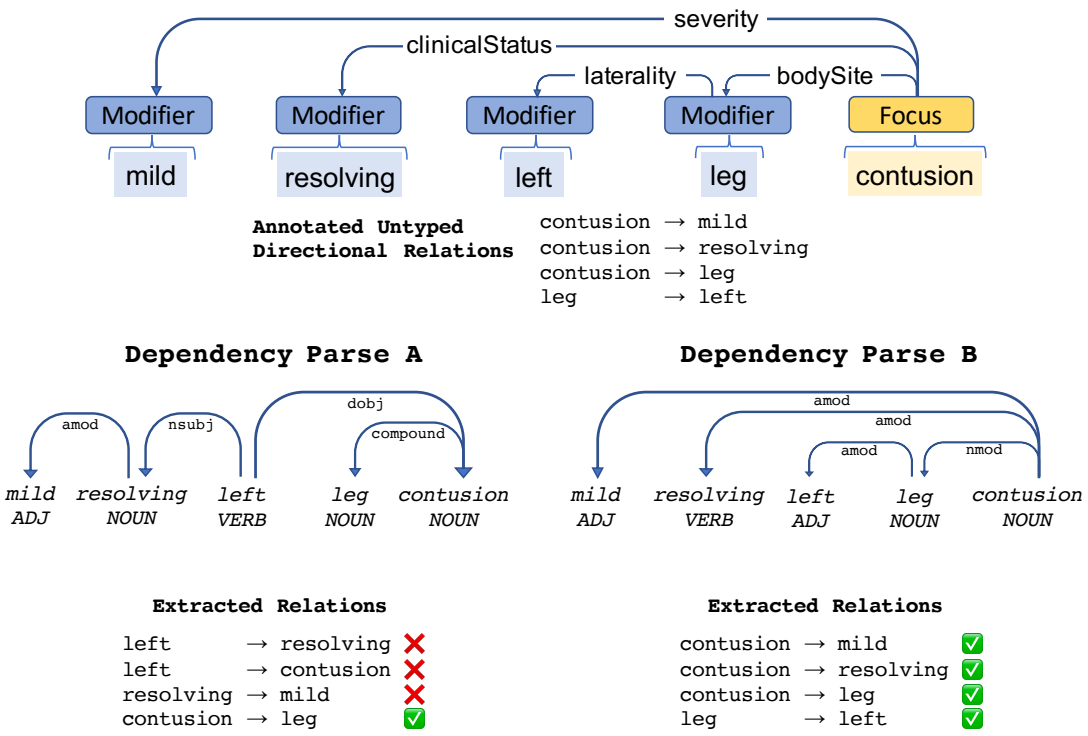


Figure 4.2: An example of the evaluation of an annotated problem description. We evaluate the ability of our dependency parsing model to learn the correct (untyped) relationships. When compared to the human-annotated example (top), **Dependency Parse A** reflects poor alignment, while **Dependency Parse B** corresponds completely.

the focus concept of the problem description. We tested this by running our dependency parse model on each problem description in the evaluation set and measuring the accuracy with which the dependency parse `ROOT` dependency corresponds to the human-annotated focus.

We further hypothesize that performance for both tasks above will increase as the dependency parse model is increasingly tuned to the domain. To test this, we ran the above evaluations using four dependency parse models. First, we evaluated two unmodified pre-trained models: (1) the Default spaCy English model, and (2) ScispaCy Biomedical, a spaCy model specifically trained on biomedical data sets.[115] Next, we evaluated two fine-tuned models as described in the Dependency Parsing step: (1) ScispaCy Biomedical fine-tuned with 141 annotated dependency parses from a random set of problem descriptions, and (2) that same fine-tuned model plus data augmentation.

Relation Classification & Data Programming

To test the ability of our framework to determine the correct semantic relationship type between entities, we next evaluated the performance of our relation classification model. We specifically tested whether or not the data programming approach can effectively be used to train a neural network model. First, we evaluated the performance of our data programming rule-based model on the test set. Next, we trained the neural network classifier via data generated from the data programming model. Finally, we compared the performance of the two models. We hypothesize that the neural network model will have better performance than the rule-based model.

4.4 Results

The gold standard annotation of the evaluation set of six hundred problem descriptions by the three annotators resulted in 1553 relationship annotations and 2057 focal concept/modifier entity annotations. We recorded Krippendorff's alpha inter-annotator

agreement scores of 0.79 for the relationships and 0.94 for the focal concepts. In the case of annotator disagreement, simple majority vote was used to adjudicate. As a result of class imbalance in the evaluation set, results from any relationships with less than fifteen supporting evaluation annotations are not reported in this study.

Table 4.2: Evaluation results from the Focus Concept Selection subtask.

model	precision	recall	f1-score
ScispaCy Biomedical + fine-tuning + data augmentation	0.95	0.94	0.94
ScispaCy Biomedical + fine-tuning	0.96	0.94	0.95
ScispaCy Biomedical	0.70	0.68	0.69
Default spaCy English (baseline)	0.68	0.66	0.67

Table 4.3: Evaluation results from the Untyped Directed Relation Extraction subtask.

model	precision	recall	f1-score
ScispaCy Biomedical + fine-tuning + data augmentation	0.89	0.90	0.90
ScispaCy Biomedical + fine-tuning	0.88	0.91	0.89
ScispaCy Biomedical	0.84	0.70	0.76
Default spaCy English (baseline)	0.84	0.65	0.73

Tables 4.2 & 4.3 show the results of the Focus Concept Selection and Untyped Directed Relation Extraction subtasks. Both focus concept and untyped directed relation F_1 scores are the highest when using the domain-specific fine-tuned dependency parse model. Note that while fine-tuning resulted in a large performance boost, data augmentation had little if any positive performance impact.

Table 4.4: Relation classification results of the neural network model trained via data programming.

relation type	precision	recall	f1-score	# annotations
laterality	0.99	0.99	0.99	167
anatomicalDirection	0.96	0.96	0.96	25
interpretation	0.94	0.94	0.94	18
historical	0.97	0.90	0.94	81
bodySite	0.98	0.90	0.94	235
certainty	1.00	0.87	0.93	45
severity	1.00	0.80	0.89	41
stage	0.95	0.83	0.88	23
course	0.76	0.85	0.80	41
occurredFollowing	1.00	0.66	0.79	29
dueTo	0.89	0.66	0.76	38
clinicalStatus	1.00	0.47	0.64	68
associatedSignAndSymptom	0.50	0.81	0.62	43
—	—	—	—	—
micro avg	0.92	0.85	0.89	854
macro avg	0.92	0.82	0.85	854
weighted avg	0.94	0.85	0.89	854

Table 4.4 shows the F_1 scores for the neural network relationship classification model for all relationship types with more than fifteen annotated relationships in the test set.

Figure 4.3 contrasts the F_1 scores of the trained neural network model as compared to the data programming rule-based model used to create the training data. This figure highlights the amount of improvement gained via data programming when using the rule-based model as a baseline.

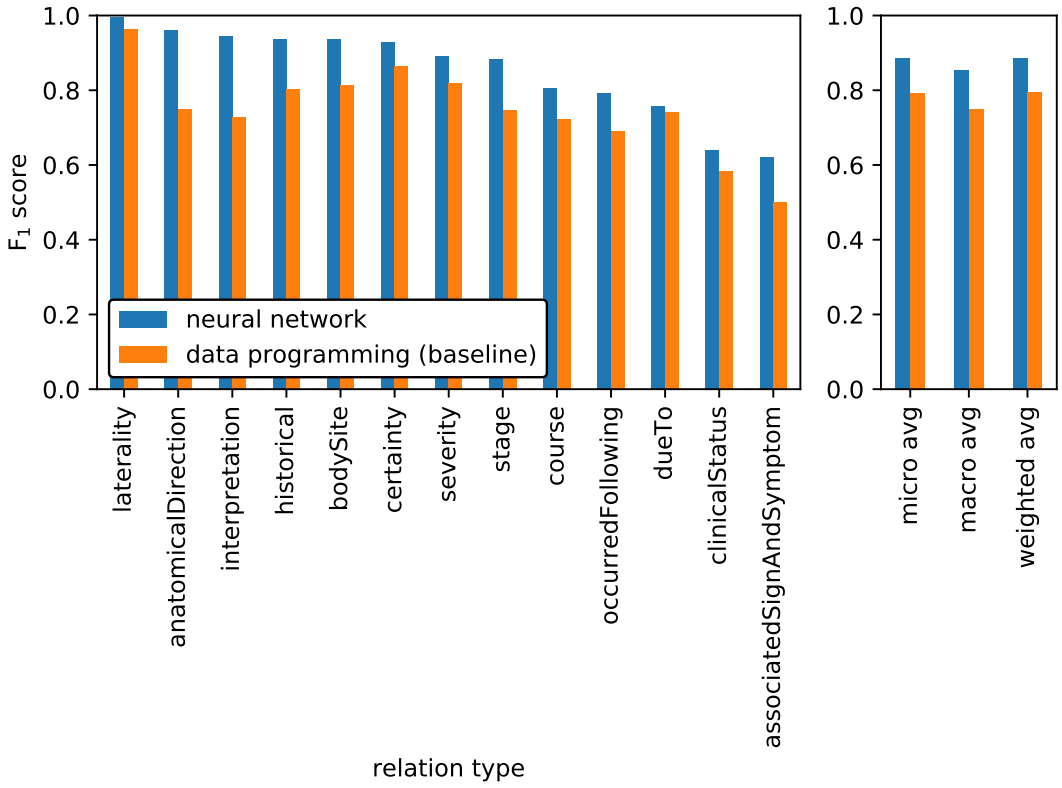


Figure 4.3: Relation classification results compared to the rule-based data programming baseline model.

Table 4.5: Shapley values for the nine features input into the relation classifier.

	Shapley value	% total contribution
Target Vector (BERT)	0.715	78.71
Dependency Parse Shortest Path	0.1057	11.64
Target Vector (cui2vec)	0.0634	6.98
Source Vector (BERT)	0.0163	1.8
Target Semantic Type	0.0041	0.45
Target Semantic Type Group	0.0013	0.14
Source Vector (cui2vec)	0.0012	0.13
Source Semantic Type	0.0009	0.1
Source Semantic Type Group	0.0004	0.05

Table 4.5 shows the Shapley values for the nine relation classification model features. BERT vectors are shown to have the most impact, and features of the target entity contribute more to the model than the source entity. Shapley values for each of the individual relationships under test are shown in Figure 4.4. While BERT vectors are prominent, there are some differences to be noted in feature importance across classes – notably, that the dependency parse shortest path contributes almost exclusively to two relationship classes and relatively little to others.

4.5 Discussion

The use of dependency parse-based methods for finding the focus concept and untyped entity relations of problem descriptions was an effective approach, as shown by Tables 4.2 & 4.3. Furthermore, these tables show that performance was significantly increased by fine-tuning the pre-trained ScispaCy Biomedical parsing model. This reinforces our first methodological focus of emphasizing transfer learning and fine-tuning, as a significant

Figure 4.4: Contrasting the Shapley values for the nine source and target entity features of the relation classifier for each of the evaluated relationship types. Note that negative Shapley values indicate that the feature had a detrimental contribution.

increase in performance was achieved with a relatively small cost of manual training data annotation. Conversely, our data augmentation algorithm did not yield a noticeable change in performance. Kobayashi also reported minimal improvement with a similar non-synonym word replacement augmentation technique,[159] leading us to conclude that more exploration is needed to determine if data augmentation can be successfully applied to this task.

Given the extracted untyped entity relationships, Table 4.4 shows that our neural network model was able to classify the correct relationship type with an overall 0.89 weighted average F_1 score. Although performance on several relationship types surpassed a 0.9 F_1 score, some variation in performance across different types is noted. Specifically, the classifier struggled with the more semantically open-ended relationship types such as ***associatedSignAndSymptom***.

Figure 4.3 shows that data programming can be an effective technique for augmenting a rule-based approach, as the neural network classifier was able to outperform the data programming rule-based classifier. This pairing of a rule-based system with a neural network model eliminated the need for creating a human-annotated training data set, a significant savings of time and effort. It also adds evidence that our second methodological focus of incorporating rule-based methods and knowledge bases is both an effective and pragmatic technique for this task.

Figure 4.4 and Table 4.5 show that BERT features dominate the Shapley value analysis of the system. It is of interest to note, however, that for two attributes ***dueTo*** and ***occurredFollowing***, the "Dependency Parse Shortest Path" feature dominates, as shown in Figure 4.4. Qualitative analysis of the results shows that these two relationship types generally have indicative words between the two entities, for example "right-sided $[\text{CHF}]_{\text{source}}$ ***caused by*** chronic $[\text{pulmonary embolism}]_{\text{target}}$" and "chronic low thoracic $[\text{pain}]_{\text{source}}$ ***after*** $[\text{fall}]_{\text{target}}$". This also reinforces the shortest path hypothesis[163] in that the words between source and target entities in the dependency parse tree primarily contribute to their relationship type. Also of note is the relatively small importance of

the cui2vec vectors when compared to the BERT representations. This observation is in line with similar findings of Kearns et al.[191] Given our last methodological focus of explainability, we can use these insights in the future to improve our model. For example, we may use the findings in Table 4.5 to justify the removal of low-impact features, simplifying our architecture. More importantly, we know that feature importance in our model is not evenly distributed between relationship types – some features such as "Dependency Parse Shortest Path" may have low average overall impact but are critical to certain relationships. This will be an important consideration as we expand our model to different relationship types.

From an implementation perspective, the methods described in this framework are intended to be generalizable to any data set of clinical problem descriptions. Given the data programming approach, large-scale annotation of training of data is not necessary, but adaptation to a particular context or data set does include the following steps:

- **Implementation of labeling functions.** Our data programming approach is heavily dependent on accurate labeling functions to produce the training data set. Implementers of this framework should expect to create a set of labeling functions using rules or heuristics specific to their data.
- **Fine-tuning the dependency parse model.** An accurate dependency parse model is an important facet of our approach. While Tables 4.2 & 4.3 show that reasonable performance can be obtained using freely available pre-trained models, at least a small amount of fine-tuning is recommended to account for data set specific variations in problem description phrasing or structure.

4.6 Conclusion

In this study we have described a framework for standardizing free-text clinical problem descriptions using HL7 FHIR. We have demonstrated that by leveraging domain-specific knowledge bases and rules, we were able to combine data programming and

neural networks to achieve higher performance than via a rule-based approach alone, all while minimizing the need for human-annotated training data. We also examined the feature set of our model and found that BERT language representations contribute significantly more to model performance compared to cui2vec's concept-based vectors. These methods ultimately allow for the alignment of free-text clinical problems into the HL7 FHIR `Condition` resource. All source code for this framework is available via https://github.com/OHNLP/clinical-problem-standardization.

Acknowledgments

This study was funded by grant NCATS U01TR02062. We thank Sunyang Fu, Donna Ihrke, and Luke Carlson for assisting in the test corpus annotation.

Chapter 5

Organizing FHIR Profile Value Sets using Containment Hierarchies and Similarity Clusters

Kevin J Peterson, Guoqian Jiang, Scott M Brue, Feichen Shen, and Hongfang Liu. Mining hierarchies and similarity clusters from value set repositories. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1372. American Medical Informatics Association, 2017

Abstract

A value set is a collection of permissible values used to describe a specific conceptual domain for a given purpose. By helping to establish a shared semantic understanding across use cases, these artifacts are important enablers of interoperability and data standardization. As the size of repositories cataloging these value sets expand, knowledge management challenges become more pronounced. Specifically, discovering value sets applicable to a given use case may be challenging in a large repository. In this study, we describe methods to extract implicit relationships between value sets, and utilize these relationships to overlay organizational structure onto value set repositories. We successfully extract two different structurings, hierarchy and clustering, and show how tooling can leverage these structures to enable more effective value set discovery. Although our ultimate goal is to organize value sets of Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) Profiles, we also evaluate our methods on value sets taken from a large public repository to ensure broad applicability.

5.1 Introduction

Controlled terminologies are the semantic underpinnings of clinical data and facilitate interoperability,[193] research,[194] and quality reporting.[195] These terminologies are shared knowledge assets, often designed to be used for a variety of purposes and use cases.[20,196] A *value set* is a grouping of codes from one or more terminologies used to express some conceptual domain.[197,198] Value sets narrow the broad semantics of controlled terminologies down to a more targeted domain, and have a variety of practical applications related to data standardization and analysis.[199–201] Like controlled terminologies, value sets are shared assets, generally published in a repository or catalog to promote discoverability and reuse.

As a value set is intended to semantically express some domain of interest, it is important that it adequately convey to the consumer the semantics to which it is bound

– or its *intension*. If this intension is not clearly stated, value set *discoverability*, or a user's ability to locate and reuse value sets relevant to their use case, will suffer. As reuse becomes more difficult, users are increasingly likely to simply create a new value set according to their needs. This compounds the problem, resulting in the proliferation of many value sets that are conceptually similar but with slightly different sets of codes. Consequently, it then becomes difficult for a user to infer whether or not these subtle differences were intended and necessary, or simply a result of different authors interpreting the same conceptual space differently. At the extreme, there may also be cases where value sets are inadvertently duplicated.[202] High redundancy may be a symptom of low reusability – a problem not dissimilar to challenges in software reuse.[203]

So how, then, do we determine the intension of a value set in order to promote reuse? If left as an exercise for the repository user, the most accessible points of inspection are the value set name and its set of codes. Although generally intended to be descriptive, a text-based value set name places the burden of interpretation on each individual user. Moreover, the name may lack precision, as names have shown to be generally insufficient at expressing complex semantics.[204] Manual review of the contained code set may provide a better representation of actual intension, but is not without its own challenges. By browsing the code set, a subject matter expert may be able to sufficiently reverse-engineer the value set intension, but this is a manual, potentially laborious process. Furthermore, value set repositories tend to organize content in flat (or nearly flat) structures. This makes it difficult to display and search for similar groups of value sets, or to scope searches to a specific context. More importantly, however, it keeps implicit relationships between value sets hidden, and awareness of these inferred connections is a valuable tool for discovery.[205]

The primiary intented application for these techniques is to organize value sets of Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR)[11] Profiles. FHIR Profiles are organaizational structures that are applied to base FHIR resources to constrain and scope semantics and structure. Value sets are integral parts of FHIR

Profiles as they can be used to scope an attribute or extension to a particular set of concepts. This is important for interoperability purposes as it allows consumers to anticipate the semantics (i.e. codes and code systems) that they must be prepared to process for a given Profile. We expect that, much like general value set repositories, large collections of FHIR Profiles will contain many similar value sets. It is important to note, however, that we intend our methods to be applicable to any value set repository, not only value sets from FHIR Profiles. As such, our methods and evaluation will include value sets from other sources as well.

The aim of this study is to utilize automated methods of capturing value set intension[202] to extract additional structure and knowledge from a value set repository. We specifically aim to extract value set *clusters* based on similarity, and *hierarchies* based on specialization/generalization. By placing value sets in the context of similar ones, we aim to promote better repository search capability,[206] and ultimately, better discoverability. Finally, we will show how these methods may be practically implemented and integrated into tooling.

5.2 Materials and Methods

We examine this value set organizational strategy in the context of two value set sources. First, we apply these methods to value sets to the Value Set Authority Center (VSAC), a National Library of Medicine (NLM) public repository for value sets referenced in Meaningful Use Clinical Quality Measures (CQMs).[45] The VSAC also includes support for value set authoring, expanding its scope beyond Meaningful Use and into other applications and domains.[207] We use this rich value set repository as a source of curated, standardized value sets. Next, we extract value sets via data-driven transformations of free-text clinical problems into HL7 FHIR `Condition` resources from a large clinical corpus (see Chapter 4 for further information). We aim to use these value sets to demonstrate the applicability of our methods beyond curated repositories.

5.2.1 VSAC Value Set Extraction

We cannot assume universal value set structure, as several standardized formats exist along with many non-standard, proprietary representations. Heterogeneously structured value sets must be consolidated via preprocessing and standardization to allow our methods to have the maximum breadth of applicability. Interacting with the VSAC requires a normalization step before further processing can be done.

The VSAC exposes value sets programmatically via the VSAC v2 REST service,[208] with value sets modeled using Sharing Value Sets (SVS),[209] an Integrating the Healthcare Enterprise® (IHE) initiative[210] focused on value sets and their representation. To allow our methods to be portable to repositories beyond the VSAC, we implemented our algorithms to operate on value sets modeled using Common Terminology Services 2 (CTS2),[211] an Object Management Group® (OMG) and Health Level Seven® (HL7) terminology standard. This necessitates an SVS to CTS2 conversion step in our process for value sets originating from the VSAC (see: https://gist.github.com/cts2/ for CTS2 $\leftrightarrow$ SVS implementation details). Our methods are also specifically scoped to *extensional* value sets, or value sets represented as enumerated lists of codes.[209]

ISO/IEC 11179 is a standard for representing metadata in registries with the goal of promoting standardization and interoperability of data.[212] Part 3 of the specification describes the basic elements of the metadata registry metamodel, and is specifically applicable to our problem space, as it provides structure and formality around our notion of a value set and its contents. For our purposes, we focus on the following elements of the metamodel:

- **Enumerated Concept Domain.** A collection of valid meanings – in this case, a set of explicitly enumerated **Value Meanings**. This aligns with what we refer to in this study as a *value set*.

  Example: VSAC value set **BMI Values** (`2.16.840.1.113883.3.600.1.889`)

- **Value Meaning.** A representation of the semantic intension, or *meaning* of a

value. For our purposes, this is a *code* drawn from a standard terminology.
Example: **ICD10CM** `Z68.1`, *Body mass index (BMI) 19 or less, adult*

An Enumerated Concept Domain is a specialization of a **Concept**, defined as a "unit of knowledge created by a unique combination of characteristics." [212] This is notable for our study, as it allows us to model Enumerated Concept Domain → Enumerated Concept Domain (or, value set → value set) relationships. It is through ISO/IEC 11179 that we base our assumption that value sets are themselves a unit of knowledge, have intrinsic meaning, and can be meaningfully related to other value sets.

5.2.2 FHIR Profile Value Set Extraction

We next focus on value sets derived via the alignment of clinical problems into the FHIR standard. This analysis is a multi-step process, with the first task being the transformation of free-text problem descriptions to FHIR `Condition` resources. This step builds specifically on the output of the FHIR standardization framework proposed in Chapter 4. Once the clinical problems have been represented as FHIR resources, FHIR Profiles were mined from the extracted FHIR `Condition` representations. The details of the steps required to mine FHIR profiles from our corpus are expanded below.

The elicitation of the FHIR Profile value sets was a data-driven process, contrasting the VSAC value sets which originated from a curated repository. To extract these Profiles, first 100,000 free-text clinical problems were randomly selected from our clinical corpus. Methods from Chapter 4 were then used to transform these free-text problem descriptions into FHIR `Condition` resources. Next, all of the transformed resources were grouped by their "code" attribute, or the primary clinical focus of the problem. Then, for each distinct code, its entire grouping was iterated over, with all FHIR attributes (such as "bodySite", "clinicalStatus", etc.[†]) aggregated into value sets. For example, given all FHIR `Condition` resources with a code of "Hypertension," the list of

[†]see https://www.hl7.org/fhir/condition.html for the full set of attributes associated with the FHIR `Condition` resource.

all values of the `stage` attribute were collected. An illustration of this grouping is shown in Figure 5.1. These attribute lists then become the value sets attached to the FHIR Profiles. Note that this analysis applies not only to base FHIR `Condition` attributes, but to all attributes that were unable to be aligned to the base FHIR schema as well. These unmapped attributes will have been aligned to the `Condition` resource as extensions, and are still eligible to have scoping value sets associated with their extension values.

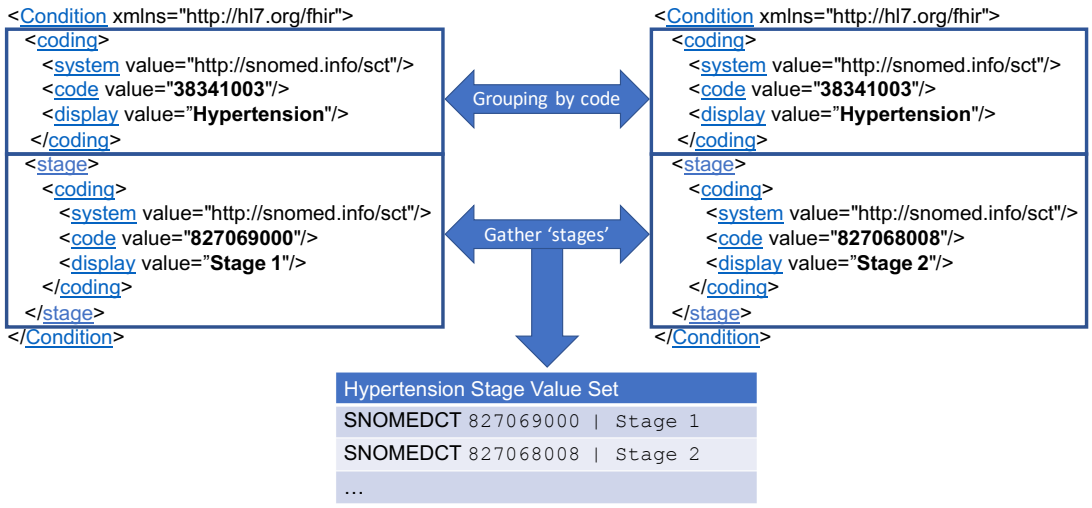


Figure 5.1: An example of FHIR Profile elicitation. In this instance, two FHIR `Condition` resources with a code of "Hypertension" are grouped, and the values of their respective "stage" attributes are merged into a value set.

5.2.3 Algorithm Selection

Our aim is to extract meaningful structure from a value set repository, and we focus our efforts on two structural constructs: hierarchy and clustering. Algorithms for extracting these structures from a repository are described below.

Containment Hierarchy. Hierarchies, or the arrangement of entities into *parent* → *child* relationships, have long been used to structure knowledge artifacts and play

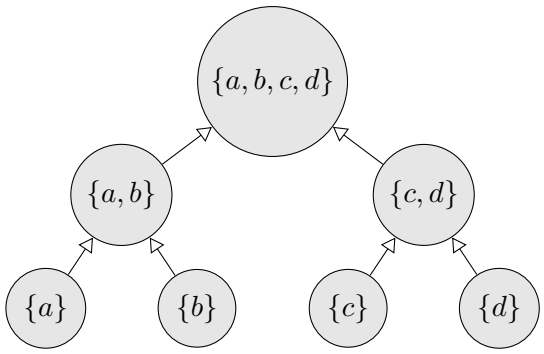
important roles in knowledge discovery.[213] Concept hierarchies have strong mathematical formalizations[214] rooted in concept maps and are an important part of knowledge management.[215] These hierarchies, however, are not always explicitly stated (or even envisioned by the content authors). Text document repositories,[216] Wikipedia entries,[217] text books,[218] and on-line dictionaries[219] are all examples of domains where hierarchies have been *extracted* via mining the existing data sources.

Containment hierarchy is a general strategy for ordering sets based on strict subsets – for example: $\{a, b, c\} \supset \{b, c\} \supset \{c\}$. For our purposes, the sets under examination are the code sets for each value set. By recursively nesting subsets, we can begin to build hierarchies. Examples of this hierarchy can be seen in Figure 5.2.

Clustering. Jansen et al. demonstrated that users searching the web for information tend to use short (2.35 terms) queries.[220] As searching is a fundamental pillar of knowledge management infrastructure,[221] it is useful for us to take this behavioral model into consideration for our study. Specifically, we recognize that a small number of search terms does not provide enough context to sufficiently pinpoint the desired results.[222]

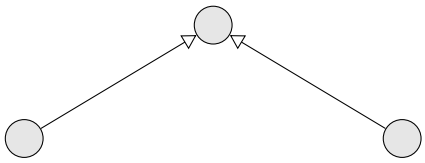
Clustering search results around a conceptual topic has been extensively studied as a way to improve search quality.[223,224] These approaches focus on grouping semantically similar documents together in order to tune the search to the perceived semantic intent of the user. Much like containment hierarchy, clustering aims to extract implicit relationships between value sets. Unlike hierarchy, however, we are not limited to containment relationships. Specifically, we can begin to envision a value set repository as an undirected graph, where the value sets are the nodes and the edges indicate some relationship between them.

Using these relationships, we can begin to analyze groups of closely connected value sets, or *clusters*, using community discovery approaches.[225,226] We define a cluster C focused on a node i in some graph G to be i and all nodes directly connected to it, or its *closed neighborhood*, denoted as $C_i = N_G[i]$. One interesting characteristic of these



(a) Each graph node represents a set of elements organized with children nodes as strict subsets of the parent.

ActStatus
aborted, active, cancelled, completed,
held, new, normal, nullified, obsolete, suspended



ProcedureActStatusCode
aborted, active,
cancelled, completed

ProblemActStatusCode
aborted, active,
completed, suspended

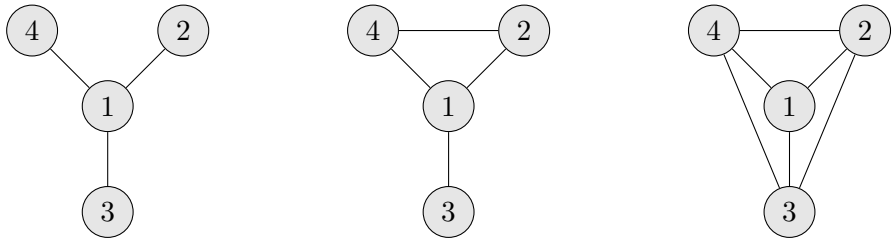
(b) A practical example highlighting three value sets. Here, **ProcedureActStatusCode** and **ProblemActStatusCode** are strict subsets of **ActStatusCode**.

Figure 5.2: Building hierarchical structures with Containment Hierarchy.

clusters is how tightly connected they are – or, how probable it is that the neighbor nodes of i are also connected. Densely connected graphs appear in several contexts, ranging from social media connections[227] to cell biology.[228] We measure the density of these connections by calculating the *clustering coefficient*, a measure of how connected the neighbors of a node are to one another.[229] Given a node i in graph G, let T_i represent the number of edges in G that connect any two neighbors of i, and let K_i represent the *degree*, or number of outgoing edges of i. The clustering coefficient CC_i can then be calculated for any node i :

$$CC_i = \frac{2T_i}{K_i(K_i - 1)}$$

Note that the clustering coefficient may only be calculated for a cluster of three or more nodes. As such, only groups of three or more related value sets were considered to be clusters.



(a) $CC_1 = 0.0$ **(b)** $CC_1 = 0.33$ **(c)** $CC_1 = 1.0$

Figure 5.3: Example Clustering Coefficient (CC_i) values for three sample graphs.

Figure 5.3 shows this calculation as applied to three example graphs. Note that as this measurement is focused around a single given node, it is referred to as the *local* clustering coefficient. We can compute the *global* clustering coefficient $\overline{CC}$ of the entire graph by averaging the clustering coefficients of each node:

$$\overline{CC} = \frac{1}{n}\sum_{i=1}^{n} CC_i$$

Thus far, we have considered the edges in our graph to be binary – either two value

sets are connected or they are not. To properly reflect our domain, we must account not only for the relationship itself, but the intensity of the connection – in our case, the amount of similarity between two value sets. Assuming we have in place algorithms to compute this similarity, we now are able to treat value set similarity not as a binary condition, but as a degree. We reflect this in our graph by using *weighted* edges. With this change we must also utilize a *weighted* clustering coefficient,[230,231] which is an adjustment to our original formula to incorporate edge weights. This is similar to the original clustering coefficient equation, but incorporates edge weights (w) by calculating the geometric mean of the weights of all *triangles* originating from the focus node i. Note that weights were normalized ($\hat{w}$) to a value in the closed interval $[0, 1]$ by dividing the raw weight by the maximum weight found in the graph: $\hat{w}_{k,i} = \frac{w_{k,i}}{max(w)}$.

$$\widehat{CC}_i = \frac{2}{K_i(K_i - 1)} \sum_{k,j} (\hat{w}_{k,i} \hat{w}_{j,i} \hat{w}_{k,j})^{1/3}$$

For each cluster, an accompanying Erdős-Rényi random graph[46] was created, with $\widehat{CC}_i$ calculated similarly for this randomly assembled cluster. If $\widehat{CC}_i$ for the actual cluster was higher than what was observed in the random cluster, we considered that cluster to be *dense*.[229]

5.2.4 Defining Value Set Similarity

Although our methods now account for varying degrees of similarity via weighted edges, the remaining challenge is to define what exactly makes two value sets *similar*. We recognize that value set similarity is context-specific and in many cases subjective. For our purposes, we define value set similarity as a measure of shared intension. We use two approaches to extract the intension of a value set: analysis of the name and code set. These methods, informed by the work of Winnenburg and Bodenreider,[202] are detailed below.

- **Value Set Name.** *Term Identification* is a Natural Language Processing (NLP) technique for extracting concepts from free text and mapping them to controlled vocabularies.[232] MetaMap,[85] an NLP tool developed by the National Library of Medicine (NLM), aims to assign Unified Medical Language System (UMLS)[30] Concept Unique Identifiers (CUIs) to biomedical free text. Passing in the value set name to the MetaMap tool yields a set of CUIs representing the normalized semantics of the name. Note that this metric only applies to VSAC value sets, as data-derived FHIR Profile value sets do not have names.

- **Code Set.** A value set's code set scopes the semantics of the value set by enumerating the permissible values for its domain. Using the code set as a representation of intension has some distinct advantages: (1) it requires no interpretation or extra processing to extract, (2) it is the most direct representation of the author's semantic intent, and (3) it yields a discrete, comparable set of values.

As a result, we represent value set intension as either a set of UMLS CUIs extracted from its name, or its member code set – in either case, a set of discrete values. The *similarity* of two value sets may then be computed as a measure of overlapping intension. Given two value sets, let A and B denote the derived UMLS CUI set or code set from each. Similarity is then defined as the Jaccard index[233] of these two sets, or their intersection size divided by their union size:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The higher the Jaccard index, the more similar we considered two value sets. Conversely, a low Jaccard index indicates dissimilarity, and similarity scores lower than a threshold of 0.25 were not considered in our cluster analysis.

The rationale for using two different similarity strategies is that we cannot necessarily rely on a single measure in isolation. For example, VSAC value sets **BMI Values** (`2.16.840.1.113883.3.600.1.889`) and **BMI values**

(2.16.840.1.113883.3.600.1.888) have names that are semantically identical, but share no common codes, as they draw from two different code systems (ICD10CM and ICD9CM, respectively). Conversely, **Procedures as Reasons for Admission to ICU Due to Pneumonia** (2.16.840.1.113762.1.4.1111.26) and **Need for Ventilator** (2.16.840.1.113762.1.4.1045.82) share the exact same code set while having semantically dissimilar names. As such, we recognize that using different similarity measures will establish potentially different similarity relationships between value sets.

5.2.5 Cluster Similarity

If the two different value set similarity measures relate value sets in different ways, we also expect them to produce different clusters of value sets. It is useful then to introduce a method to calculate just *how* dissimilar these clusters are – or *cluster similarity*. To determine this, we first compute two value set graphs, one with edges weighted by the code set similarity function and the other using value set name similarity. Next, we extract clusters from both graphs pairwise by a common focus node i, yielding clusters C_i and C'_i. Finally, we compute similarity at a cluster level by inspecting how many nodes (or value sets) the clusters have in common by calculating the Jaccard index:

$$\texttt{cluster\_similarity}(i) = J(C_i, C'_i)$$

5.3 Results

First, we analyzed 3820 total value sets from the VSAC repository, focusing on extracting implicit value set to value set relationships. Our methods sought to mine the overarching structures implied by these relationships – specifically hierarchies and clusterings of similar value sets. The metrics gathered were focused on quantifying the characteristics of these structures in an effort to better understand how they may be ultimately leveraged.

Analysis of the value set clustering characteristics gave insight into how interconnected the value sets were – or, how readily they formed into related groups. A cluster was defined as three or more connected value sets, and two sets of clusters were built using the two different similarity measures. Of the 3820 total VSAC value sets analyzed, 3185 clusters were found using the name similarity measurement, while 1190 were found via code set similarity. Table 5.1 reflects the results of the clustering coefficient calculation, a measure of how densely interconnected the clusters were. Valid clustering coefficients range from $0 - 1$, with 0 indicating no neighbor connections and 1 being all neighbors interconnected (or a *clique*). This analysis was computed twice: once for each similarity algorithm, and the subtables reflect the result of both computations.

The clustering coefficient result for each cluster was also compared to a clustering coefficient for a similar, randomly assembled cluster. Through this calculation we were able to obtain the number of *dense* clusters, or clusters observed to have a weighted clustering coefficient greater than would be expected to occur randomly. For clusters grouped by value set name, 3170 out of 3185 clusters were considered dense. For code set groupings, it was 964 out of 1190.

Table 5.1: Clustering summary statistics for VSAC value sets. The clustering coefficient is a measure of how densely value sets tended to group together based on the two similarity measures.

	Mean	Median	Min	Max	Std. Dev.
$\widehat{CC}_i$	0.4067	0.4257	0.0	1.0	0.2834
$\lvert C_i \rvert$	4.9067	4.0	3.0	21.0	2.6021

(a) Code Set Similarity

	Mean	Median	Min	Max	Std. Dev.
$\widehat{CC}_i$	0.586	0.4874	0.0	1.0	0.2847
$\lvert C_i \rvert$	15.1086	8.0	3.0	89.0	16.0366

(b) Name Similarity

$\widehat{CC}_i$: weighted local clustering coefficient
$|C_i|$: number of nodes in a cluster

Relationships between value sets indicated a degree of similarity. Our methods assumed that differences in how we computed value set similarity would result in different relationships, and consequently, different clusters. Table 5.2 shows in summary how similar the clusters from the two similarity measures were. Cluster similarity was computed as the Jaccard index of a cluster from each similarity measure, compared pairwise by a common focus node. For this measurement, the permissible cluster similarity range was $0-1$, with 0 indicating no shared value sets in the clusters (which is impossible, as via pairwise comparison by common focus the clusters will at least share one node), and 1 indicating that the clusters were identical. Analysis of these relationships was also

Table 5.2: Similarity comparison of clusters computed from the two different similarity measures for VSAC value sets. Cluster similarity was calculated by pairwise comparison of clusters produced by both the value set name and code set similarity measures.

	Mean	Median	Min	Max	Std. Dev.
Cluster Similarity	0.2916	0.2143	0.0111	1.0	0.243

leveraged to explore implicit hierarchical structures within the repository. Of the 3820 VSAC value sets analyzed, 1546 were found to be roots, or value sets with no computed hierarchical parents. To measure the extent of the extracted hierarchy, the longest path to a leaf for each root value set was calculated. Figure 5.4 summarizes our findings, reflecting the levels of hierarchies found in the repository, from 0 levels (meaning a value set with no children) to 5, the maximum level observed.

Next we present the results from the FHIR Profile value sets. Figure 5.5 shows the hierarchy levels extracted from four types of value sets extracted from FHIR Profiles, while Table 5.3 presents the clustering coefficients for these value sets. These figures are analogous to the VSAC analysis Figure 5.4 and Table 5.1, respectively.

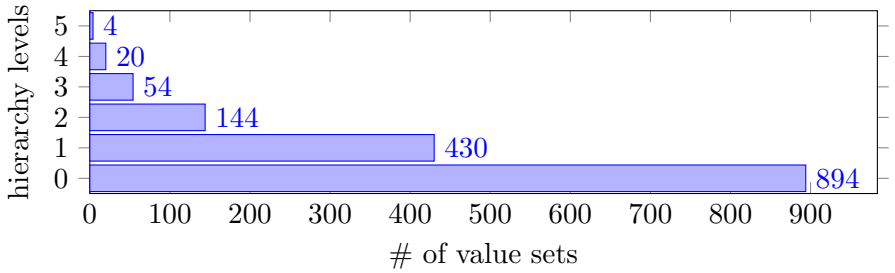


Figure 5.4: Analysis of extracted hierarchy levels for VSAC value sets. For each root value set (or value set with no computed parents), its *height* (or longest path to a leaf) was calculated, indicating the level of hierarchy.

Table 5.3: Clustering summary statistics for FHIR Profile value sets. The clustering coefficient was measured for value sets extracted from two base FHIR `Condition` attributes (`bodySite` and `stage`) along with two FHIR Extensions (`dueTo` and `related`).

	Mean	Median	Min	Max	Std. Dev.
$\widehat{CC}_i$	0.4453	0.4597	0.0	1.0	0.2225
$\lvert C_i \rvert$	10.7289	9.0	3	41	7.1373

(a) `bodySite` (FHIR base attribute) Code Set Similarity

	Mean	Median	Min	Max	Std. Dev.
$\widehat{CC}_i$	0.6215	0.6871	0.0	1.0	0.2246
$\lvert C_i \rvert$	86.5312	18.0	3	178	82.673

(b) `dueTo` (FHIR extension) Code Set Similarity

	Mean	Median	Min	Max	Std. Dev.
$\widehat{CC}_i$	0.3955	0.4114	0.0	1.0	0.1904
$\lvert C_i \rvert$	10.587	6.0	3	67	10.8424

(c) `related` (FHIR extension) Code Set Similarity

	Mean	Median	Min	Max	Std. Dev.
$\widehat{CC}_i$	0.5779	0.5818	0.0	1.0	0.2313
$\lvert C_i \rvert$	35.7653	23.0	3	104	30.0384

(d) `stage` (FHIR base attribute) Code Set Similarity

$\widehat{CC}_i$: weighted local clustering coefficient
$|C_i|$: number of nodes in a cluster

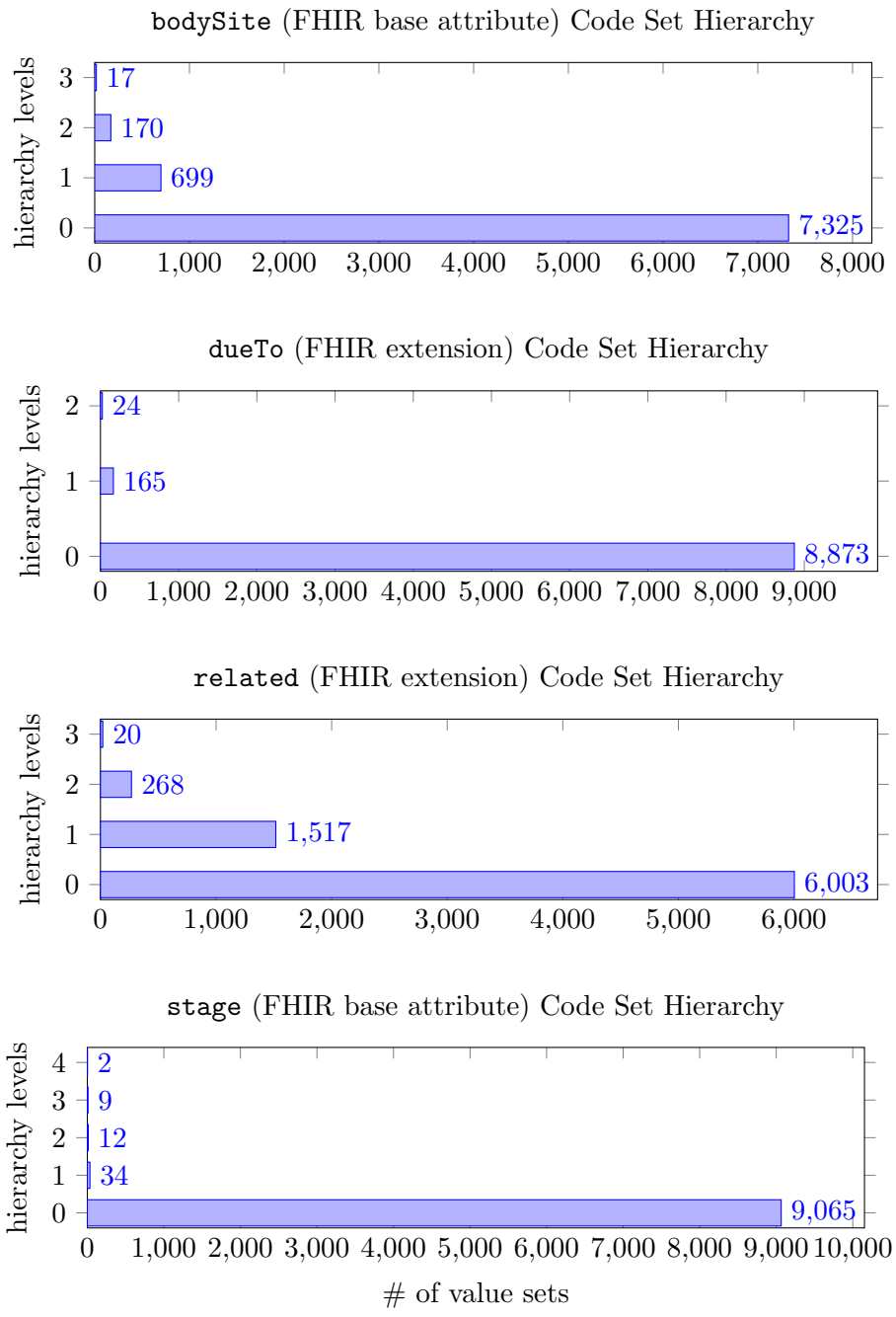


Figure 5.5: Analysis of extracted hierarchy levels for FHIR Profile value sets found for two base FHIR `Condition` attributes (`bodySite` and `stage`) along with two FHIR Extensions (`dueTo` and `related`).

5.4 Discussion

We have described methods to extract two different organization structures from value sets, and we begin our discussion with an examination the structures found, starting with the VSAC repository. Table 5.1 demonstrates high clustering coefficient values, indicating that value set clusters tend to be highly interconnected. We also find that the majority of clusters were *dense*, or exhibited a clustering coefficient value higher than would be expected for a similar random graph. This clustering is especially pronounced in clusters built using the value set name similarity measure, where almost all clusters (99.53%) were dense. This shows that not only are we able to extract clusters of similar value sets from a repository, but that value sets, at least in the VSAC repository, tend to be very closely clustered.

Our ability to infer connections between value sets is predicated on our ability to measure how connected, or *similar* they are. Utilizing two different similarity measures was assumed to produce different similarity connections, and thus, different clusters of related value sets. Table 5.2 shows that on average clusters computed using the two different similarity measures contained roughly 29% similar value sets. This indicates that while different similarity measures do indeed produce measurably different clusters, the clusters do show some degree of congruence. As value set *similarity* itself is inherently subjective and context dependent, we may be able to find underlying trends by looking at where different similarity measures agree.

A hierarchical structure was also successfully extracted from the VSAC repository. Figure 5.4 demonstrates that while the extracted hierarchy is relatively flat on average, some hierarchical structure does exist, and at times, can be as deep as 5 levels (see Figure 5.6). This is certainly an improvement organizationally over a flat list, especially since it required no manual curation or classification.

Similar results were observed for value sets extracted from the mined FHIR Profiles. Figure 5.5 shows that some hierarchical characteristics emerged from the FHIR Profile

value sets, although these hierarchical structures are less pronounced as compared to the VSAC value sets. Table 5.3 does, however, show generally high degrees of clustering for FHIR Profile value sets – with clustering coefficients as high or higher than what was observed for the VSAC.

Our study purpose challenged us not only to extract structure, but to apply it in useful ways. As such, enhancing tooling with improved affordances for knowledge discovery is our end goal. In our previous work, we outlined the architecture for a value set management tool focused on usability.[199] We now leverage this toolset as our primary implementation platform. For hierarchy, Sunburst charts, or visual representations of hierarchy with radiating levels, have shown to be effective representations of hierarchical structures.[234] Figure 5.6 is an implementation of a Sunburst chart displaying an extracted VSAC value set hierarchy.

Clustering can be leveraged to suggest alternatives or possibly related value sets to users as they browse. Figure 5.7 depicts search results from a keyword search. Results found by traditional text-based searching are augmented by using *See Also* suggestions, where other members of the cluster are displayed as possibilities for further exploration.

5.5 Conclusion

In this study we have shown that data mining techniques can be employed to extract implicit relationships between value sets. These relationships in turn can be leveraged to add meaningful structure and organization to a repository. Ultimately, we show that these structures have practical applications in user-facing tooling, enhancing users' ability to discover the correct value set for their use case, and thus, increasing value set reuse and general repository utility. Meaningful organiazation emerged from both the VSAC and FHIR Profiles value sets, indicating that these organizational techniques can be applied to both curated repositories and value sets extracted automatically via data-driven techniques.

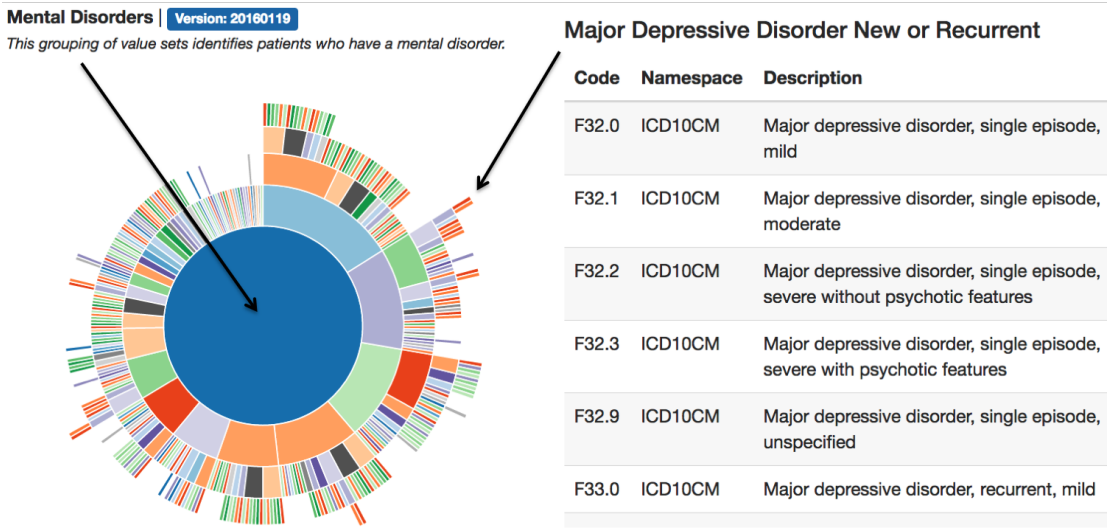


Figure 5.6: A Sunburst chart representing a value set hierarchy. Levels of hierarchy are represented as concentric circles radiating outward from a focus. In this example, the focus value set **Mental Disorders** is surrounded by radiating levels of hierarchy (shown left). On mouse-over, details of the value set **Major Depressive Disorder New or Recurrent** are displayed (shown right). The hierarchical path between the two is: **Mental Disorders → Mental Disorders ICD10CM → Mental Health Diagnoses → BH Condition involving unipolar depression ICD-10-CM → Major Depression → Major Depressive Disorder New or Recurrent**.

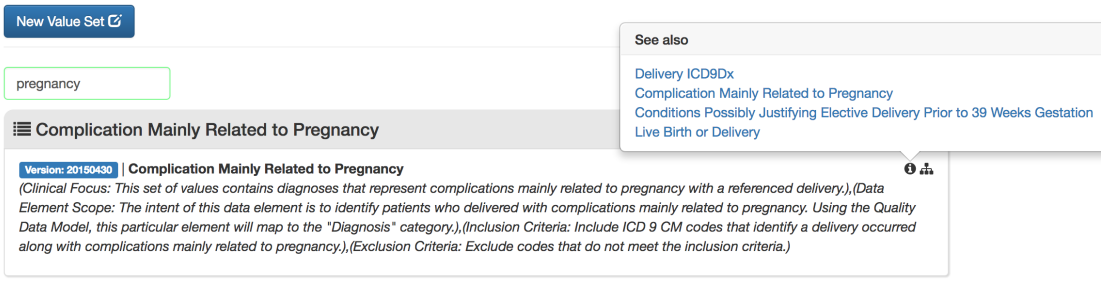


Figure 5.7: Showing *See Also* suggestions on search to suggest similar value sets by using value set clustering.

5.5.1 Limitations and Future Work

A limitation of our analysis is that hierarchy calculated by set containment algorithms may not always reflect the actual semantic hierarchies. There are certainly instances where a value set may contain a strict subset of codes from another but may not be a logical child, depending on context. Figure 5.8 illustrates one such example. In this case, the codes of **Medication Fill Status** are a strict subset of the codes of **ProcedureActStatusCode** and **ProblemActStatusCode**. It may be illogical (or at least context-dependent) to state that these value sets share a semantic relationship.

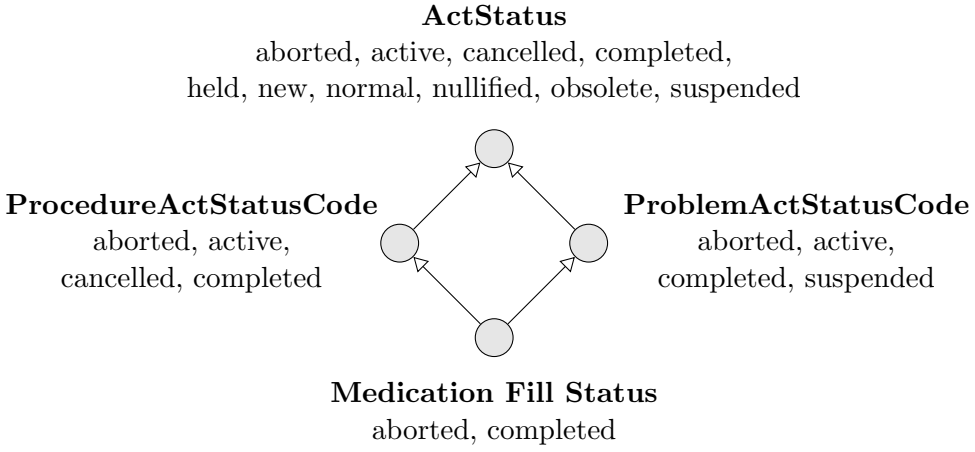


Figure 5.8: Examining a Containment Hierarchy calculation for semantic correctness. Here, **Medication Fill Status** is calculated to have a hierarchical relationship with both **ProcedureActStatusCode** and **ProblemActStatusCode**.

Future directions may include further integration of ISO/IEC 11179. Specifically, we look to consider data elements as a measure of value set similarity, such as grouping value sets by the data they describe (for example, grouping by object class or property).

Acknowledgments

This study was supported in part by the LHSNet project (CDRN-1501-26638), the caCDE-QA project (U01 CA180940), and the BD2KonFHIR project (U01 HG009450). We also acknowledge Mayo Clinic's Knowledge Management and Delivery effort, as well as the Rochester Epidemiology Project.

Chapter 6

Diachronic Language Change in Free-Text Clinical Problems

Kevin J Peterson and Hongfang Liu. An examination of the statistical laws of semantic change in clinical notes. *AMIA Informatics Summit Proceedings*, 2021. (in press)

Abstract

Evolving language patterns may impact the performance of clinical Natural Language Processing (NLP) systems. In this study, we propose methods to detect linguistic change in clinical diagnosis text over time. Using four different methods of change quantification, we find that clinical diagnoses do exhibit gradual year-by-year change when tested against a large clinical corpus. We also find our methods can detect a sudden shift in language caused by a single event – in our case, an Electronic Health Record implementation change. We intend this work to be used as a template for implementing a robust clinical language change monitoring strategy for deployed NLP systems.

6.1 Introduction

Even with the increasing digitization of the medical record by modern Electronic Health Records (EHRs), a majority of patient information is still encoded using unstructured, free-text forms.[236,237] This clinical narrative is the mechanism through which the patient "story" is conveyed,[238] providing background and context for patients' pertinent health issues. The linguistic characteristics of this narrative are varied and diverse,[239] and effective use of language in the clinical setting can be a contributing factor to clinical outcomes.[240] As such, understanding the linguistic characteristics of clinical language is an important part of understanding how information is communicated in the clinical domain as a whole.[53]

The language used to describe clinical care is not static, however. Human language is continually evolving,[241] and the observable linguistic differences over time are known as the "diachronic change" of language. In the biomedical domain, given the prominence of free-text, diachronic change is an important consideration for the application of Natural Language Processing (NLP) techniques. Robust NLP systems must be able to detect changes in grammar, syntax, or semantics in order to adapt to changing medical practices.[48]

The goal of this study is to analyze the linguistic changes in clinical language over time. Specifically, we examine the change over time of free-text clinical diagnosis statements. We explore several methods for quantifying language change that can be used to construct a robust representation of temporal linguistic differences. We apply our methods to a large clinical notes corpus spanning seventeen years and examine them using a case study of a current Mayo Clinic Clinical Decision Support (CDS) project.

6.2 Background & Significance

There are several broad approaches to quantifying diachronic language change. One direction is to create mathematical representations of language, or *language models*,

and quantitatively compare them. With this technique, linguistic differences over time are quantified by using a language model trained at a particular point in time to predict some text in the future.[242] To evaluate the language model differences, an information-theoretical approach can be taken to look at the relative entropy of two language models over some time span.[243]

Another approach is to focus more on the semantics, or how the meaning of individual words change. This approach generally focuses on words in isolation and how the changes in word senses can be shown to drift over time. This analysis may be conducted in a variety of ways, including using Baysian techniques[244] and more recently neural language models,[245] and may consider aspects such as changes in frequency of word usage or change in part of speech.[246]

Regardless of how diachronic change is ultimately quantified, not taking general language changes into account when processing text spanning time periods can lead to incorrect or invalid conclusions.[27] Language change must be actively accounted for in deployed information systems in order to avoid performance degradation of fundamental downstream NLP tasks.[28] It has also been anticipated that the increased pace of digitization of older data will exacerbate the problem, forcing existing NLP systems to adopt data-driven approaches to solving the problem of normalizing outdated language, as opposed to manual or one-off intervention by subject matter experts.[247,248]

In spite of the potential impact to clinical NLP systems, there are few instances of these methods being applied to the domain of clinical diagnoses. One recent work proposes methods to track disease change via Wikipedia articles,[249] which attempts to quantify changes in disease meaning using edit activity on public Wikipedia pages. Other studies explore the issue through their impact on publicity available clinical terminologies,[250,251] showing that without effort, these resources can become out of date or fail to keep pace with quickly emerging diseases or treatments. To account for this drift, these terminologies often respond to change through reactive or "top-down" means that

may lead to ambiguities that are subtle and difficult to detect.[252,253] Applying data-driven techniques to quantifying these changes would help the domain be more proactive in terms of adjusting to this drift.

For this work, the motivation for quantifying language change in clinical problems is pragmatic: understanding the characteristics of this change will allow us to plan appropriately for changes to running clinical NLP systems. Sufficient monitoring may even allow NLP systems, such as the standardization framework proposed in this dissertation, to become more robust to this drift, as there is evidence that even a small-to-moderate effort to account for language change has a positive impact to downstream NLP tasks.[254] In effect, a robust quantification of the change can lead not only to improved detection and monitoring, but to normalization efforts such as reconciling differences in spelling[255] or vocabulary[256] over time.

6.3 Methods

All of our methods utilize a large clinical notes corpus including approximately 50,000 patients over seventeen years. Notes were drawn from the years 2000 to 2017 for consistency, as in 2018 a new EHR implementation significantly changed note structure. Roughly seven million total clinical notes in total are included in the corpus. From these notes we extract text from the following sections for analysis in this study:

- **Diagnosis.** The summary-level description of the clinical problem. These are terse representations of the diagnosis entered as free-text by the clinician. They are often represented as complex noun phrases as opposed to full narrative text.[58]

- **Impression, Report, and Plan (IRP).** The narrative section of the clinical note that further describes the specific diagnosis. This section generally provides much more context and detail regarding the diagnosis, along with current or proposed treatment plans.

These two sections were chosen for two main reasons. First, we are able to draw a one-to-one relation between the two sections - meaning, for each diagnosis there is an attendant IRP section in the note. Second, we expect that the semantics of these sections are partially linked – or, that the interpretation of the text in the diagnosis section is meant, at least in part, to be understood through the extra detail and context described in the IRP section. For example, the term "hypertension" in the diagnosis section is in part defined by its intrinsic meaning as a well-known medical term, but a significant portion of its clinical meaning can only be understood through interpretation of the context and details noted in its associated IRP section.

Text from these two sections was then stratified by years to allow for temporal processing. Our corpus was segmented by grouping all clinical notes by the calendar year in which they were originally composed. We generally quantify language change as a function of "year distance," or the relative amount of time in years between two given year groups. Year distance is considered the independent variable in all our language change analysis, while the dependent variable will correspond to one of several specific facets of change we explored, which are described below.

Notes were randomly selected across all specialties and departments. Our clinical notes corpus includes over one hundred types of notes, including progress notes, evaluations, discharge summaries, and so on. Although there are many possible types, the bulk of our corpus is made up of a limited set. Figure 6.1 shows the total distribution of note types (limited to the top ten), and Figure 6.2 shows the distribution by year.

The note type abbreviations for the ten most frequent note types found in our corpus (shown in Figures 6.1 & 6.2) are defined below:

SV	Subsequent Visit	SUM	Dismissal Summary
LE	Limited Exam	MIS	Miscellaneous
CON	Consult	DCS	Discharge Summary
ME	Multi-system Evaluation	SUP	Supervisory
TOM	Test-Oriented Miscellaneous	ADM	Hospital Admission Note

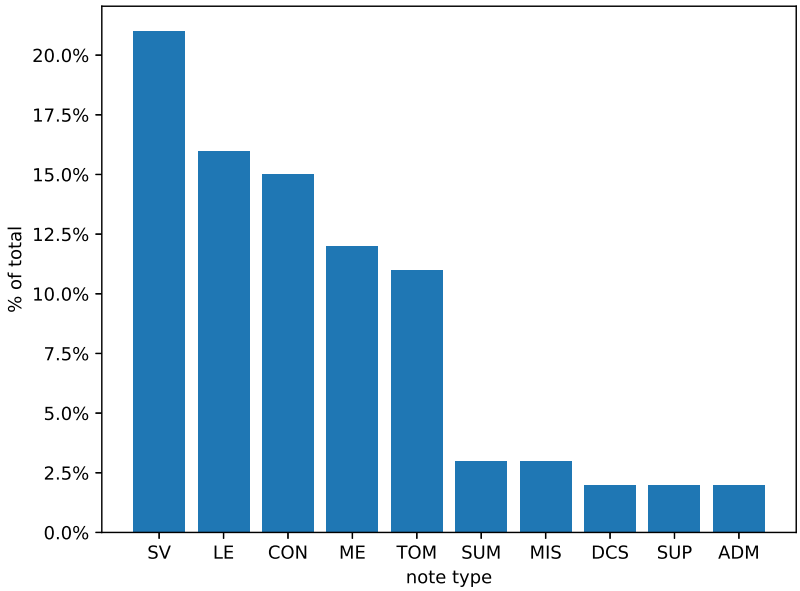


Figure 6.1: The total distribution of the top ten most frequent note types drawn from our clinical corpus.

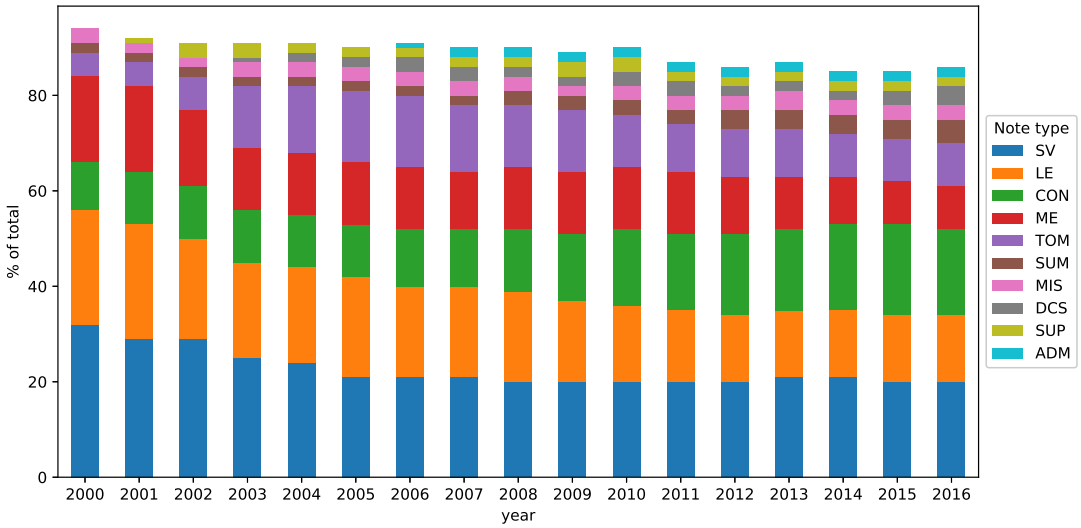


Figure 6.2: The by-year distribution of the top ten most frequent note types drawn from our clinical corpus.

Our analysis of diachronic change in clinical diagnosis statements is separated into several main facets where we explore the change from multiple directions. First, in Section 6.3.1 we take an information-theoretical approach where language model entropy is compared across years. Next, Section 6.3.2 explores the semantic change of words over time by comparing word vector representations and the movement of these vectors over time. We then explore how changing terminologies and vocabularies within the domain impact codification of clinical diagnosis statements. This analysis, detailed in Section 6.3.3, explores how changing concept definitions in the publicly available biomedical terminologies can cause changes to how these statements are coded. In Section 6.3.4 we explore beyond the summary-level clinical diagnosis statement and into the larger narrative clinical note. The intent here is to look at the context around the diagnosis, specifically the IRP section, to explore the "pragmatic" change in semantics of the diagnosis. We conclude with a case study in Section 6.3.5, where our methods are applied to a running clinical decision support system for two diseases of interest.

6.3.1 Language Model Perplexity

A language model is a statistical method used to model the probability of sequences that can be composed given the vocabulary of a language. It can be used to capture the probability distribution of sequences of words in a corpus, or $P(w_1, ..., w_n)$, given some target word w_n and all preceding words. N-grams, or strings of words of length n, are commonly used as the basis for language modeling.[257] By using *bi*grams - or n-grams of length 2, we can approximate this probability calculation by only considering a word and its immediately preceding word, or:

$$P(w_1, ..., w_n) \approx \prod_{i=1}^{n} P(w_i|w_{i-1})$$

where this simplification can be made under a Markov assumption for our model – or, that the probability distribution for a word in a sequence is conditioned solely on the

previous word.[258]

Ultimately, these models give language a mathematical foundation through which linguistic characteristics can be quantified. A consequence of this is that two or more language models can be effectively compared using an information theory-based approach using model *perplexity*.[49,50] Perplexity is the measure of how well a language model predicts some unseen corpus of words W, or $PP(W)$. It can be thought of as a restatement of the entropy of the language model such that:

$$PP(W) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i|w_{i-1})}}$$

and can help us to quantify how well an existing language model "fits" some unseen corpus of text. We used the Python Natural Language Toolkit (NLTK)[259] to conduct this analysis, leveraging its built-in Kneser-Ney implementation for smoothing.[260]

For our purposes, we intend to use language models and perplexity to quantify changes in language over time. To do this, we first train two language models for each year of our corpus – one for the diagnosis section and one for the IRP section. For each year's trained language model, we test it on text from each subsequent year and measure the resulting perplexity. Through this technique we wish to learn how well the language model of a given year can predict language of future years. Our hypothesis is that a language model will have increasing difficulty when predicting text from future years. In other words, we hypothesize that language is in fact continually changing, and thus, the older the language model, the more difficulty it will have with contemporary language patterns. The null hypothesis is that all language models will predict text from all years with equivalent perplexity. This would mean that a language model trained on any year could correctly model any other year with no degradation in quality of representation – implying that the language over time is remaining static.

We ran the perplexity analysis in two phases: once using the full corpus, and once for the text of each note type individually (See Figures 6.1 & 6.2). This was done to

account for variable distributions of note types over the years. We hypothesize that any general perplexity trend noted during the analysis of the full corpus will also be seen for each individual note type.

6.3.2 Embedding Drift

Word embeddings are transformations of discrete words into a continuous vector space. Embeddings allow words to be compared mathematically via cosine similarity or other measures. Many word embedding techniques are rooted in the distributional hypothesis,[261] or that semantically similar words will be found in similar contexts. In terms of word embeddings, distributional semantics can be leveraged to ensure semantically similar words will be closer together in the vector space. This section focuses on how word embeddings can be used to detect changing word semantics over time.

Word2vec,[51] a prominent unsupervised machine learning method for word embedding, uses neural networks to map words to vectors, and is the embedding model used for this study. Word2vec represents each word as a single, high-dimensional vector, regardless of the number of senses the word may carry. As a consequence, polysemous words are represented as a single vector that is either skewed toward one sense used predominately in the corpus, or some aggregation of the multiple sense vectors.[262] This is usually seen as a disadvantage and detrimental to downstream tasks – as such, several techniques have been introduced to allow for sense-specific embeddings.[263–265] For our purposes, however, having one vector per word (as opposed to one per word sense) is desirable, as changes to the vector of a polysemous word can show average movement to or from certain senses over time – a facet of change we wish to capture.

For a given word, if no change in meaning or semantics happens over time, we assume the word2vec embedding vector will also remain constant. This also applies to polysemous words – the vector representing the composition of the various senses of a word should remain fixed. We would, however, expect the vector to change in two circumstances. First, if the meaning of a word changes to something new over time.

This would cause the vector to shift to a new portion of the vector space. Second, if a polysemous word undergoes shift to or from one or more of the senses. This would cause the vector to move away from the less-used senses and toward the preferred usage. Either way, leveraging the fact that word2vec only allows for one vector per word, we will be able to detect either of these two scenarios.

To see these vector changes over time, we first need to train a separate word2vec model for every time period under study – in our case, for every year. There is an issue regarding this technique, however. Two word2vec models cannot be directly compared, as vectors in any two word2vec models are subject to randomizations within the training algorithm and will not align, even if the training corpus and hyperparameters are held constant.[266] We assume, however, that even though absolute vector positioning between models cannot be compared, vectors do maintain their relative position as compared to other vectors.[246] This means that a linear transformation applied to one vector space could be used to align it to another space. As such, a common solution to the problem is to learn a linear transformation that maps each vector in a source vector space to a target space.[267,268]

This transformation can be accomplished through an orthogonal Procrustes matrix. Using this method, an orthogonal matrix is learned such that the sum of squares of word vector distances from one vector space to another is minimized.[269] The following equation represents this transformation:

$$R = \underset{\Omega:\ \Omega^T\Omega=I}{\arg\min} \sum_{w \in V} ||\Omega \mathbf{v}_{w,i} - \mathbf{v}_{w,j}||^2$$

where i and j represent two different vector spaces (and, different word2vec models). The alignment is done based on word2vec vectors $\mathbf{v}$ for a given word w from a given year i or j. The vocabulary used in the Procrustes alignment, or V, is the intersection of the vocabulary of the words for the years of interest.

We can then calculate the diachronic embedding similarity of a word's meaning

across two years i and j by taking the word vector from year i and applying the Procrustes transformation R to transform it into the target vector space. We then compute the cosine similarity of the resulting vector with the corresponding word vector from year j. We also subtract an error-correcting scalar c, which is described further below.

$$distance(w, R) = 1 - cosine\_sim(R\mathbf{v}_{w,i}, \mathbf{v}_{w,j}) - c$$

Although the Procrustes linear transformation aligns the two vector spaces from each year such that word vectors across years should be more comparable, we know, however, that not all words change uniformly.[268] As such, the Procrustes transformation could be influenced by outliers,[270] or words in the vocabulary that have large differences in their vector representation over time. As a result, the resulting learned Procrustes linear transformation may be biased toward a larger correction, and thus, may over-correct vectors. To account for this, we adjust the resulting cosine similarity by an error-correcting value c. To calculate this adjustment term, we leverage the *Law of Conformity*, which hypothesizes that the semantic rate of change for a given word is inversely related to its relative frequency.[268] The intuition is that if this law holds, the most frequently occurring words in the corpus should show very little change over time. In other words, given the Law of Conformity, we expect that $cosine\_sim(R\mathbf{v}_{w,i}, \mathbf{v}_{w,j}) \approx 1$ will hold as long as the words w are among the most frequently used words in the corpus. Given that, we calculate a correcting term c as:

$$c = \frac{1}{|L|} \sum_{w \in L} cosine\_sim(R\mathbf{v}_{w,i}, \mathbf{v}_{w,j})$$

where L is some subset of the vocabulary V such that L consists of the top n words in the corpus by frequency. For this study we chose n to be the top 1% most frequently occurring words in the intersection of the vocabularies of each corpus.

Figure 6.3 shows how an example Procrustes transform can align two different

word2vec vector spaces. In this example, we learn a transform that moves the year 2000 vectors (shown as the solid lines) as close as possible to the year 2010 vectors (the dashed lines), with the small arrows representing the transformation. A single linear transform is able to align most vectors, meaning that most words will be represented by similar vectors (and thus, have similar meanings) once the vector spaces are aligned. Some words, however, will not align, signifying a vector change beyond that which could be accounted for from the alignment of the vector spaces. The word "portal" in this example is one of those cases. As seen, the alignment transformation places the predicted 2010 vector far away from the actual 2010 vector. We can infer from this that the meaning of the word "portal" did in fact change over this time span.

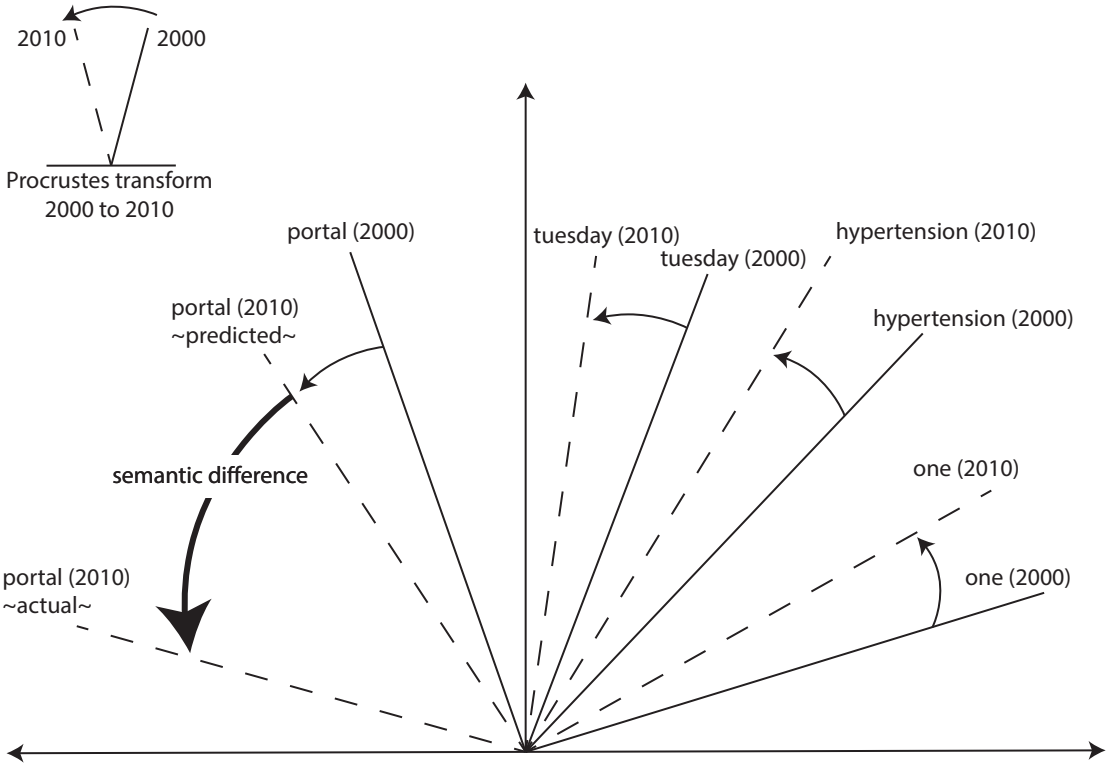


Figure 6.3: An example alignment of two word2vec vector spaces using a Procrustes linear transform. The "semantic difference" arrow here illustrates a temporal vector difference that is not accounted for via vector space alignment. We interpret this discrepancy as an indication of semantic drift.

6.3.3 Concept Drift

We next focus our analysis of semantic change on the effect of evolution in standardized terminologies. The Unified Medical Language System (UMLS) is a large biomedical terminology used to enable efficient and standardized use of medical language.[30] Distributed via the National Library of Medicine (NLM), the UMLS aggregates a variety of terminologies spanning multiple biomedical subdomains. For a resource with such broad scope within the domain, representing and reacting to change in clinical practice and biomedical knowledge is critical, and the UMLS has a robust change model built in for just this purpose.[251]

It is known that the UMLS changes over time, in terms of both the number and composition of concepts.[271] For applications that rely on the UMLS to provide biomedical language normalization, changes in the UMLS will inevitably propagate to downstream tooling. This section aims to determine the impact of UMLS change over time on the codification of diagnoses. Specifically, we look at how changing UMLS versions impacts the output of downstream concept extraction tooling. Similar methods have been proposed by Cardoso et al., where journal articles were annotated using older terminologies and differences were compared,[272] but here our focus is on clinical diagnosis text over a sixteen-year span of UMLS releases.

To examine how changes in the UMLS impact codification, we first set up a testing environment to examine codifications given multiple years of the UMLS. First, we downloaded and installed sixteen previous UMLS versions (the AA releases from 2004 to 2019). All terminologies matching the Level 0 + SNOMED CT configuration were installed for each year. Next, we installed QuickUMLS,[52] a tool that leverages a UMLS installation to extract UMLS concepts from text. As QuickUMLS is tied to a specific version of UMLS, this installation was repeated for each of the sixteen UMLS versions.

After the environment setup, the analysis of codification differences was conducted. First, we looped over a random sample of 100,000 clinical diagnoses taken from our

large clinical corpus. Then, we extracted the UMLS concepts for each diagnosis using all versions of UMLS, and compared them in pairwise fashion. Because two different UMLS years may result in different sets of extracted concepts, we computed the similarity of the concept extraction for the two years using Jaccard similarity,[233] which is the number of concepts in common over the union set of concepts for each of the two years:

$$Jaccard(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

A Jaccard similarity score of 1 indicates perfect alignment of the extracted concept sets, while a score of 0 signifies that the sets were disjoint. Findings will be reported in terms of Jaccard distance, which is equivalent to $1 - Jaccard(A, B)$.

6.3.4 Pragmatic Drift

Language pragmatics is the study of how linguistic meaning is understood through the lens of context and intent.[273] A focus on pragmatics in this section augments our previous analyses, which focused primarily on the meaning derived from the linguistic forms of the diagnoses (i.e. the semantics). An analysis of pragmatics is necessary for a robust accounting of language change, as semantic and pragmatic meaning are not always completely congruent – or, as Bender explains, "meaning derived from form is different from meaning derived from context of use."[274]

In general, pragmatics helps to define the link between an expression and its interpretation.[275] This is important for all NLP domains, but especially so for clinical diagnoses, as a diagnosis is not noted in isolation, but is entered under the context of the clinician's examination, assumptions, and interpretations. For example, if a patient has a diagnosis of "hypertension" on their problem list, variations in clinical practice, interpretation of symptoms, or plan of treatment may significantly impact the interpretation of that diagnosis. In other words, the text "hypertension" in the diagnosis section may intrinsically convey some information given its interpretation as a common

medical term, but the term alone does not capture the full clinical intent.

For this task, we use the IRP section as a window into the larger context and interpretation of the diagnosis text. Specifically, given a set of diagnoses, we are interested in how their associated IRP sections change over time. To examine this, we first selected the top ten diagnoses (by count) from the diagnosis section over all years. Then, for each of these diagnoses, we collected their associated IRP sections, segmented by year. We then examined each diagnoses' attendant IRP sections for diachronic linguistic change. We expect that even when the summary-level diagnosis text remains constant over time (such as "hypertension"), the intent, interpretation, and context of that diagnosis noted in the IRP section will change.

The null hypothesis would be that for each diagnosis there would be no change in the associated IRPs over time. We expect, however, that changes in clinical practice will cause the language in the IRPs to drift. We used Term Frequency Inverse Document Frequency (TF-IDF), a technique to compute text similarity, to measure the change in IRP sections over time based on weighted term frequencies.[276] TF-IDF transforms documents into vectors by computing a frequency distribution of terms – where a *document* for our purposes is defined as one IRP section. The advantage of TF-IDF over only term frequency is that it places higher importance on terms that occur in small numbers of documents, highlighting indicative terms in the corpus. TF-IDF can be derived as follows:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$
$$idf(t, D) = log \frac{|D|}{1 + |\{d \in D : t \in d\}|}$$

where $tfidf$ is a product of the term frequency (tf) and the inverse document frequency (idf) over a set of documents D for a given term t. For each document, this process can be repeated for every term in the corpus vocabulary V to yield an $|V|$-dimensional vector representing the composition of each document.

6.3.5 Case Study

We conclude our methods with a case study that examines linguistic change in the context of a current clinical project. We use an ongoing Mayo Clinic CDS project, MayoExpertAdvisor (MEA),[277] as a case study, as MEA uses unstructured data to deliver point-of-care recommendations to physicians and relies heavily on NLP results. We focus on two conditions that MEA currently monitors: heart failure and atrial fibrillation. This is in contrast to our previous methods described above, where clinical language change was explored broadly across all conditions and clinical problems.

For this case study we focus on two aspects of language change. First, we explore whether or not clinical diagnosis text for the two clinical conditions monitored by MEA does in fact change steadily over time, or *Incremental Diachronic Drift*. Next, we determine if we can detect an abrupt shift in the clinical practice, or *Sudden Shift*. In our case, the precipitating event for the sudden change was the implementation of a new Electronic Health Record. We suspect that such a shift will bring larger than expected linguistic change. Diagnosis text was extracted from a large corpus of clinical notes using results from MedTagger,[278] the Mayo Clinic NLP system that powers MEA.

Incremental Diachronic Drift

In order to test shifting diagnosis text over time, a corpus of approximately 100,000 clinical notes with mentions of heart failure and atrial fibrillation were collected over the years 2005 to 2017. To show diachronic drift, we hypothesize that chronologically closer years will have more similar language characteristics. In terms of language models, this means a model trained on a given year will "fit" next year's text less well, and even less for the year after that – signifying an incremental drift. The methods for this analysis generally follow the approach outlined in Section 6.3.1.

Our experiment was initiated by training a bigram language model for each year in our data set. For each year, we then computed the perplexity of the model against three

different test sets: (1) a held-out portion of text from the training year, (2) text from the next year, and finally (3) two years after the training year. We expect perplexity to increase as the test set year gets further from the training set year.

Sudden Shift

In 2018 Mayo Clinic changed EHR implementations, bringing changes to many aspects of clinical data capture including free-text clinical notes. To determine if our methods could retrospectively detect this change, we split our corpus into two segments: all text before 2018 (`Pre-2018`) and all text after the 2018 EHR change (`Post-2018`). Then, given a language model trained on the `Pre-2018` corpus, we measured the perplexity of the model against a held-out portion of the `Pre-2018` corpus as well as the `Post-2018` corpus to see if there has been a larger-than-expected shift.

6.4 Results

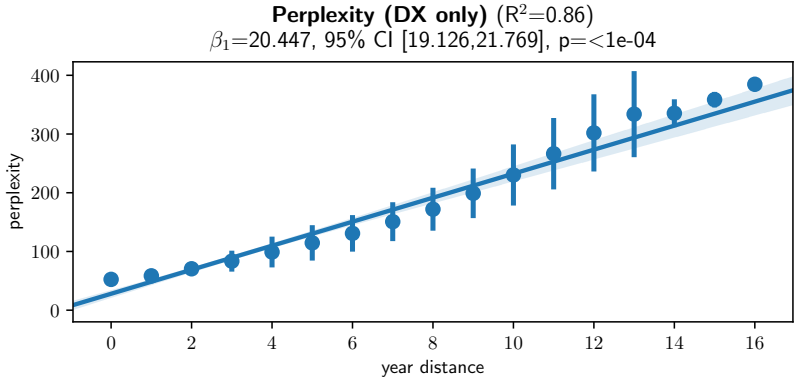
Results for each of the diachronic change analysis methods are shown below. Where applicable, change is represented as a function chronological distance between the two groups of text under analysis, noted as "year distance" on the x-axis of plots. Simple linear regression is used to show change characteristics over time where applicable, with β_1 denoting the coefficient of the non-intercept term. Error bars shown on plots represent the standard deviation for the observations.

6.4.1 Language Model Perplexity

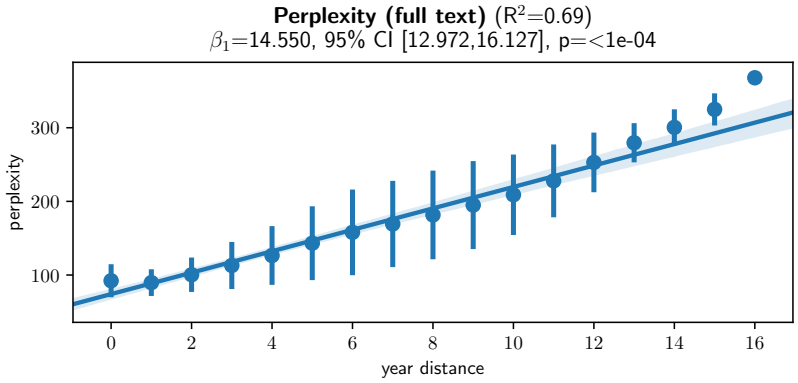
Figure 6.4a shows the trend of perplexity given year distance between train and test corpora for text in the diagnosis section. This figure contrasts the amount of perplexity (y-axis) plotted against the number of years distance between the language model and the testing corpus (x-axis).

Figure 6.4b shows a similar perplexity analysis for the IRP section. As shown, for

both diagnoses and IRPs, language perplexity increases as the test corpus gets farther (in terms of time) from the training corpus.



(a) Perplexity of the Diagnosis section.



(b) Perplexity of the IRP section.

Figure 6.4: Language model perplexity change over time, as measured as a function of time (in years) between text.

Figure 6.5 shows the same perplexity analysis as above, but repeated separately for the top five most frequently occurring note types. We conducted individual analysis of any note type that constituted more than 10% of the corpus total. The perplexity trends of the individual note types seem to agree with trends seen in the analysis of the full corpus.

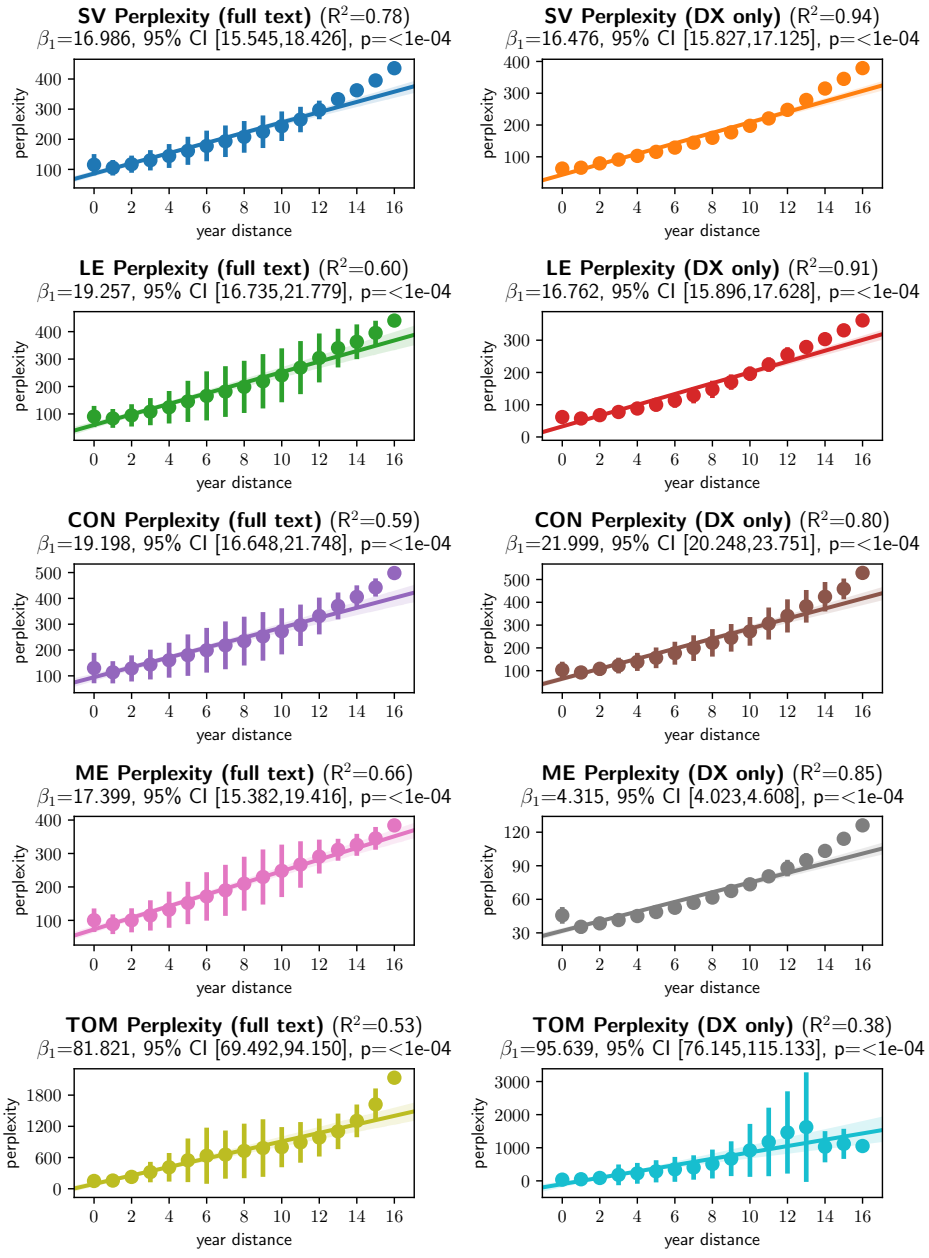


Figure 6.5: Language model perplexity over time for text stratified by note type.

6.4.2 Embedding Drift

Embedding vector changes over time for a selection of words are shown in Figure 6.6. As shown, different words show different rates of change within the word2vec vector space. For example, words like "dr." and "tuesday" change very little over time, while the words "guideline" and "portal" show considerable amounts of change.

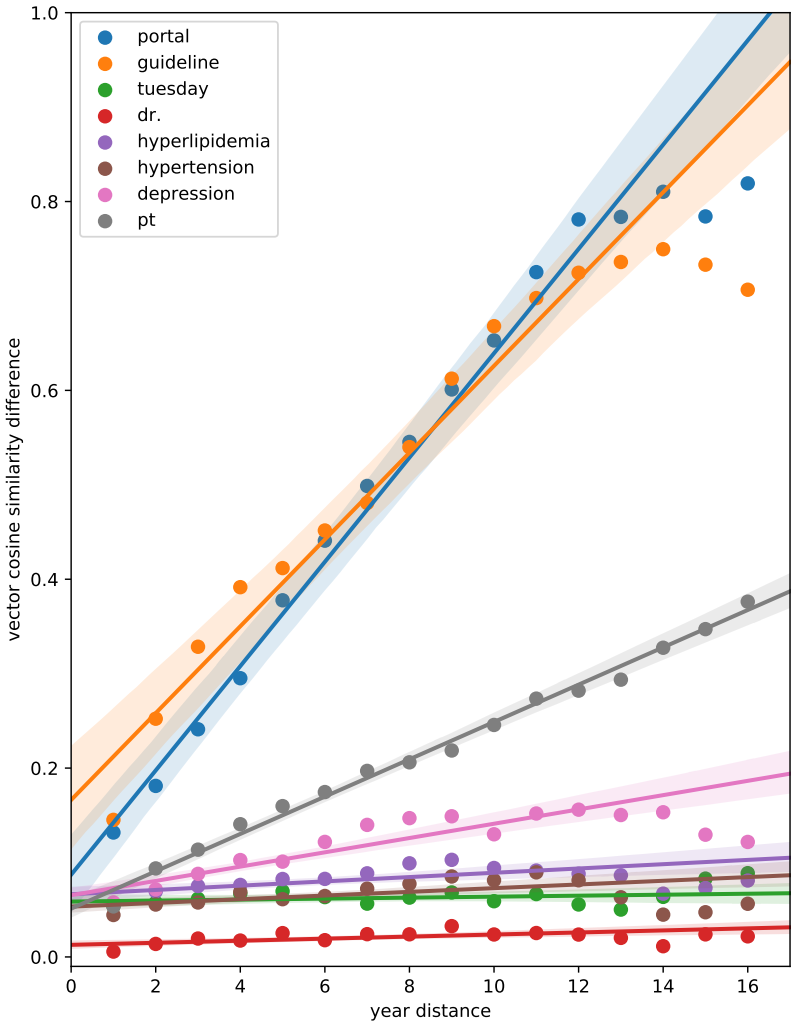


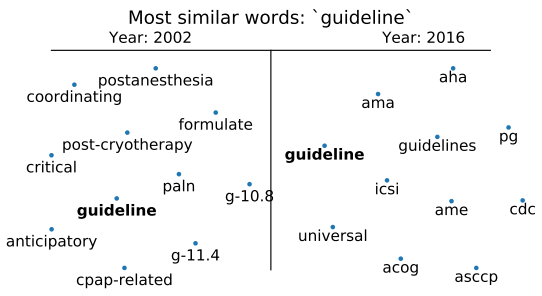
Figure 6.6: Word2vec vector cosine similarity difference for a select set of words over compared over time.

Table 6.1 shows regression information corresponding to Figure 6.6. Again, different words exhibit different rates and characteristics of drift. For example, the 95% confidence interval for the rate of change for the words "tuesday" and "dr." includes zero, indicating that no change over time is expected and the meaning is renaming stable. Other words, such as "portal," for example, show much higher rates of change.

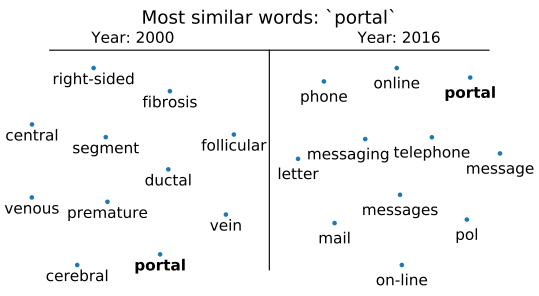
Table 6.1: Linear regression details for Figure 6.6. The coefficient β_1 represents the amount of expected decrease in cosine similarity per year.

word	R^2	β_1	β_1 CI (95%)	β_1 p-value
portal	0.589	0.055	[0.047,0.063]	<1e-04
guideline	0.460	0.046	[0.037,0.054]	<1e-04
pt	0.818	0.020	[0.018,0.021]	<1e-04
depression	0.373	0.008	[0.006,0.009]	<1e-04
hyperlipidemia	0.115	0.002	[0.001,0.003]	<1e-04
hypertension	0.096	0.002	[0.001,0.003]	0.0002
dr.	0.072	0.001	[0.000,0.002]	0.0016
tuesday	0.007	0.001	[-0.001,0.002]	0.3436

The word clouds for two words "guideline" and "portal" are shown in Figure 6.7. For the word "guideline" in Figure 6.7a, differences in the two word clouds can be interpreted as differences over time in patient care guideline usage. As shown, as care guidelines change, the word "guideline" similarly changes to refer to different things.



(a) Similar words for "guideline" from year 2002 (left) vs. 2016 (right).



(b) Similar words for "portal" from year 2000 (left) vs. 2016 (right).

Figure 6.7: Contrasting the neighborhood of similar words for two words exhibiting substantial change in meaning over time.

Figure 6.7b shows how the word "portal" has changed in meaning. In this example, it moved from having a mostly anatomical meaning to representing various emerging forms of tele-health, specifically the Mayo Clinic Patient Portal.[279]

6.4.3 Concept Drift

Figure 6.8 shows the degree of change for UMLS diagnosis codifications (in terms of Jaccard distance) using QuickUMLS over 16 years of UMLS distributions. As shown, the farther the UMLS distributions are apart in years, the greater the differences in codification are on average.

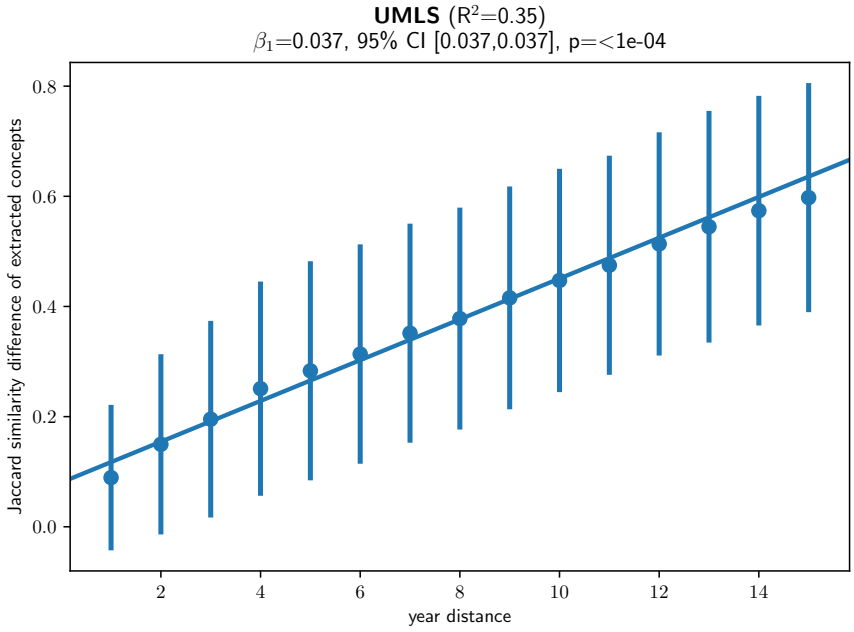


Figure 6.8: The Jaccard distance of UMLS concepts when codified using two different years of the UMLS. The x-axis represents the distance (in years) between the UMLS distributions.

The similarity distributions for each year distance period are shown in Figure 6.9. The closer the UMLS installations are in years, the more the distribution is skewed toward zero difference. As the year gap increases, more of the probability density

moves towards higher degrees of dissimilarity, moving the center of the Jaccard distance probability distribution farther to the right.

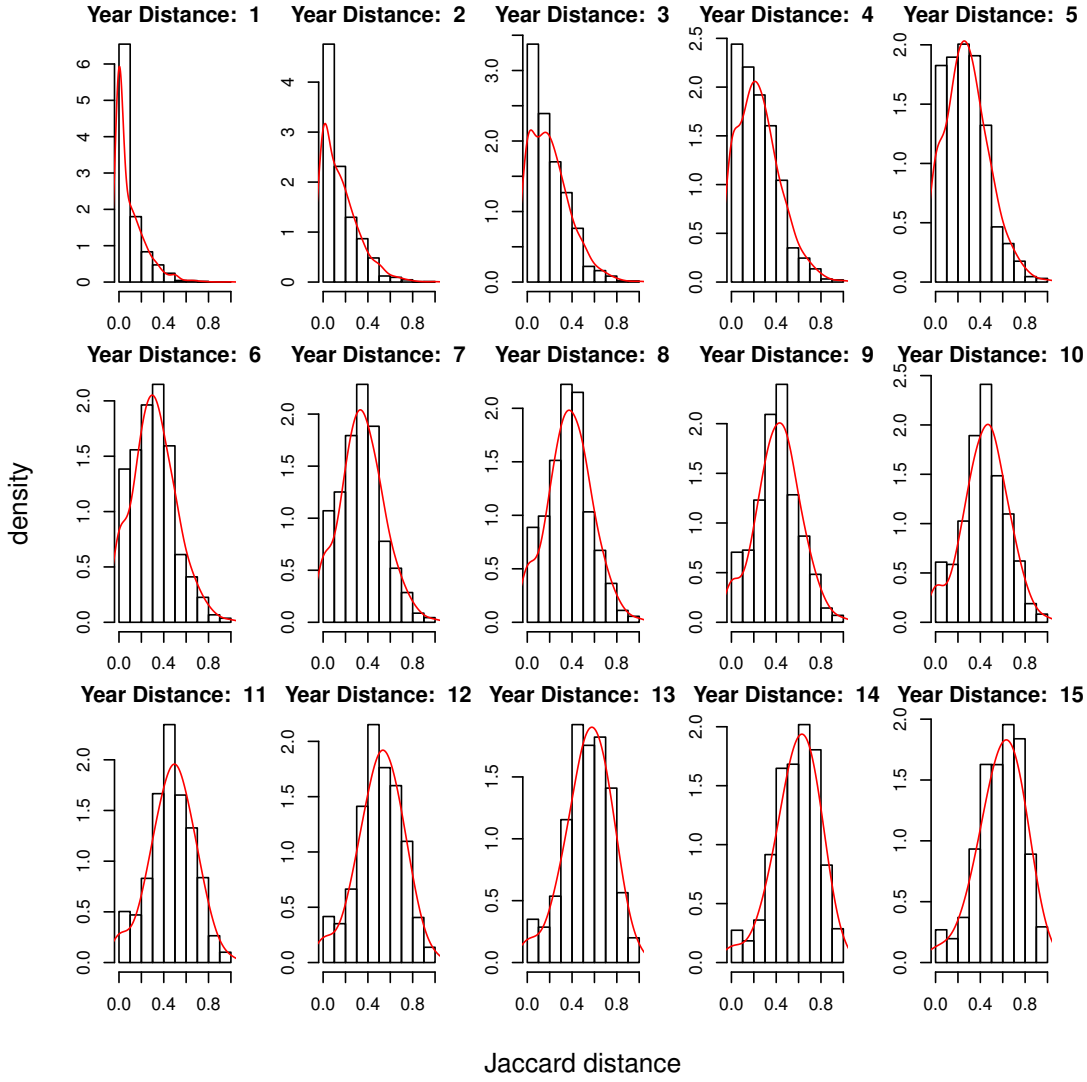


Figure 6.9: The distribution of Jaccard distance of UMLS concepts when codified using two different years of the UMLS. The shift from right to left skew in the probability density distribution corresponds to increasingly different UMLS codifications over time.

6.4.4 Pragmatic Drift

Pragmatic drift measures the changes in context and intent of a diagnosis through examination of the IRP section. Figure 6.10 shows the top ten diagnoses in our corpus and how their corresponding IRPs change over time. This shows all IRPs are changing, or growing more dissimilar over time, with some changing at different rates.

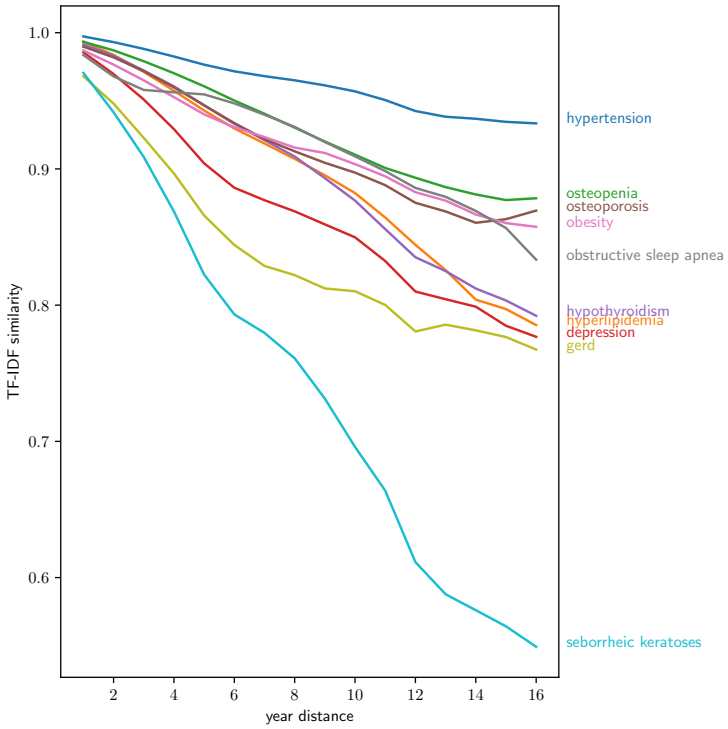


Figure 6.10: TF-IDF similarity as a function of year distance of IRP for the ten most common diagnoses. This figure illustrates the drift of the clinical narrative even if the diagnosis text stays constant.

Figure 6.11 shows the individual regression fits in more detail. As shown, time accounts for variable amounts of change, as shown by the differences in R^2. Also, differences in the β_1 coefficient indicate different rates of change. A higher β_1 value indicates more change, while lower values suggest that the context and intent of that diagnosis are staying relatively stable over time.

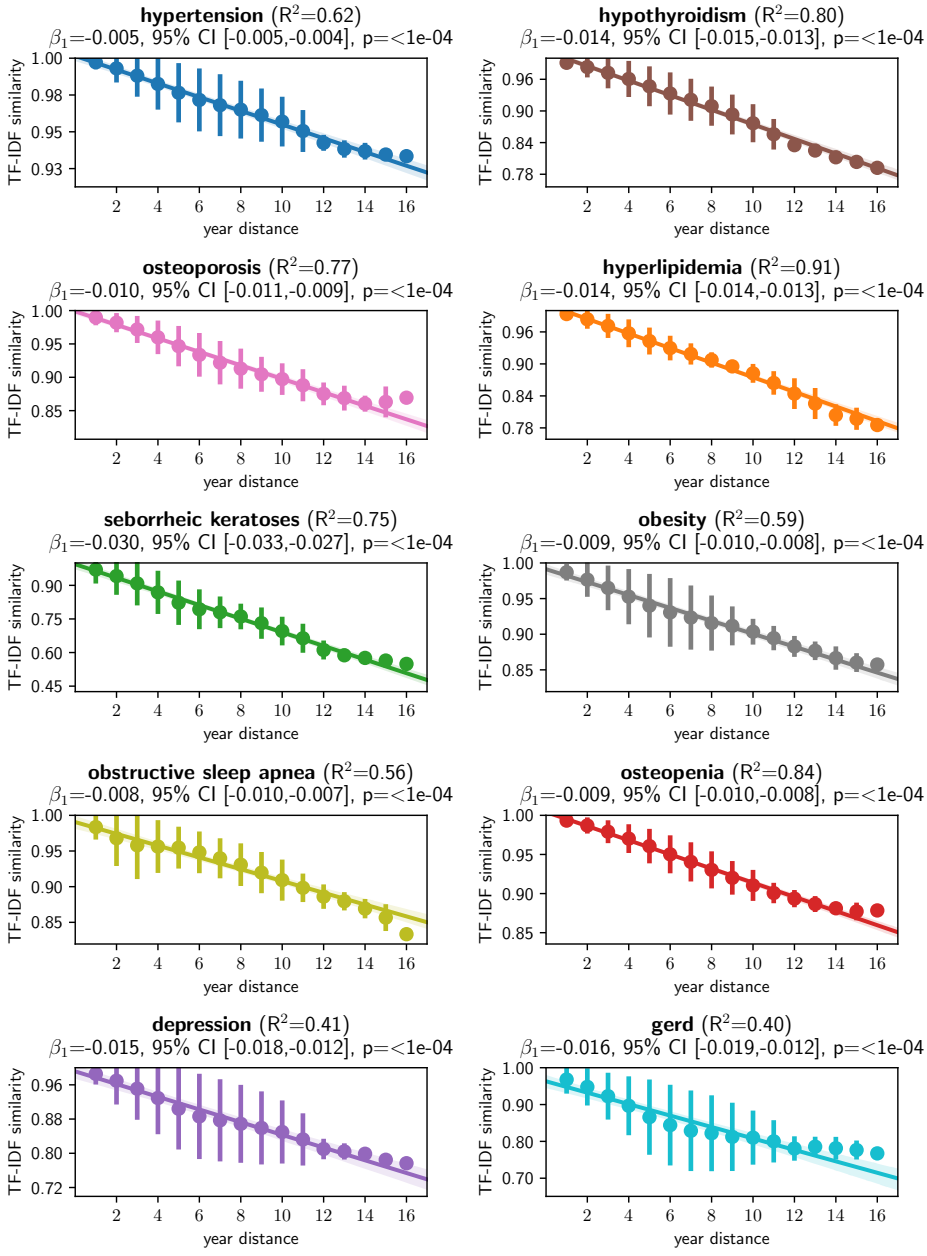


Figure 6.11: An in-depth representation of Figure 6.10 highlighting each diagnosis with its simple linear regression information regarding TF-IDF change over time.

6.4.5 Case Study

Figures 6.12 & 6.13 show linguistic change over time based on model perplexity for both clinical conditions examined in our case study. For these figures, the plotted lines indicate perplexity scores for each year's language model when tested against three test corpora: text from the same year in which it was trained, train year + 1, and train year + 2. Given text from each year, a trained language model was best able to predict a held-out portion of text from the year it was trained on ("train year"), as shown by the low perplexity score. Fitting the next year ("train year + 1") was slightly worse, as was two years in the future ("train year + 2").

For this analysis, the null hypothesis is that all three plot lines would overlap, meaning a language model trained on a given point in time will have no additional difficulty modeling future language. Our experiment suggests otherwise, indicating that years closer chronologically have more similar language characteristics.

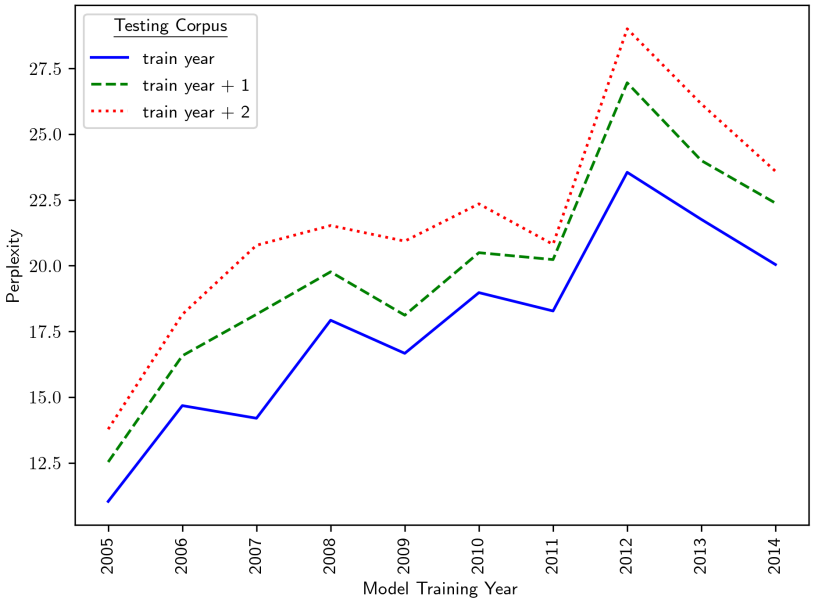


Figure 6.12: Language model perplexity for atrial fibrillation plotted against three test corpora: train year, train year + 1, and train year + 2.

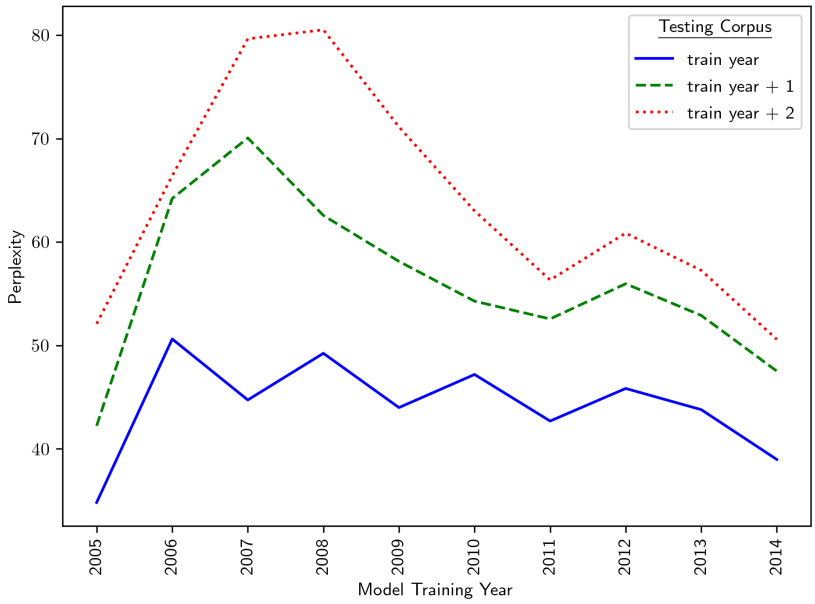


Figure 6.13: Language model perplexity for heart failure plotted against three test corpora: train year, train year + 1, and train year + 2.

Table 6.2 shows the `Pre-2018` language model tested against its own years (baseline) vs. text from after the EHR change (`Post-2018`). Both conditions show a drastic jump in perplexity using the `Pre-2018` language model to predict `Post-2018` language. This suggests that language patterns from before 2018 are poor predictors of language after 2018, signifying sudden language change.

Table 6.2: Detecting sudden language shift in diagnosis text across a major clinical change event (a 2018 EHR change).

Train Corpus / Test Corpus	Perplexity	
	Atrial fibrillation	Heart failure
`Pre-2018` (train) / `Pre-2018` (test) – (baseline)	7.44	13.66
`Pre-2018` (train) / `Post-2018` (test)	149.16	55.86

6.5 Discussion

For both the diagnosis and the IRP sections, a language model trained on a given year exhibits decaying performance as time progresses. This can be seen in Figures 6.4a & 6.4b, where the year distance between language model and test data is correlated with higher perplexity. This supports the hypothesis that clinical language is changing over time – both in terms of the diagnosis text itself, and the supporting context and detail in the IRP section. Figure 6.5 is further evidence of this change, showing that diachronic change can be detected even when controlling for note types, generally matching what was observed when analyzing all note types together.

Figure 6.6 demonstrates that different words do change at different rates. We would not expect semantic change for common, unambiguous words such as "dr." and "tuesday," and this is confirmed in the flat rate of change. Other words such as "guideline" and "portal" are subject to considerable amounts of change, as shown by the increasing average vector distance over time. Table 6.1 shows this sample of words ordered by descending rates of vector change over time.

Figure 6.7 examines the word clouds for the two words from our embedding comparison subset with the highest change rates. For the word "guideline," some of this change can be explained by the movement of Mayo Clinic from internal department-based guidelines to an enterprise system called AskMayoExpert (AME), a centralized system to disseminate clinical knowledge in the form of standardized care guidelines.[280,281] This is shown as "ame" in the word cloud. Similarly, Figure 6.7b shows the shift of the word "portal" to an entirely different sense focused on messaging and patient interaction via technology, following general trends of technology change including the introduction of the Mayo Clinic Patient Portal and the expanding role of online patient services at the Mayo Clinic over that time.

Figure 6.8 shows the impact that UMLS changes over time can have on diagnosis codification. Although codification change over time is unsurprising given that the

UMLS itself changes over time, it is an important factor that must be taken into account by downstream applications.[271] As shown in Figure 6.8, the Jaccard similarity of the resulting concept sets decreases by around 0.037 for every year difference between UMLS versions. An interpretation of this result is best explained through a scenario: If we upgrade UMLS by one year, how many concepts can we expect to change during QuickUMLS concept extraction? Or, concretely, how many concepts n could we extract such that we could expect on average one difference, where "one difference" is assumed to be a single concept change. We can express this in terms of the Jaccard distance $\frac{|A \bigcap B|}{|A \bigcup B|}$, where a single concept difference in n concepts could be represented by setting the intersection count to $n-1$ and the union count to $n+1$:

$$
\begin{aligned}
1 - \beta_1 t &= \frac{|A \bigcap B|}{|A \bigcup B|} \\
1 - 0.037(t) &= \frac{|A \bigcap B|}{|A \bigcup B|} \\
1 - 0.037(1) &= \frac{n-1}{n+1} \\
&\dots \\
n &\approx 53
\end{aligned}
$$

Given this, we see that by upgrading UMLS by one year we could reasonably expect that out of every 53 concepts codified, one would be codified differently. This effect becomes more pronounced over time, however. For example, if the change in UMLS was ten years:

$$
\begin{aligned}
1 - 0.037(10) &= \frac{n-1}{n+1} \\
&\dots \\
n &\approx 4
\end{aligned}
$$

we see a much smaller number of concepts before we would observe one difference in codification. This means that if end users were to refresh a ten-year-old UMLS installation, they could expect one out of every four concepts from clinical diagnosis statements to be codified differently. Such a drastic change may overwhelm downstream analytics, so it is important that users consider the implications of delaying terminology upgrades.

We also present evidence that the underlying intent, interpretation, and context of diagnosis statements are not necessarily fixed even if the summary-level diagnosis text remains constant. Figures 6.10 & 6.11 show that even when the underlying summary diagnosis is stable, the associated IRP section does in fact change over time. This implies that the pragmatics of diagnosis statements are subject to drift as well as their semantics. It is important to note that change due to time is only one source of change, as the IRP sections for some diagnoses seem inherently more variable even after taking change over time into account. As shown, the relatively low R^2 scores for some of the diagnoses in Figure 6.11 indicates that for some diseases, variation in the IRP section is influenced heavily by aspects other than change over time (such as clinical diagnostic variation, heterogeneity of symptoms/treatment, or other factors).

Finally, our case study reinforces our larger findings, as we note that our two diseases of interest, heart failure and atrial fibrillation, show similar change characteristics compared to the larger notes corpus. Table 6.2 highlights a facet of change not explored elsewhere in this study: the possibility of sudden language change, as opposed to gradual drift over time. This type of change could be particularly problematic, as there would be less time for running NLP systems to both detect the change and to react appropriately.

6.6 Conclusion

In this study we conducted a multi-faceted examination of diachronic linguistic change in clinical diagnoses. We find that though the change characteristics of clinical text are varied and nuanced, evidence of change over time is found through all methods presented here. It is our intent that the methods described here can be used to help running NLP systems (including implementations of the standardization framework proposed in this dissertation) to detect and adapt to change over time, resulting in more robust and resilient processing of clinical text.

Acknowledgments

This study was funded in part by NCATS U01TR002062.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This work presents a standardization framework for clinical problem descriptions. The preceding chapters have outlined a standardization progression starting with basic entity/relationship models and culminating in alignment to international healthcare data standards. Aside from standardization, we also have devoted two chapters to pragmatic issues encountered during standardization activities: organizing constraining value sets and accounting for semantic drift over time. It is our intent that addressing these concerns will make standardization more approachable and practically achievable for implementers.

Chapter 2 began the standardization progression, and focused on capturing semantic patterns in terms of clinical entities and their relationships. The data-driven techniques proposed not only allowed unsupervised extraction of prominent clinical language patterns, but also enabled comparison to industry-standard code systems. This chapter established our Web Ontology Language (OWL) representation of clinical problems, and

showed that data-driven extracted semantic patterns do have some alignment with pre-coordinated patterns found in a common industry standard, the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) CORE Problem List Subset.

Chapter 3 introduced the next level of standardization for clinical problems, using SNOMED CT expressions to codify problems with relevant modifiers and context. This chapter outlined the details of a deep-learning framework to learn entity relationships, an integral part of building SNOMED CT expressions and other downstream standardized representations. We found that this framework was effective, outperforming non-deep learning approaches and the previously reported state-of-the-art.

Our standardization methods ultimately concluded in Chapter 4, where clinical problems were transformed into Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR). This built on the work of Chapter 3, but introduced several new methods to deal with model generalizability. A multi-faceted, semi-supervised approach was introduced to help transform free text into FHIR `Condition` resources, and showed considerable improvement over Chapter 3 when generalizing to real-world clinical text.

The final two chapters addressed potential implementation challenges when applying and implementing this framework. First, Chapter 5 introduced two novel organizational methods for value sets, important artifacts that are used in the FHIR standardization process and elsewhere. Our results showed that these techniques can be successfully used to overlay organizational structure on large collections of value sets. We also proposed some user-focused interface design strategies that can be coupled with our methods to provide a better user experience through tooling. Next, Chapter 6 explored a common source of semantic misalignment – changing meaning over time. In this chapter, four different methods were employed to detect change in clinical problem language over time using a large clinical repository spanning seventeen years. We showed that all of these methods were able to detect meaningful linguistic and semantic change over time. Our techniques were also able to detect large, sudden changes in language caused by drastic shifts in clinical data capture techniques due to the adoption of new technologies.

7.2 Future Work

One important direction for future work is to explore concrete applications of these methods through pairings with real-world clinical use cases. Possible avenues for deployment would include the following scenarios:

- **Assisting abstractors.** Manual abstraction of free-text clinical notes is a labor-intensive undertaking requiring specifically-trained subject matter experts. Increasing the efficiency of this task is an active area of research, with Natural Language Processing (NLP) techniques playing a large role.[282] A potential use case for this framework is to be used in coordination with the abstraction process, where subject matter experts could highlight free-text problems in the clinical narrative, which could then be standardized.

- **Allowing clinical knowledge engineers to mine prominent conditions and modifiers.** Engineering clinical knowledge artifacts such as care guidelines and other decision aids requires robust knowledge of the underlying domain, knowledge that can be mined via data-driven ontologies.[283] Our framework could be leveraged to assist in data-driven knowledge extraction, especially given the focus on OWL representations introduced in Chapter 2.

- **Integration with point-of-care systems to assist in coding clinical problems.** Despite rapid advances in technology and the widespread adoption of Electronic Health Records (EHRs), integrating these tools into the clinical workflow continues to be a challenge. Specifically, the codification of clinical problems is regularly the responsibility of clinicians, not informaticians, which in the context of a clinical encounter is inefficient and error-prone.[10] Integration of this framework into point-of-care data capture systems could help bridge the gap between clinician-friendly free text and standardized forms required by underlying EHRs.

- **Evaluation of gaps in existing ontologies.** Chapter 2 introduces some comparisons between data-driven concepts and standardized representations in the SNOMED CT CORE Problem List Subset. This approach could be expanded to other coding systems and positions this technique as a framework for comparing local semantics with any number of external terminologies or ontologies.

References

[1] Ivo D Dinov. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. *Gigascience*, 5(1):s13742–016, 2016.

[2] John S Luo. Electronic medical records. *Primary Psychiatry*, 13(2):20–23, 2006.

[3] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):3, 2014.

[4] Lawrence L Weed. Medical records that guide and teach. *New England Journal of Medicine*, 278(11):593–600, 1968.

[5] Adam Wright, Dean F Sittig, Julie McGowan, Joan S Ash, and Lawrence L Weed. Bringing science to medicine: an interview with Larry Weed, inventor of the problem-oriented medical record. *Journal of the American Medical Informatics Association*, 21(6):964–968, 2014.

[6] Alvan R Feinstein. The problems of the problem-oriented medical record. *Annals of Internal Medicine*, 78(5):751–762, 1973.

[7] S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186, 2011.

[8] Chad M Hodge and Scott P Narus. Electronic problem lists: a thematic analysis of a systematic literature review to identify aspects critical to success. *Journal of the American Medical Informatics Association*, 25(5):603–613, 2018.

[9] SNOMED International. SNOMED CT Sept 2018. https://www.nlm.nih.gov/healthit/snomedct/us_edition.html, 2018. [Accessed: 01-01-2019].

[10] Eva S Klappe, Nicolette F de Keizer, and Ronald Cornet. Factors influencing problem list use in electronic health records—application of the unified theory of acceptance and use of technology. *Applied Clinical Informatics*, 11(03):415–426, 2020.

[11] Duane Bender and Kamran Sartipi. HL7 FHIR: an Agile and RESTful approach to healthcare information exchange. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 326–331. IEEE, 2013.

[12] Barry Smith and Werner Ceusters. HL7 RIM: an incoherent standard. *Studies in Health Technology and Informatics*, 124:133–138, 2006.

[13] Carl F Cargill. Why standardization efforts fail. *Journal of Electronic Publishing*, 14(1), 2011.

[14] Tim Benson and Grahame Grieve. Why interoperability is hard. In *Principles of Health Interoperability*, pages 19–35. Springer, 2016.

[15] M-M Bouamrane, C Tao, and IN Sarkar. Managing interoperability and complexity in health systems. *Methods of Information in Medicine*, 54(01):01–04, 2015.

[16] Jan Walker, Eric Pan, Douglas Johnston, Julia Adler-Milstein, David W Bates, and Blackford Middleton. The value of health care information exchange and

interoperability: there is a business case to be made for spending money on a fully standardized nationwide system. *Health Affairs*, 24(Suppl1):W5–10, 2005.

[17] Zilma Silveira Nogueira Reis, Thais Abreu Maia, Milena Soriano Marcolino, Francisco Becerra-Posada, David Novillo-Ortiz, and Antonio Luiz Pinho Ribeiro. Is there evidence of cost benefits of electronic medical records, standards, or interoperability in hospital information systems? overview of systematic reviews. *JMIR Medical Informatics*, 5(3), 2017.

[18] Rachel L Richesson and Jeffrey Krischer. Data standards in clinical research: gaps, overlaps, challenges and future directions. *Journal of the American Medical Informatics Association*, 14(6):687–696, 2007.

[19] Noushin Ashrafi, Jean-Pierre Kuilboer, and Tristan Stull. Semantic interoperability in healthcare: challenges and roadblocks. In *STPIS@ CAiSE*, pages 119–122, 2018.

[20] James J Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(04/05):394–403, 1998.

[21] Peter L Elkin, Steven H Brown, Casey S Husser, Brent A Bauer, Dietlind Wahner-Roedler, S Trent Rosenbloom, and Ted Speroff. Evaluation of the content coverage of SNOMED CT: ability of SNOMED Clinical Terms to represent clinical problem lists. *Mayo Clinic Proceedings*, 81(6):741–748, 2020/08/03 2006.

[22] Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, 2(11):e309, 2004.

[23] Hian Chye Koh and Gerald Tan. Data mining applications in healthcare. *Journal of Healthcare Information Management*, 19(2):65, 2011.

[24] Alexander Turchin, Maria Shubina, Eugene Breydo, Merri L Pendergrass, and Jonathan S Einbinder. Comparison of information content of structured and narrative text data sources on the example of medication intensification. *Journal of the American Medical Informatics Association*, 16(3):362–370, 2009.

[25] Marcelline R Harris, Laura Heermann Langford, Holly Miller, Mary Hook, Patricia C Dykes, and Susan A Matney. Harmonizing and extending standards from a domain-specific and bottom-up approach: an example from development through use in clinical applications. *Journal of the American Medical Informatics Association*, 22(3):545–552, 2015.

[26] Catalina Martínez-Costa, Ronald Cornet, Daniel Karlsson, Stefan Schulz, and Dipak Kalra. Semantic enrichment of clinical models towards semantic interoperability. the heart failure summary use case. *Journal of the American Medical Informatics Association*, 22(3):565–576, 2015.

[27] Alexander Koplenig. Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. *Digital Scholarship in the Humanities*, 32(1):159–168, 2017.

[28] Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. Diachronic evaluation of NER systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 97–107. Bochumer Linguistische Arbeitsberichte, 2016.

[29] Hongfang Liu, Kavishwar Wagholikar, and Stephen Tze-Inn Wu. Using SNOMED-CT to encode summary level data–a corpus analysis. *AMIA Summits on Translational Science Proceedings*, 2012:30–37, 2012.

[30] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004.

[31] Deborah L McGuinness and Frank Van Harmelen. OWL Web Ontology Language overview. *W3C Recommendation*, 10(10):2004, 2004.

[32] Graham Klyne and Jeremy J Carroll. Resource Description Framework (RDF): concepts and abstract syntax. 2006.

[33] Steve Harris, Andy Seaborne, and Eric Prud'hommeaux. SPARQL 1.1 query language. *W3C Recommendation*, 21(10), 2013.

[34] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

[35] Gabor Angeli, Victor Zhong, Danqi Chen, Arun Tejasvi Chaganty, Jason Bolton, Melvin Jose Johnson Premkumar, Panupong Pasupat, Sonal Gupta, and Christopher D Manning. Bootstrapped self training for knowledge base population. In *Proceedings of the Eighth Text Analysis Conference (TAC2015)*, pages 1–7, 2015.

[36] National Library of Medicine. The CORE problem list subset of SNOMED CT. https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html, 2010. [Accessed: 05-05-2020].

[37] Kin Wah Fung, Clement McDonald, and Suresh Srinivasan. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. *Journal of the American Medical Informatics Association*, 17(6):675–680, 2010.

[38] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[39] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[40] Irina Rish. An empirical study of the Naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, volume 3, pages 41–46, 2001.

[41] Rohit J Kate. Towards converting clinical phrases into SNOMED CT expressions. *Biomedical Informatics Insights*, 6:BII–S11645, 2013.

[42] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, pages 3567–3575, 2016.

[43] OpenCEM. OpenCEM Browser. http://www.opencem.org/, [Accessed: 11-10-2019].

[44] Dipak Kalra, Thomas Beale, and Sam Heard. The openEHR Foundation. *Studies in Health Technology and Informatics*, 115:153–173, 2005.

[45] Olivier Bodenreider, Duc Nguyen, Pishing Chiang, Philip Chuang, Maureen Madden, Rainer Winnenburg, Rob McClure, Steve Emrick, and Ivor D'Souza. The NLM Value Set Authority Center. *Studies in Health Technology and Informatics*, 192:1224, 2013.

[46] Pál Erdős and Alfréd Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[47] Theodora Bynon. *Historical linguistics*. Cambridge University Press, 1977.

[48] Gaurav Vashisth, Jan-Niklas Voigt-Antons, Michael Mikhailov, and Roland Roller. Exploring diachronic changes of biomedical knowledge using distributed concept representations. In *Proc. 18th BioNLP Workshop and Shared Task*, pages 348–358, 2019.

[49] Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. Evaluation metrics for language models. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop (1998)*, 1998.

[50] Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.

[51] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[52] Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*, pages 1–4, 2016.

[53] Carol Friedman, Pauline Kra, and Andrey Rzhetsky. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235, 2002.

[54] Bonaventura Coppola, Aldo Gangemi, Alfio Gliozzo, Davide Picca, and Valentina Presutti. Frame detection over the semantic web. In *European Semantic Web Conference*, pages 126–142. Springer, 2009.

[55] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. Semantic web machine reading with FRED. *Semantic Web*, 8(6):873–893, 2017.

[56] Na Hong, Andrew Wen, Daniel J Stone, Shintaro Tsuji, Paul R Kingsbury, Luke V Rasmussen, Jennifer A Pacheco, Prakash Adekkanattu, Fei Wang, Yuan Luo, Jyotishman Pathak, Hongfang Liu, and Guoqian Jiang. Developing a FHIR-based EHR phenotyping framework: a case study for identification of patients

with obesity and multiple comorbidities from discharge summaries. *Journal of Biomedical Informatics*, 99:103310, 2019.

[57] Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851, 2009.

[58] Kevin J Peterson and Hongfang Liu. The sublanguage of clinical problem lists: a corpus analysis. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1451. American Medical Informatics Association, 2018.

[59] RH Baud, Anne-Marie Rassinoux, Judith Corinna Wagner, Christian Lovis, C Juge, LL Alpay, Pierre-André Michel, Patrice Degoulet, and Jean-Raoul Scherrer. Representing clinical narratives using conceptual graphs. *Methods of Information in Medicine*, 34(01/02):176–186, 1995.

[60] Stephen B Johnson. A semantic lexicon for medical language processing. *Journal of the American Medical Informatics Association*, 6(3):205–218, 1999.

[61] Tielman T Van Vleck, Adam Wilcox, Peter D Stetson, Stephen B Johnson, and Noémie Elhadad. Content and structure of clinical problem lists: a corpus analysis. In *AMIA Annual Symposium Proceedings*, volume 2008, page 753. American Medical Informatics Association, 2008.

[62] Adam Wright, Francine L Maloney, and Joshua C Feblowitz. Clinician attitudes toward and use of electronic problem lists: a thematic analysis. *BMC Medical Informatics and Decision Making*, 11(1):36, 2011.

[63] Hongfang Liu, Stephen T Wu, Dingcheng Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Wagholikar, Peter J Haug, Stanley M Huff, and Christopher G Chute. Towards a semantic lexicon for clinical natural language processing. In

AMIA Annual Symposium Proceedings, volume 2012, page 568. American Medical Informatics Association, 2012.

[64] Sunghwan Sohn, Yanshan Wang, Chung-Il Wi, Elizabeth A Krusemark, Euijung Ryu, Mir H Ali, Young J Juhn, and Hongfang Liu. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *Journal of the American Medical Informatics Association*, 25(3):353–359, 2018.

[65] Peter L Elkin, Steven H Brown, Michael J Lincoln, Michael Hogarth, and Alan Rector. A formal representation for messages containing compositional expressions. *International Journal of Medical Informatics*, 71(2-3):89–102, 2003.

[66] Zellig Harris. *Theory of Language and Information: A Mathematical Approach*. Oxford University Press UK, 1991.

[67] Zellig S Harris. The structure of science information. *Journal of Biomedical Informatics*, 35(4):215–221, 2002.

[68] Amit P Sheth. Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In *Interoperating Geographic Information Systems*, pages 5–29. Springer, 1999.

[69] Naomi Sager, Margaret Lyman, Christine Bucknall, Ngo Nhan, and Leo J Tick. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160, 1994.

[70] Carol Friedman, George Hripcsak, William DuMouchel, Stephen B Johnson, and Paul D Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1):83–108, 1995.

[71] James J Cimino, Tiffani J Bright, and Jianhau Li. Medication reconciliation using natural language processing and controlled terminologies. In *Medinfo 2007:*

Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems, page 679. IOS Press, 2007.

[72] Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606, 2011.

[73] Collin F Baker, Charles J Fillmore, and John B Lowe. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics-Vol. 1*, pages 86–90. Association for Computational Linguistics, 1998.

[74] Charles J Fillmore. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32, 1976.

[75] Charles J Fillmore and Collin Baker. A frames approach to semantic analysis. In *The Oxford Handbook of Linguistic Analysis*, pages 313–340. Oxford University Press, 2010.

[76] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating NLP using linked data. In *International Semantic Web Conference (ISWC)*, pages 98–113. Springer, 2013.

[77] Louise Deléger, Leonardo Campillos, Anne-Laure Ligozat, and Aurélie Névéol. Design of an extensive information representation scheme for clinical narratives. *Journal of Biomedical Semantics*, 8(1):37, 2017.

[78] Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He, and Dongyan Zhao. Natural language question answering over RDF: a graph data driven approach. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 313–324, New York, NY, USA, 2014. ACM.

[79] Weiguo Zheng, Lei Zou, Xiang Lian, Jeffrey Xu Yu, Shaoxu Song, and Dongyan Zhao. How to build templates for RDF question/answering: an uncertain graph similarity join approach. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1809–1824. ACM, 2015.

[80] Marta Tatu, Mithun Balakrishna, Steven Werner, Tatiana Erekhinskaya, and Dan Moldovan. Automatic extraction of actionable knowledge. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 396–399. IEEE, 2016.

[81] Asma Ben Abacha and Pierre Zweigenbaum. MEANS: a medical question-answering system combining NLP techniques and semantic web technologies. *Information Processing & Management*, 51(5):570–594, 2015.

[82] Feichen Shen, Hongfang Liu, Sunghwan Sohn, David W Larson, and Yugyung Lee. BmQGen: biomedical query generator for knowledge discovery. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1092–1097. IEEE, 2015.

[83] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[84] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

[85] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *AMIA Annual Symposium Proceedings*, pages 17–21. American Medical Informatics Association, 2001.

[86] Marie-Catherine De Marneffe and Christopher D Manning. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics, 2008.

[87] Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. SNOMED Clinical Terms: overview of the development process and project status. In *AMIA Annual Symposium Proceedings*, page 662. American Medical Informatics Association, 2001.

[88] Henry Wasserman and Jerome Wang. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. In *AMIA Annual Symposium Proceedings*, volume 2003, page 699. American Medical Informatics Association, 2003.

[89] Olivier Bodenreider. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of Medical Informatics*, page 67, 2008.

[90] Alan L Rector and Sebastian Brandt. Why do it the hard way? the case for an expressive description logic for SNOMED. *Journal of the American Medical Informatics Association*, 15(6):744–751, 2008.

[91] Alexa T McCray. The UMLS Semantic Network. In *Proceedings. Symposium on Computer Applications in Medical Care*, pages 503–507. American Medical Informatics Association, 1989.

[92] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(suppl_2):W541–W545, 2011.

[93] Orri Erling and Ivan Mikhailov. RDF support in the Virtuoso DBMS. In *Networked Knowledge-Networked Media*, pages 7–24. Springer, 2009.

[94] Ora Lassila and Deborah McGuinness. The role of frame-based representation on the semantic web. *Linköping Electronic Articles in Computer and Information Science*, 6(5):2001, 2001.

[95] Alan Rector. Representing specified values in OWL: "value partitions" and "value sets". *W3C WG note*, 17, 2005.

[96] Nathaniel R Greenbaum, Yacine Jernite, Yoni Halpern, Shelley Calder, Larry A Nathanson, David Sontag, and Steven Horng. Contextual autocomplete: A novel user interface using machine learning to improve ontology usage and structured data capture for presenting problems in the emergency department. *bioRxiv*, page 127092, 2017.

[97] Ankur Agrawal, Zhe He, Yehoshua Perl, Duo Wei, Michael Halper, Gai Elhanan, and Yan Chen. The readiness of SNOMED problem list concepts for meaningful use of electronic health records. *Artificial Intelligence in Medicine*, 58(2):73–80, 2013.

[98] International Health Terminology Standards Development Organization (IHTSDO). SNOMED CT Expression constraint language specification and guide, v1.0, 2015.

[99] Kevin J Peterson and Hongfang Liu. Automating the transformation of free-text clinical problems into SNOMED CT expressions. In *AMIA Summits on Translational Science Proceedings*, pages 497–506. American Medical Informatics Association, 2020.

[100] Cathy M Price, Amanda C de C Williams, Blair H Smith, and Alex Bottle. Implementation of patient-reported outcomes (PROMs) from specialist pain clinics

in England and Wales: experience from a nationwide study. *European Journal of Pain*, 23(7):1368–1377, 2019.

[101] Ian S Jaffe, Karen Chiswell, and Ephraim L Tsalik. A decade on: systematic review of clinicaltrials.gov infectious disease trials, 2007-2017. In *Open Forum Infectious Diseases*, pages 1–9, 2019.

[102] Julian Zelingher, David M Rind, Enrique Caraballo, Mark S Tuttle, NE Olson, and Charles Safran. Categorization of free-text problem lists: an effective method of capturing clinical data. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 416. American Medical Informatics Association, 1995.

[103] Yichuan Wang, LeeAnn Kung, and Terry Anthony Byrd. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126:3–13, 2018.

[104] Casey Holmes. The problem list beyond meaningful use: part I: the problems with problem lists. *Journal of AHIMA*, 82(2):30–33, 2011.

[105] International Health Terminology Standards Development Organization (IHTSDO). Compositional grammar specification and guide, v2.3.1, 2016.

[106] James R Campbell and Thomas H Payne. A comparison of four schemes for codification of problem lists. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 201. American Medical Informatics Association, 1994.

[107] James R Campbell, Junchuan Xu, and Kin Wah Fung. Can SNOMED CT fulfill the vision of a compositional terminology? analyzing the use case for problem list. In *AMIA Annual Symposium Proceedings*, volume 2011, page 181. American Medical Informatics Association, 2011.

[108] Peter L Elkin, M Tuttle, Kevin Keck, K Campbell, G Atkin, and Christopher G Chute. The role of compositionality in standardized problem list generation. *Studies in Health Technology and Informatics*, 52:660–664, 1998.

[109] Stefan Schulz, Daniel Schober, Djamila Raufie, and Martin Boeker. Pre-and post-coordination in biomedical ontologies. In *OBML 2010 Workshop Proceedings*, pages L1–L4, 2010.

[110] Sergey Brin. Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, pages 172–183. Springer, 1998.

[111] Jose Antonio Miñarro-Giménez, Catalina Martínez-Costa, Pablo López-García, and Stefan Schulz. Building SNOMED CT post-coordinated expressions from annotation groups. *Studies in Health Technology and Informatics*, 235:446–450, 2017.

[112] Pablo López-García and Stefan Schulz. Structural patterns under x-rays: is SNOMED CT growing straight? *PLoS ONE*, 11(11):e0165619, 2016.

[113] Kin Wah Fung, William T Hole, Stuart J Nelson, Suresh Srinivasan, Tammy Powell, and Laura Roth. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. *Journal of the American Medical Informatics Association*, 12(4):486–494, 2005.

[114] Irena Spasić, Padraig Corcoran, Andrei Gagarin, and Andreas Buerki. Head to head: semantic similarity of multi–word terms. *IEEE Access*, 6:20545–20557, 2018.

[115] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: fast and robust models for biomedical natural language processing. In *Proceedings of the*

18th BioNLP Workshop and Shared Task, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[116] Tim Benson. *Principles of Health Interoperability HL7 and SNOMED*. Springer Science & Business Media, 2012.

[117] Marc Moreno Lopez and Jugal Kalita. Deep learning applied to NLP. *arXiv preprint arXiv:1703.03091*, 2017.

[118] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive Modeling*, 5(3):1, 1988.

[119] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, November 2016.

[120] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[121] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1):198, 2017.

[122] Yoon Kim. Convolutional neural networks for sentence classification. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, October 2014.

[123] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 2335–2344, August 2014.

[124] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[125] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[126] Ruth Reátegui and Sylvie Ratté. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making*, 18(3):74, 2018.

[127] William TF Goossen. Detailed clinical models: representing knowledge, data and semantics in healthcare information technology. *Healthcare Informatics Research*, 20(3):163–172, 2014.

[128] Adam S Rothschild, Harold P Lehmann, and George Hripcsak. Inter-rater agreement in physician-coded problem lists. In *AMIA Annual Symposium Proceedings*, volume 2005, page 644. American Medical Informatics Association, 2005.

[129] Kevin J Peterson, Guoqian Jiang, and Hongfang Liu. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. *Journal of Biomedical Informatics*, page 103541, 2020.

[130] Peter Salmon, Ann Rappaport, Mike Bainbridge, Glyn Hayes, and John Williams. Taking the problem oriented medical record forward. In *Proceedings of the AMIA Annual Fall Symposium*, page 463. American Medical Informatics Association, 1996.

[131] Beth Acker, June Bronnert, Teresa Brown, Jill S Clark, Betty Dunagan, Tracy Elmer, Suzanne Goodell, Kate Green, Pamela Heller, Casey Holmes, Margo Imel, Kathy Johnson, Crystal Kallem, Melanie Loucks, Sheetal Patel, Debbie A Reed,

Rita Scichilone, and Anne L Tegen. Problem list guidance in the EHR. *Journal of AHIMA*, 82(9):52, 2011.

[132] Sereh MJ Simons, Felix HJM Cillessen, and Jan A Hazelzet. Determinants of a successful problem list to support the implementation of the problem-oriented medical record according to recent literature. *BMC Medical Informatics and Decision Making*, 16(1):102, 2016.

[133] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24, 2019.

[134] Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Annals of Family Medicine*, 15(5):419–426, 2017.

[135] Philip J Kroth, Nancy Morioka-Douglas, Sharry Veres, Stewart Babbott, Sara Poplau, Fares Qeadan, Carolyn Parshall, Kathryne Corrigan, and Mark Linzer. Association of electronic health record design and use factors with clinician stress and burnout. *JAMA Network Open*, 2(8):e199609–e199609, 2019.

[136] Peter L Elkin, Kent R Bailey, Philip V Ogren, Brent A Bauer, and Christopher G Chute. A randomized double-blind controlled trial of automated term dissection. In *Proceedings of the AMIA Symposium*, pages 62–66. American Medical Informatics Association, 1999.

[137] JE Rogers and Alan L Rector. Terminological systems: bridging the generation gap. In *Proceedings of the AMIA Annual Fall Symposium*, pages 610–614. American Medical Informatics Association, 1997.

[138] Alan L Rector. Clinical terminology: why is it so hard? *Methods of Information in Medicine*, 38(04/05):239–252, 1999.

[139] Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. SemEval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, Denver, Colorado, June 2015. Association for Computational Linguistics.

[140] Joey Coyle, Yan Heras, Tom Oniki, and Stan Huff. Clinical Element Model. *University of Utah*, 2008.

[141] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[142] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics*, 77:34 – 49, 2018.

[143] Charles N Mead. Data interchange standards in healthcare IT-computable semantic interoperability: Now possible but still difficult. do we really need a better mousetrap? *Journal of Healthcare Information Management*, 20(1):71–78, 2006.

[144] CIMI. Clinical Information Modeling Initiative. https://cimi.hl7.org/, [Accessed: 12-12-2019].

[145] Na Hong, Andrew Wen, Feichen Shen, Sunghwan Sohn, Sijia Liu, Hongfang Liu, and Guoqian Jiang. Integrating structured and unstructured EHR data using an

FHIR-based type system: A case study with medication data. *AMIA Summits on Translational Science Proceedings*, 2018:74–83, 2018.

[146] George Despotou, Yannis Korkontzelos, Nicholas Matragkas, Eda Bilici, and Theodoras N Arvanitis. Structuring clinical decision support rules for drug safety using natural language processing. In *16th International Conference on Informatics, Management, and Technology in Healthcare (ICIMTH 2018)*, pages 89–92, 2018.

[147] Jessica Germaine Shull. Digital health and the state of interoperable electronic health records. *JMIR Medical Informatics*, 7(4):e12712, 2019.

[148] David Ferrucci and Adam Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

[149] Na Hong, Andrew Wen, Feichen Shen, Sunghwan Sohn, Chen Wang, Hongfang Liu, and Guoqian Jiang. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *JAMIA Open*, 2(4):570–579, 10 2019.

[150] Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 3*, COLING '92, pages 977–981, USA, 1992. Association for Computational Linguistics.

[151] Barbara Plank. What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:1608.07836*, 2016.

[152] David Vadas and James R Curran. Parsing noun phrases in the Penn Treebank. *Computational Linguistics*, 37(4):753–809, 2011.

[153] Yuichiro Sawai, Hiroyuki Shindo, and Yuji Matsumoto. Semantic structure analysis of noun phrases using abstract meaning representation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 851–856, Beijing, China, July 2015. Association for Computational Linguistics.

[154] Manabu Torii, Elly W Yang, and Son Doan. A preliminary study of clinical concept detection using syntactic relations. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1028. American Medical Informatics Association, 2018.

[155] Robin Kurtz, Daniel Roxbo, and Marco Kuhlmann. Improving semantic dependency parsing with syntactic features. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 12–21, Turku, Finland, September 2019. Linköping University Electronic Press.

[156] Sunghwan Sohn, Stephen Wu, and Christopher G Chute. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science Proceedings*, 2012:1, 2012.

[157] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657, Cambridge, MA, USA, 2015. MIT Press.

[158] Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 271–276, USA, 2011. Association for Computational Linguistics.

[159] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[160] Frank Reichartz, Hannes Korte, and Gerhard Paass. Semantic relation extraction with kernels over typed dependency trees. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 773–782, New York, NY, USA, 2010. Association for Computing Machinery.

[161] Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2006.

[162] Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-supervised Learning in NLP*, pages 10–18. Association for Computational Linguistics, 2012.

[163] Razvan C Bunescu and Raymond J Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731. Association for Computational Linguistics, 2005.

[164] Sun Kim, Haibin Liu, Lana Yeganova, and W John Wilbur. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of Biomedical Informatics*, 55:23–30, 2015.

[165] Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. A dependency-based neural network for relation classification. In *Proceedings of the*

53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 285–290, Beijing, China, July 2015. Association for Computational Linguistics.

[166] Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv preprint arXiv:1601.03651*, 2016.

[167] Yuan Luo, Özlem Uzuner, and Peter Szolovits. Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Briefings in Bioinformatics*, 18(1):160–178, 2016.

[168] Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, 2016:baw032, 03 2016.

[169] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068, 05 2016.

[170] Eryu Xia, Wen Sun, Jing Mei, Enliang Xu, Ke Wang, and Yong Qin. Mining disease-symptom relation from massive biomedical literature and its application in severe disease diagnosis. In *AMIA Annual Symposium Proceedings*, volume 2018, pages 1118–1126. American Medical Informatics Association, 2018.

[171] Changqin Quan, Meng Wang, and Fuji Ren. An unsupervised text mining method for relation extraction from biomedical literature. *PloS One*, 9(7):1–8, 7 2014.

[172] François Chollet et al. Keras. `https://keras.io`, [Accessed: 01-02-2020], 2015.

[173] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[174] Han Xiao. bert-as-service. `https://github.com/hanxiao/bert-as-service`, [Accessed: 01-02-2020], 2018.

[175] Andrew L Beam, Benjamin Kompa, Inbar Fried, Nathan P Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. *arXiv preprint arXiv:1804.01486*, 2018.

[176] Zhiheng Li, Zhihao Yang, Chen Shen, Jun Xu, Yaoyun Zhang, and Hua Xu. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. *BMC Medical Informatics and Decision Making*, 19(1):22, 2019.

[177] Alexa T McCray. An upper-level ontology for the biomedical domain. *Comparative and Functional Genomics*, 4(1):80–84, 2003.

[178] Alexa T McCray, Anita Burgun, and Olivier Bodenreider. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in Health Technology and Informatics*, 84(Pt 1):216–220, 2001.

[179] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, pages 269–282, 2017.

[180] Fei Xia and Meliha Yetisgen-Yildiz. Clinical corpus annotation: challenges and strategies. In *Proceedings of the Third Workshop on Building and Evaluating*

Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 2012.

[181] Hongfang Liu, Suzette J Bielinski, Sunghwan Sohn, Sean Murphy, Kavishwar B Wagholikar, Siddhartha R Jonnalagadda, KE Ravikumar, Stephen T Wu, Iftikhar J Kullo, and Christopher G Chute. An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings*, 2013:149, 2013.

[182] Yanshan Wang, Sunghwan Sohn, Sijia Liu, Feichen Shen, Liwei Wang, Elizabeth J Atkinson, Shreyasee Amin, and Hongfang Liu. A deep representation empowered distant supervision paradigm for clinical information extraction. *arXiv preprint arXiv:1804.07814*, 2018.

[183] Laura Chiticariu, Yunyao Li, and Frederick R Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 827–832, 2013.

[184] Andrew Wen, Sunyang Fu, Sungrim Moon, Mohamed El Wazir, Andrew Rosenbaum, Vinod C Kaggal, Sijia Liu, Sunghwan Sohn, Hongfang Liu, and Jungwei Fan. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ Digital Medicine*, 2(1):1–7, 2019.

[185] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[186] Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[187] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model

predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[188] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

[189] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

[190] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[191] William Kearns, Wilson Lau, and Jason Thomas. UW-BHI at MEDIQA 2019: An analysis of representation methods for medical natural language inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 500–509, Florence, Italy, August 2019. Association for Computational Linguistics.

[192] Kevin J Peterson, Guoqian Jiang, Scott M Brue, Feichen Shen, and Hongfang Liu. Mining hierarchies and similarity clusters from value set repositories. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1372. American Medical Informatics Association, 2017.

[193] David J Brailer. Interoperability: the key to the future health care system. *Health Affairs*, 24:W5, 2005.

[194] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary

use of EHR: data quality issues and informatics opportunities. *AMIA Summits on Translational Science Proceedings*, 2010:1–5, 2010.

[195] David Blumenthal and Marilyn Tavenner. The "meaningful use" regulation for electronic health records. *New England Journal of Medicine*, 363(6):501–504, 2010.

[196] James J Cimino, George Hripcsak, Stephen B Johnson, and Paul D Clayton. Designing an introspective, multipurpose, controlled medical vocabulary. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 513–518. American Medical Informatics Association, 1989.

[197] Jyotishman Pathak, Guoqian Jiang, Sridhar O Dwarkanath, James D Buntrock, and Christopher G Chute. LexValueSets: an approach for context-driven value sets extraction. In *AMIA Annual Symposium Proceedings*, volume 2008, page 556. American Medical Informatics Association, 2008.

[198] Jyotishman Pathak, Janey Wang, Sudha Kashyap, Melissa Basford, Rongling Li, Daniel R Masys, and Christopher G Chute. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *Journal of the American Medical Informatics Association*, 18(4):376–386, 2011.

[199] Kevin J Peterson, Guoqian Jiang, Scott M Brue, and Hongfang Liu. Leveraging terminology services for extract-transform-load processes: a user-centered approach. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1010. American Medical Informatics Association, 2016.

[200] Rachel L Richesson, James E Andrews, and Jeffrey P Krischer. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. *Journal of the American Medical Informatics Association*, 13(5):536–546, 2006.

[201] Christopher G Chute, Scott A Beck, Thomas B Fisk, and David N Mohr. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *Journal of the American Medical Informatics Association*, 17(2):131–135, 2010.

[202] Rainer Winnenburg and Olivier Bodenreider. Metrics for assessing the quality of value sets in clinical quality measures. In *AMIA Annual Symposium Proceedings*, page 1497. American Medical Informatics Association, 2013.

[203] David Garlan, Robert Allen, and John Ockerbloom. Architectural mismatch or why it's hard to build systems out of existing parts. In *Proceedings of the 17th International Conference on Software Engineering*, pages 179–185. ACM, 1995.

[204] David Remsen. The use and limits of scientific names in biological informatics. *ZooKeys*, (550):207, 2016.

[205] Jonathan D Wren, Raffi Bekeredjian, Jelena A Stewart, Ralph V Shohet, and Harold R Garner. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, 20(3):389–398, 2004.

[206] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: the concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414. ACM, 2001.

[207] Emir Khatipov, Maureen Madden, Pishing Chiang, Philip Chuang, Duc Nguyen, Ivor D'Souza, Rainer Winnenburg, Olivier Bodenreider, Julia Skapik, Rob McClure, and Steve Emrick. Creating, maintaining and publishing value sets in the VSAC. In *AMIA Annual Symposium Proceedings*, page 1459. American Medical Informatics Association, 2014.

[208] US National Library of Medicine. VSAC SVS API v2.

https://www.nlm.nih.gov/vsac/support/usingvsac/vsacsvsapiv2.html. [Accessed: 02-02-2017].

[209] IHE International, Inc. IHE IT Infrastructure (ITI) Technical Framework Volume 1 (ITI TF-1) Integration Profiles. http://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_TF_Vol1.pdf, 2016. [Accessed: 02-02-2017].

[210] Eliot L Siegel and David S Channin. Integrating the healthcare enterprise: a primer: part 1. introduction 1. *Radiographics*, 21(5):1339–1341, 2001.

[211] OMG. Common Terminology Services 2. http://www.omg.org/spec/CTS2/, 2012. [Accessed: 08-08-2017].

[212] ISO/IEC JTC1 SC32 WG2. ISO/IEC 11179. http://metadata-standards.org/11179/. [Accessed: 02-02-2017].

[213] Barbara H Kwasnik. The role of classification in knowledge representation and discovery. *Library Trends*, 48(1):22, 1999.

[214] Rudolf Wille. Formal concept analysis as mathematical theory of concepts and concept hierarchies. In *Formal Concept Analysis*, pages 1–33. Springer, 2005.

[215] Joseph D Novak and Alberto J Cañas. The theory underlying concept maps and how to construct and use them. *Florida Institute for Human and Machine Cognition*, 2008, 2008.

[216] Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213. ACM, 1999.

[217] Lev Muchnik, Royi Itzhack, Sorin Solomon, and Yoram Louzoun. Self-emergence of knowledge trees: extraction of the Wikipedia hierarchies. *Physical Review E*, 76(1):016106, 2007.

[218] Shuting Wang, Chen Liang, Zhaohui Wu, Kyle Williams, Bart Pursel, Benjamin Brautigam, Sherwyn Saul, Hannah Williams, Kyle Bowen, and C Lee Giles. Concept hierarchy extraction from textbooks. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, pages 147–156. ACM, 2015.

[219] Martin S Chodorow, Roy J Byrd, and George E Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics*, pages 299–304. Association for Computational Linguistics, 1985.

[220] Bernard J Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. Real life information retrieval: a study of user queries on the web. In *ACM SIGIR Forum*, volume 32, pages 5–17. ACM, 1998.

[221] Maryam Alavi and Dorothy E Leidner. Review: knowledge management and knowledge management systems: conceptual foundations and research issues. *MIS Quarterly*, 25(1):107–136, 2001.

[222] Kenneth Wai-Ting Leung, Wilfred Ng, and Dik Lun Lee. Personalized concept-based clustering of search engine queries. *IEEE Transactions on Knowledge and Data Engineering*, 20(11):1505–1518, 2008.

[223] Anton V Leouski and W Bruce Croft. An evaluation of techniques for clustering search results. Technical report, DTIC Document, 2005.

[224] Stanislaw Osinski and Dawid Weiss. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54, 2005.

[225] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[226] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[227] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E Tarjan. Clustering social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 56–67. Springer, 2007.

[228] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[229] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[230] Jukka-Pekka Onnela, Jari Saramäki, János Kertész, and Kimmo Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, 2005.

[231] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[232] Michael Krauthammer and Goran Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526, 2004.

[233] Paul Jaccard. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge, 1901.

[234] John Stasko, Richard Catrambone, Mark Guzdial, and Kevin McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies*, 53(5):663–694, 2000.

[235] Kevin J Peterson and Hongfang Liu. An examination of the statistical laws of semantic change in clinical notes. *AMIA Informatics Summit Proceedings*, 2021. (in press).

[236] Angus Roberts. Language, structure, and reuse in the electronic health record. *AMA Journal of Ethics*, 19(3):281–288, 2017.

[237] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352, 2013.

[238] Cecil G Helman. *Doctors and Patients-An Anthology*. CRC Press, 2018.

[239] Qing T Zeng, Doug Redd, Guy Divita, S Jarad, C Brandt, and JR Nebeker. Characterizing clinical text and sublanguage: A case study of the VA clinical notes. *Journal of Health & Medical Informatics*, 3:2, 2011.

[240] Gibson Ferguson. *English for Medical Purposes*, chapter 13, pages 243–261. John Wiley & Sons, Ltd, 2012.

[241] W Tecumseh Fitch. Empirical approaches to the study of language evolution. *Psychonomic Bulletin & Review*, 24(1):3–33, 2017.

[242] Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. Diachronic degradation of language models: insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[243] Stefania Degaetano-Ortlieb and Elke Teich. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, 2018.

[244] Lea Frermann and Mirella Lapata. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016.

[245] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*, 2014.

[246] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, 2015.

[247] Adam Jatowt and Kevin Duh. A framework for analyzing semantic change of words across time. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 229–238. IEEE, 2014.

[248] Krzysztof Jassem and Paweł Skórzewski. Processing historical texts with contemporary NLP tools. *Proceedings of the 8th Language and Technology Conference*, pages 152–157, 2017.

[249] Gerardo Lagunes-García, Alejandro Rodríguez-González, Lucía Prieto-Santamaría, Eduardo P García del Valle, Massimiliano Zanin, and Ernestina Menasalvas-Ruiz. How Wikipedia disease information evolve over time? an analysis of disease-based articles changes. *Information Processing & Management*, 57(3):102225, 2020.

[250] Gintare Grigonyte, Fabio Rinaldi, and Martin Volk. Change of biomedical domain terminology over time. In *Baltic HLT*, pages 74–81, 2012.

[251] Diane E Oliver, Yuval Shahar, Edward H Shortliffe, and Mark A Musen. Representation of change in controlled medical terminologies. *Artificial Intelligence in Medicine*, 15(1):53–76, 1999.

[252] JJ Cimino. High-quality, standard, controlled healthcare terminologies come of age. *Methods of Information in Medicine*, 50(02):101–104, 2011.

[253] Amy Y Wang, James W Barrett, Tim Bentley, David Markwell, Colin Price, Kent A Spackman, and Michael Q Stearns. Mapping between SNOMED RT and Clinical Terms version 3: a key component of the SNOMED CT development process. In *Proceedings of the AMIA Symposium*, page 741. American Medical Informatics Association, 2001.

[254] Eva Pettersson, Beáta Megyesi, and Joakim Nivre. Parsing the past: identification of verb constructions in historical text. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 65–74. Association for Computational Linguistics, 2012.

[255] Marcel Bollmann and Anders Søgaard. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. *arXiv preprint arXiv:1610.07844*, 2016.

[256] Alexandre Allauzen and Jean-Luc Gauvain. Diachronic vocabulary adaptation for broadcast news transcription. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[257] Brian Roark, Murat Saraclar, and Michael Collins. Discriminative n-gram language modeling. *Computer Speech & Language*, 21(2):373–392, 2007.

[258] Lawrence Saul and Fernando Pereira. Aggregate and mixed-order Markov models for statistical language processing. *arXiv preprint cmp-lg/9706007*, 1997.

[259] Edward Loper and Steven Bird. NLTK: the Natural Language Toolkit. *arXiv preprint cs/0205028*, 2002.

[260] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language

modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE, 1995.

[261] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[262] Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? *arXiv preprint arXiv:1506.01070*, 2015.

[263] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. SensEmbed: learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, 2015.

[264] Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[265] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

[266] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[267] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

[268] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[269] Peter H Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[270] Yeshwanth Cherapanamjeri, Prateek Jain, and Praneeth Netrapalli. Thresholding based efficient outlier robust PCA. *arXiv preprint arXiv:1702.05571*, 2017.

[271] Olivier Bodenreider and Lee B Peters. Characterizing the semantic composition of the UMLS Metathesaurus over time. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2016.

[272] Silvio Domingos Cardoso, Cédric Pruski, Marcos Da Silveira, Ying-Chi Lin, Anika Groß, Erhard Rahm, and Chantal Reynaud-Delaître. Leveraging the impact of ontology evolution on semantic annotations. In *European Knowledge Acquisition Workshop*, pages 68–82. Springer, 2016.

[273] Carol Friedman and Stephen B Johnson. Natural language and text processing in biomedicine. In *Biomedical Informatics*, pages 312–343. Springer, 2006.

[274] Emily M Bender. 100 things you always wanted to know about semantics & pragmatics but were afraid to ask. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2018.

[275] Chris Cherpas. Natural language processing, pragmatics, and verbal behavior. *The Analysis of Verbal Behavior*, 10(1):135–147, 1992.

[276] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[277] Jane L Shellum, Rick A Nishimura, Dawn S Milliner, Charles M Harper Jr, and John H Noseworthy. Knowledge management in the era of digital medicine: a programmatic approach to optimize patient care in an academic medical center. *Learning Health Systems*, 1(2):e10022, 2017.

[278] Hongfang Liu, Kavishwar B Wagholikar, Siddhartha Jonnalagadda, and Sunghwan Sohn. Integrated cTAKES for concept mention detection and normalization. In *CLEF (Working Notes)*, 2013.

[279] Frederick North, Sarah J Crane, Rajeev Chaudhry, Jon O Ebbert, Karen Ytterberg, Sidna M Tulledge-Scheitel, and Robert J Stroebel. Impact of patient portal secure messages and electronic visits on adult primary care office visits. *Telemedicine and e-Health*, 20(3):192–198, 2014.

[280] James A Dilling, Stephen J Swensen, Michele R Hoover, Gene C Dankbar, Amerett L Donahoe-Anshus, M Hassan Murad, and Jeff T Mueller. Accelerating the use of best practices: the Mayo Clinic model of diffusion. *Joint Commission Journal on Quality and Patient Safety*, 39(4):167–176, 2013.

[281] Frederick North, Samuel Fox, and Rajeev Chaudhry. Clinician time used for decision making: a best case workflow study using cardiovascular risk assessments and Ask Mayo Expert algorithmic care process models. *BMC Medical Informatics and Decision Making*, 16(1):96, 2016.

[282] David S Carrell, Scott Halgrim, Diem-Thy Tran, Diana S M Buist, Jessica Chubak, Wendy W Chapman, and Guergana Savova. Using natural language processing

to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *American Journal of Epidemiology*, 179(6):749–758, 01 2014.

[283] Katrina F Hurley and Syed Sibte Raza Abidi. Ontology engineering to model clinical pathways: towards the computerization and execution of clinical pathways. In *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*, pages 536–541. IEEE, 2007.